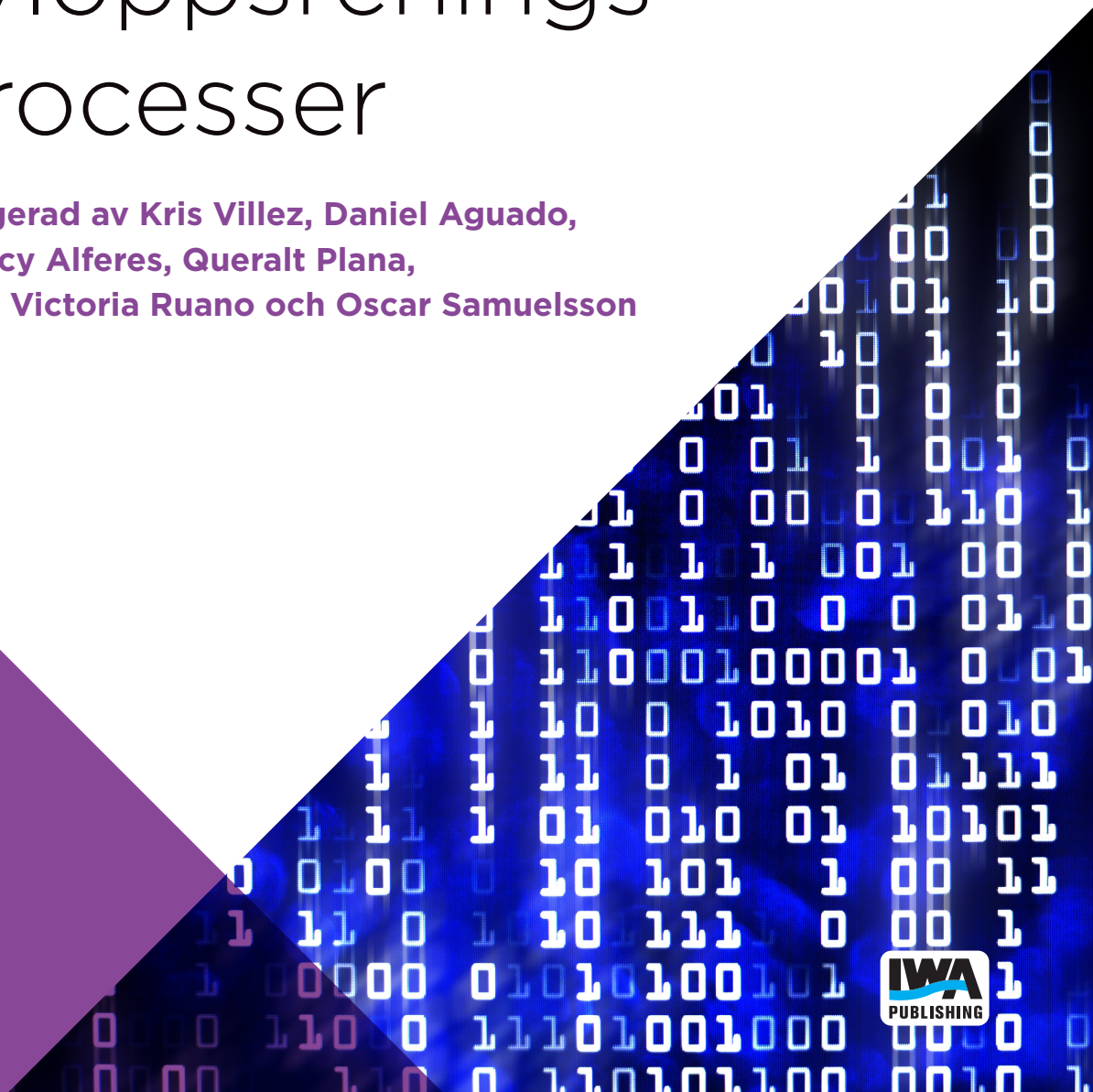
Scientific and Technical Report Series No. 31a

Bästa möjliga data från givare i avloppsrenings-processer

Redigerad av Kris Villez, Daniel Aguado,
Janelcy Alferes, Queralt Plana,
Maria Victoria Ruano och Oscar Samuelsson



IWA PUBLISHING

Bästa möjliga data från givare i avloppsreningsprocesser

Bästa möjliga data från givare i avloppsreningsprocesser

Redigerad av

Kris Villez, Daniel Aguado, Janelcy Alferes, Queralt Plana, Maria Victoria Ruano och Oscar Samuelsson

IWA PUBLISHING

Published by **IWA Publishing**
Unit 104–105, Export Building
1 Clove Crescent
London E14 2BA, UK
Telephone: +44 (0)20 7654 5500
Fax: +44 (0)20 7654 5555
Email: publications@iwap.co.uk
Web: www.iwaponline.com

English edition first published 2024
© 2024 IWA Publishing

Swedish edition first published 2025
© 2025 IWA Publishing

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the UK Copyright, Designs and Patents Act (1998), no part of this publication may be reproduced, stored or transmitted in any form or by any means, without the prior permission in writing of the publisher, or, in the case of photographic reproduction, in accordance with the terms of licenses issued by the Copyright Licensing Agency in the UK, or in accordance with the terms of licenses issued by the appropriate reproduction rights organization outside the UK. Enquiries concerning reproduction outside the terms stated here should be sent to IWA Publishing at the address printed above.

The publisher makes no representation, express or implied, with regard to the accuracy of the information contained in this book and cannot accept any legal responsibility or liability for errors or omissions that may be made.

Disclaimer
The information provided and the opinions given in this publication are not necessarily those of IWA and should not be acted upon without independent consideration and professional advice. IWA and the Editors and Authors will not accept responsibility for any loss or damage suffered by any person acting or refraining from acting upon any material contained in this publication.

British Library Cataloguing in Publication Data
A CIP catalogue record for this book is available from the British Library

ISBN: 9781789065107 (eBook)
ISBN: 9781789065114 (ePub)

This eBook was made Open Access in June 2025.

Doi: 10.2166/9781789065107

© 2025 IWA Publishing
This is an Open Access book distributed under the terms of the Creative Commons Attribution Licence (CC BY-NC-ND 4.0), which permits copying and redistribution for non-commercial purposes with no derivatives, provided the original work is properly cited (https://creativecommons.org/licenses/by-nc-nd/4.0/). This does not affect the rights licensed or assigned from any third party in this book.

Innehållsförteckning

Sammanfattning xiii

Förord xv

Kapitel 1 1
Inledning 1
D. Aguado, J. Alferes, Q. Plana, M. V. Ruano, O. Samuelsson och K. Villez

1.1 Rapportens fokus: data och metadata 1
1.2 Incitament för att samla in och organisera metadata och behovet av denna rapport 2
1.3 Läsanvisningar 3
1.4 Tackord och avslutande kommentarer 3

Kapitel 2 7
Datatyper: terminologi och exempel 7
Q. Plana och K. Villez

2.1 Läsanvisningar 7
2.2 Inledning 8
2.3 Signaler på ett typiskt reningsverk 8
2.4 Signaldata, laboratoriedata och metadata 9
2.4.1 Signaldata 9
2.4.1.1 Datapunktens dimensionalitet 9
2.4.1.2 Definitioner 9
2.4.1.3 Dataförluster 11
2.4.1.4 Hur lagrar man information som beskriver förlorade data? 11
2.4.1.5 Ytterligare kommentarer 12
2.4.2 Exempel på signaldata 13
2.4.2.1 Exempel E2.1: börvärde för luftflödesstyrning 13
2.4.2.2 Exempel E2.2: mätning av luftflöde 13
2.4.2.3 Exempel E2.3: absorbansspektrofotometri med ultraviolett ljus 14

2.4.2.4 Exempel E2.4: flödescytometri . . . 15
2.4.3 Laboratoriedata . . . 16
2.4.3.1 Definition . . . 17
2.4.3.2 Exempel E2.5: laboratoriemätningar . . . 17
2.4.4 Metadata . . . 18
2.5 Strukturerade och ostrukturerade data . . . 19
2.5.1 Strukturerade data . . . 19
2.5.1.1 Definition . . . 19
2.5.1.2 Ytterligare anmärkningar . . . 19
2.5.1.3 Exempel E2.1 och E2.2 (fortsättn.): univariata signaler . . . 20
2.5.1.4 Exempel E2.3 (fortsättn.): absorbansspektrofotometri med ultraviolett ljus . . . 20
2.5.1.5 Exempel E2.5 (fortsättn.): koncentrationsmätningar . . . 20
2.5.2 Ostrukturerade data . . . 20
2.5.2.1 Definition . . . 20
2.5.2.2 Hantering av ostrukturerade data . . . 20
2.5.2.3 Ytterligare kommentarer . . . 21
2.5.2.4 Exempel E2.4 (fortsättn.): flödescytometri . . . 22
2.5.2.5 Exempel E2.5 (fortsättn.): text . . . 22
2.5.2.6 Exempel E2.6: balansmatris för att representera anläggningens layout och flödesschema . . . 23
2.6 Versionshantering av data . . . 25
2.6.1 Hantering av dynamiska datamängder . . . 25
2.6.2 Definitioner . . . 26
2.6.2.1 Originaldata eller rådata . . . 26
2.6.2.2 Primärdata . . . 26
2.6.2.3 Sekundärdata . . . 26
2.6.2.4 Spårbarhet av data: dokumenterad historik från rådata till sekundärdata 27
2.6.2.5 Versionshantering av data och mjukvara: likheter och skillnader . . . 28
2.7 Slutsater . . . 28
Tackord . . . 29

Kapitel 3 . . . 31
Metadata för en systematisk beskrivning av signaldata . . . 31
J. Alferes, J. A. Baeza, N. Nicolaï, J-D. Therrien, Q. Plana och K. Villez

3.1 Läsanvisningar . . . 31
3.2 Metadata som beskriver signalgenerering: information om utrustning och processer för insamling av givardata . . . 32
3.2.1 Prioritering av signalgenererade metadata . . . 33
3.2.2 Metadata som beskriver givardata . . . 33
3.2.2.1 Tidsstämpelns ursprung . . . 33
3.2.2.2 Tidsupplösning eller samplingsfrekvens . . . 33
3.2.2.3 Numerisk upplösning (kvantisering) . . . 35
3.2.2.4 Givarens placering . . . 35
3.2.2.5 Mätenhet . . . 36
3.2.2.6 Mätområde . . . 36
3.2.2.7 Variabelns fysiska intervall . . . 36
3.2.2.8 Signalöverföring (analogt/digitalt/manuellt) . . . 36
3.2.2.9 Mätprincip . . . 37

3.2.2.10 Syfte med mätningen . . . 37
3.2.3 Metadata som beskriver den signalproducerande enheten (instrumentet) . . . 38
3.2.3.1 Produktspecifikationer och guider . . . 38
3.2.3.2 Tillverkare, modellnamn och serie . . . 38
3.2.3.3 Unik identifierare . . . 38
3.2.3.4 Instrumentets historia . . . 39
3.2.4 Metadata som beskriver modifieringar av signalen . . . 39
3.2.4.1 Förändring av enheter . . . 40
3.2.4.2 Modifiering med informationsförlust (aggregering, medelvärdesbildning, filtrering, komprimering) . . . 40
3.2.4.3 Ersättning av data (imputation, interpolering, synkronisering) . . . 41
3.2.4.4 Datafusion och softsensorer . . . 41
3.3 Metadata som beskriver kvaliteten på givardata: noggrannhet, riktighet, precision . . . 41
3.3.1 Beskriva mätfel genom deras observerade symtom . . . 42
3.3.1.1 Fel som avvikelser från referensvärden . . . 42
3.3.1.2 Statisk beskrivning av datakvalitet - grunderna: noggrannhet, riktighet och precision . . . 43
3.3.1.3 Statisk beskrivning av datakvalitet – avancerad nivå: repeterbarhet och reproducerbarhet . . . 50
3.3.1.4 Redovisning av dynamik: responstid, chock, drift och extremvärden . . . 50
3.3.2 Beskriva mätfel genom deras grundorsaker . . . 54
3.3.2.1 Översikt . . . 54
3.3.2.2 Systematiska fel . . . 55
3.3.2.3 Tillfälliga fel . . . 56
3.3.2.4 Slumpmässiga fel . . . 57
3.3.3 Diagnostik: identifiera de orsaker som förklarar symtomen . . . 57
3.4 Metadata som ger kontextuell information: annotering . . . 58
3.4.1 Syfte . . . 59
3.4.2 Grundläggande definitioner . . . 59
3.4.3 Strategier för att hantera utmaningar med annoteringar . . . 60
3.4.3.1 Koppla samman annoteringar med andra metadata och signaler . . . 60
3.4.3.2 Flexibla annoteringsmetoder leder till heterogena metadata . . . 60
3.4.3.3 Kommunikation mellan olika expertdomäner, arbetsteam och över tid . . . 61
3.4.3.4 Standardisering av annoteringar . . . 61
3.4.3.5 Automatisera annoteringar . . . 61
3.4.4 Kategorisering av annoteringar . . . 61
3.4.4.1 Kategorisering utifrån tid och plats . . . 62
3.4.4.2 Kategorisering utifrån förklaringsgrad . . . 62
3.4.4.3 Kategorisering utifrån informationskälla . . . 63
3.4.4.4 Kategorisering enligt annoteringsprocessen . . . 65
3.5 Slutsatser . . . 65
Tackord . . . 65

Kapitel 4 . . . 69
Strukturerad lagring av signaler samt deras metadata . . . 69
N. Nicolaï, K. Villez och Q. Plana

4.1 Läsanvisningar . . . 69
4.2 Vikten av att strukturera data . . . 69

4.2.1 Strukturera för nuläget: den pågående digitala revolutionen . . . 70
4.2.1.1 Data som en resurs för information . . . 70
4.2.1.2 Bättre beslutsfattande och kostnadsbesparingar . . . 71
4.2.2 Strukturera data för framtiden . . . 71
4.2.2.1 Förberedelser för det okända (s.k. unknown uknowns) . . . 71
4.2.2.2 Integrering av interna och externa data (“data brokerage”) . . . 71
4.3 Datalivscykeln: vilka data ska lagras och vilka ska behållas? . . . 72
4.4 Teknik för implementering . . . 73
4.4.1 Motiv . . . 73
4.4.2 Fildatabas (“flat-file database”) . . . 74
4.4.3 Relationsdatabaser . . . 75
4.4.4 Tidsseriedatabas . . . 78
4.4.5 Historikdatabaser . . . 80
4.4.5.1 Allmän beskrivning . . . 80
4.4.5.2 Komprimeringstekniker som används i historikdatabaser . . . 82
4.4.6 Datalager . . . 83
4.4.7 Data lake . . . 86
4.4.8 Branschexempel . . . 88
4.4.8.1 Exempel 1: pilEAUte pilotanläggning för avloppsvattenrening . . . 88
4.4.8.2 Exempel 2: förenklat flödesschema för processdata på Hampton Roads Sanitation District (HRSD) anläggningar . . . 89
4.5 Förståelse för relationsdatabaser . . . 91
4.5.1 Definition . . . 91
4.5.2 De vanligaste datarelationerna . . . 93
4.5.2.1 [datapunkt] producerad av [enhet] . . . 93
4.5.2.2 [datapunkt] producerad vid [plats] . . . 93
4.5.2.3 [kvalitet] baserat på [datapunkter] . . . 93
4.5.3 Relationer som beskriver datakvalitet . . . 94
4.5.4 Definiera en relationsstruktur . . . 96
4.5.5 Två branschexempel . . . 97
4.5.5.1 Exempel E4.1: WaterHubDataPool . . . 97
4.5.5.2 Exempel E4.2: dateEAUbase . . . 99
4.6 Slutsatser . . . 100

Kapitel 5 . . . 103
Praktiska metoder för validering av datakvalitet från givare . . . 103
J. Alferes, O. Grievson, C. Heffernan, D. Lumley, T. Massiri, A. Menniti, O. Samuelsson, K. Villez och E. Zegers

5.1 Läsanvisingar . . . 104
5.2 Inledning . . . 104
5.3 Terminologi för datakvalitet från givare . . . 105
5.3.1 Definitioner . . . 105
5.3.2 Exempel E5.2: kalibrering av offset- och skalningsfel . . . 109
5.4 Värdet av metadata för givarunderhåll . . . 110
5.4.1 Optimering av kalibreringsfrekvensen . . . 110
5.4.1.1 Optimering av kalibreringsfrekvens utifrån drift . . . 110
5.4.1.2 Exempel E5.3: övervakning av offset och lutning för två pH-givare . . . 111
5.4.2 Optimal rengöringsfrekvens . . . 112

5.4.3 Delning av data inom reningsverket . . . 112
5.4.4 Delning av data mellan organisationer och modellbygge . . . 112
5.5 Utformning av valideringsrutiner för givare . . . 113
5.5.1 Ett strukturerat arbetssätt för validering av givare . . . 113
5.5.2 Allmänna rekommendationer för tillförlitliga referensmätningar . . . 114
5.5.3 Mallar och exempel på standardiserade arbetsrutiner för validering av givare . . . 115
5.6 Ett praktisk exempel på automatisk givarjustering och insamling av metadata . . . 116
5.7 Slutsatser . . . 117
Tackord . . . 118
Referenser . . . 118
Bilaga 5.A: Standardiserade valideringsrutiner för givare . . . 119
5.A.1 Validering av syregivare . . . 119
5.A.1.1 DO-rutin-1: Kvantifiering av bias med en handhållen referensgivare . . . 119
5.A.1.2 DO-rutin-2a: Tvåpunktsvalidering med vattenmättad luft . . . 120
5.A.1.3 DO-Procedur-2b: Tvåpunktsvalidering med syremättat vatten . . . 123
5.A.1.4 Beskrivande information för validering av syregivare . . . 125
5.A.2 Procedurer för validering av pH-givare . . . 126
5.A.2.1 pH-rutin-1: Enpunktsvalidering av pH-givare vid aktuell driftspunkt . . . 126
5.A.2.2 pH-rutin-2: Tvåpunktsvalidering av pH-givare i buffert . . . 128
5.A.2.3 pH-rutin-3: Flerpunktsvalidering av pH-givare i buffertlösningar . . . 130
5.A.3 Valideringsrutiner för ammoniumgivare . . . 132
5.A.3.1 TAN-rutin-1: Enpunktsvalidering av ammoniumgivare vid aktuell driftspunkt . . . 132
5.A.3.2 TAN-Procedure-2: procedur för validering av flerpunktsgivare för ammoniak vid aktuell driftspunkt . . . 134
5.A.3.3 Kontextuell information för validering av ammoniumgivare . . . 136
5.A.4 Valideringsrutin för givar som mäter total suspenderade halt . . . 137
5.A.4.1 TSS-rutin-1: Funktionskontroll av givaren . . . 137
5.A.4.2 TSS-rutin-2: Kvantifiering av bias . . . 138
5.A.4.3 TSS-rutin-3: Tvåpunktsvalidering vid noll susphalt och en vald koncentration . . . 139
5.A.4.4 TSS-rutin-4: Flerpunktsvalidering av susphaltsgivare . . . 141
5.A.4.5 Kontextuell information om validering av suspgivare . . . 143
5.A.5 Valideringsrutiner för givare med absorbansmätning av UV-ljus . . . 144
5.A.5.1 UV/Vis-rutin-1: Kontroll av givarfunktion . . . 144
5.A.5.2 UV/Vis-rutin-2: Enpunktsvalidering vid aktuell driftpunkt . . . 146
5.A.5.3 Kontextuell information om spektrala absorbansmätningar av UV-ljus . . . 147
5.A.5.4 Allmänna rekommendationer för validering av spektrometrar . . . 148
5.A.6 Mall för valideringsrutiner . . . 149
5.A.6.1 Mall för valideringsrutin av givare . . . 149
5.A.6.2 Kontextuell information . . . 150
Bilaga 5.B: Validering av vattenflödes-, nivå-, tryck- och temperaturgivare . . . 151
5.B.1 Inledning . . . 151
5.B.2 Validering av flödesgivare . . . 151
5.B.2.1 Elektromagnetiska flödeshastighetsgivare . . . 151
5.B.2.2 Flödesmätning i öppen kanal . . . 152
5.B.2.3 Flödesgivare med area-hastighetsmätning . . . 154
5.B.2.4 Flödesgivare med flygtidsmätning . . . 154
5.B.2.5 Andra tekniker för validering av flödesmätare . . . 154
5.B.3 Validering av nivågivare . . . 155

5.B.4 Validering av tryckgivare 156
5.B.5 Validering av temperaturgivare 158
5.B.5.1 Termoelement 158
5.B.5.2 Resistansgivare 159
5.B.5.3 Termistorer 160
5.B.5.4 Installation 160
5.B.5.5 Validering och kalibrering av temperaturgivare 160

Kapitel 6 161
Dataanalytiska metoder för kvalitetsbedömning av givardata 161
D. Aguado, J. Alferes, F. Arteaga, L. Belia, J. B. Copp, L. Corominas, F. Corona, A. Ferrer, H. Haimi, P. Kazemi, Q. H. Le, I. Miletic, M. Mulas, A. Robles, M. V. Ruano, S. Russo, O. Samuelsson, J. P. Steyer, K. Villez, E. I. P. Volcke, M. J. Wade och J. Zambrano

6.1 Läsanvisningar 162
6.2 Översikt 162
6.3 Redundans: kärnan i datadriven kvalitetsbedömning 163
6.3.1 Redundans via en referensmätning 164
6.3.2 Referensmätningar utifrån historiska data 165
6.3.3 Modellbaserad redundans 166
6.4 Ett typiskt dataflöde vid databehandling 166
6.5 Preliminär dataanalys och visualisering 170
6.6 Förbehandling av data 174
6.6.1 Avvikelsedetektion 174
6.6.1.1 Syfte 174
6.6.1.2 Exempel på metoder för avvikelsedetektion 175
6.6.2 Ersättning av data 175
6.6.2.1 Syfte 175
6.6.2.2 Exempel på metoder för att ersätta data 176
6.6.3 Filtrering och utjämning av data 181
6.6.3.1 Syfte 181
6.6.3.2 Exempel på filtreringsmetoder 181
6.6.4 Karaktärsdrag i data 185
6.6.4.1 Syfte 185
6.6.4.2 Exempel på metoder 185
6.6.5 Datatransformation 187
6.6.5.1 Syfte 187
6.6.5.2 Exempel på transformationsmetoder 187
6.6.6 Skalning av data 188
6.6.6.1 Syfte 188
6.6.6.2 Exempel på skalningsmetoder 188
6.6.7 Datastrukturering: metoder för att hantera ostrukturerade data 190
6.6.7.1 Syfte 190
6.6.7.2 En kommentar om icke-reversibel datastrukturering 190
6.6.7.3 Metoder för att strukturera mätdata för partiklar 190
6.6.7.4 Metoder för att strukturera textdata 194
6.7 Analys av datakvalitet 194
6.7.1 Mekanistiska metoder 194

6.7.1.1 Syfte och tillämpning för bedömning av datakvalitet 194
6.7.1.2 Metoder 195
6.7.1.3 Potential med datakorrigering på reningsverk 198
6.7.2 Datadrivna metoder. 200
6.7.2.1 Syfte och tillämpning för bedömning av datakvalitet 200
6.7.2.2 Övervakade tillvägagångssätt 200
6.7.2.3 Oövervakade metoder 205
6.7.3 Riktlinjer för val av metod 216
6.7.3.1 Ramverk för att analysera datakvalitet med datadrivna metoder 216
6.7.3.2 Val av metod(er) 217
6.7.3.3 Ytterligare praktiska anmärkningar 218
6.7.4 Prestandautvärdering av datadrivna feldetektionsmetoder 221
6.7.4.1 Arbetsflöde 221
6.7.4.2 Ett praktisk exempel på prestandautvärdering. 225
6.7.4.3 Nyttor och fallgropar vid prestandautvärdering av datadrivna metoder. 228
6.8 Slutsatser. 229

Kapitel 7 239
Slutsatser och framtidsutblick 239
D. Aguado, J. Alferes, A. Menniti, Q. Plana, M. V. Ruano, O. Samuelsson, K. Villez och E. Zegers

7.1 Identifierade utmaningar med metadata. 239
7.1.1 Varför? – hur man motiverar metadata. 239
7.1.1.1 Delning av data. 240
7.1.1.2 Främja en metadatabejakande organisation 240
7.1.1.3 Kostnaden för metadata 240
7.1.1.4 Minskad osäkerhet och ökat förtroende för ett datadrivet beslutsfattande 241
7.1.2 Utmaningar med att implementera metadatastrategier. 241
7.1.2.1 Förankra värdet med metadata 241
7.1.2.2 Främja en metadatakultur. 241
7.1.2.3 Tekniska utmaningar för att implementera god metadatapraxis 242
7.2 Bakgrunden till denna rapport och hantering av observerade metadatautmaningar 242
7.2.1 Förstå vad metadata är eller kan vara 242
7.2.2 Förståelse för hur metadata bäst organiseras 242
7.2.3 En generell praxis för metadata från givare 243
7.2.4 En översikt över datadrivna metoder för validering av givardata 243
7.3 Användarperspektivet: att komma igång med data 244
7.3.1 Skapa en kultur för datahantering – data governance 244
7.3.2 System för hantering av datakvalitet. 244
7.3.3 Dataarkitektur och integrationer. 245
7.3.4 Digital mognadsgrad 246
7.4 En plan för datatransformation 246
7.4.1 Steg 1: Diagnos. 247
7.4.2 Steg 2: Genomförande. 248
7.4.2.1 Automatisera manuell insamling av data och metadata 248
7.4.2.2 Utforma ett ramverk för datahantering 248
7.4.2.3 Inrätta ett system för hantering av datakvalitet 248

7.4.2.4 Dataintegration och implementering av ett datahanteringssystem 248
7.4.2.5 Smarta beslutsstöd 249
7.4.2.6 Fortsatt digitalisering. 249
7.5 Framtidsutblick 250

Ordlista 253

Sammanfattning

Utvecklingen med instrumentering på avloppsreningsverk har revolutionerat möjligheterna för processövervakning och reglering. Effektiv dataöverföring och kraftigt reducerade kostnader för datalagring har inneburit att reningsverk i dag förfogar över stora mängder data – eran med Big data har nått VA-branschen. Samtidigt har de senaste årens utveckling inom AI och maskininlärning gjort det möjligt att utvinna värdefull information ur dessa stora datamängder.

Men trots digitaliseringens möjligheter så krävs både förmågan att hantera stora datamängder och tillräcklig datakvalitet för att fullt ut nyttiggöra AI och datadriven automatisering på ett effektivt och framgångsrikt sätt. Dokumentering och lagring av metadata, exempelvis karakterisering av datakvaliteten, är ofta en bristvara. Det leder till att stora delar av insamlade data riskerar att gå till spillo eftersom de med tiden blir svårare och svårare att tolka och därmed snabbt blir utdaterade. Den krassa verkligheten är att datadrivna beslutsstöd med AI och maskininlärning aldrig kommer att realiseras om reningsverkets organisation saknar kunskap och förmåga att hantera de fyra V som Big data innebär: datamängd (*Volume*), hastighet (*Velocity*), variation (*Variability*) samt noggrannhet och tillförlitlighet (*Veracity*).

Rapporten *Bästa möjliga data från givare i avloppsreningsprocesser* är en svensk översättning av den engelska rapporten *Metadata Collection and Organization in Wastewater Treatment and Wastewater Resource Recovery Systems*. Rapporten är en guide till att hantera dagens datatekniska utmaningar och syftar till att beskriva vad metadata innebär och hur de rent praktiskt kan användas för att nyttiggöra reningsverksdata. Rapporten ger råd om hur metadata bör samlas in, tolkas, struktureras och lagras utifrån befintliga standarder och de krav som datadrivna algoritmer ställer på data. Den är en bra utgångspunkt för läsare som vill förbättra sina datahanteringsrutiner och förstå hur de kan utveckla metoder för datadrivna beslutsstöd. Den riktar sig till en bred målgrupp såsom instrumenttekniker, driftpersonal, driftchefer och allmänt dataintresserade personer på reningsverk.

Förord

Miljöbalken kräver att verksamhetsutövaren använder bästa möjliga teknik (BMT) för att minska miljöpåverkan. Men en förutsättning för att dagens avancerade avloppsreningsprocesser ska prestera på “bästa möjliga sätt” är att de styrs utifrån tillförlitliga onlinegivare. Eller som Bengt Carlsson, professor emeritus i reglerteknik brukar säga – en process kan ge godtyckligt dålig prestanda om den regleras utifrån felaktiga mätvärden. Tillförlitliga mätvärden från onlinegivare är alltså en del i, och en förutsättning för, att uppnå BMT och därmed också en grundförutsättning för hög reningsgrad på våra reningsverk.

Men data från onlinegivare kan användas till mer än processreglering. I takt med digitaliseringens framfart blir datadrivna beslutsstöd, digitala tvillingar och AI-modeller allt viktigare. Dessa teknologier kräver stora mängder data av hög kvalitet. Men en utmaning är att kunna beskriva, validera, migrera och omvandla data effektivt. Utan metadata – data som beskriver data – blir det svårt, eller till och med omöjligt, att tolka och använda data som någon annan har samlat in.

Det ökande behovet av metadata blev tydligt under konferensen Instrumentation Control and Automation 2017, vilket ledde till att Kris Villez initierade en s.k. task group inom International Water Association (IWA) med namnet Metadata Collection and Organization (MetaCO). Syftet med MetaCO var att sammanställa de bästa metoderna för att generera, lagra och nyttja metadata från onlinegivare i en teknisk-vetenskaplig rapport. När förslaget skickades in till IWA för sex år sedan så var begreppet machine learning ett nischområde på universiteten. I dag används AI i alla discipliner och verksamheter. Men ett behov kvarstår – tillgången till bästa möjliga data.

SVU-projektet *Bästa möjliga data* startades parallellt med MetaCO för att bidra och skapa synergier på flera sätt, bland annat genom att översätta den tekniskt-vetenskapliga slutrapporten till svenska. Översättningen har gjorts med DeepL:s språkmodell för maskinöversättning som jag sedan har justerat för att passa målgruppen processingenjörer och dataanvändare på reningsverk. Målsättningen har varit att översätta det tekniska innehållet korrekt, snarare än ordagrant. Många engelska tekniska begrepp saknar svenska motsvarigheter. I dessa fall har jag valt att behålla det ursprungliga engelska namnet antingen i texten eller med kursiv stil inom parentes, för att underlätta för läsaren att söka sig vidare med korrekt sökterm.

Jag vill rikta ett stort tack till alla processingenjörer, instrumenttekniker och allmänt dataintresserade personer på VA-organisationerna runt om i landet som bidragit med tid och erfarenhet. Ert engagemang har gjort denna rapport till ett värdefullt verktyg. Ett särskilt tack till Mjölby kommun, Stockholm vatten och avfall, Eskilstuna Strängnäs energi och miljö och Mälarenergi vatten som tillsammans

med Svenskt vatten utveckling och Stiftelsen för IVL har finansierat projektet. Tack vare er kan denna rapport nu spridas som Open Access på svenska.

Jag hoppas att VA-Sverige kommer ha nytta av rapporten så att de data som med stor möda samlas in verkligen kommer till användning, och i slutändan bidrar till bättre vattenrening. Svenska vattendrag är värda bästa möjliga teknik (och data).

Oscar Samuelsson, Sjundemåla, Mars 2025

doi: 10.2166/9781789065107_0001

Kapitel 1
Inledning

D. Aguado[1], J. Alferes[2], Q. Plana[3,4], M. V. Ruano[5], O. Samuelsson[6] och K. Villez[7]

[1]*CALAGUA – Unidad Mixta UV-UPV, Institut Universitari d'Investigació d'Enginyeria de l'Aigua i Medi Ambient – IIAMA, Universitat Politècnica de València, València, Spanien*
[2]*VITO – WaterClimateHub, Ostende, Belgien*
[3]*modelEAU, Université Laval, Québec, QC, Kanada*
[4]*Eurecat – Kataloniens teknikcentrum, Barcelona, Spanien*
[5]*CALAGUA – Unidad Mixta UV-UPV, Departament d'Enginyeria Química, Universitat de València, Burjassot, Spanien*
[6]*IVL Svenska Miljöinstitutet, Stockholm, Sverige*
[7]*Oak Ridge National Laboratory, Oak Ridge, TN, USA*

Denna rapport gör en djupdykning i s.k. metadata från givare på avloppsreningsverk, inklusive metoder för att generera metadata. I detta inledande kapitel beskrivs drivkrafterna bakom den IWA task group MetaCO (Meta-data Collection and Organization) som har skrivit denna rapport. Drivkrafterna för rapporten inkluderar historiska datatekniska utmaningar (stora datamängder och dataflöden) och mer aktuella utmaningar såsom datastrukturer (hantering av olika typer och strukturer av data) och datakvalitet (validering av tillförlitlighet hos data). Kapitlet avslutas med ett stort tack till alla som bidragit till rapporten.

1.1 RAPPORTENS FOKUS: DATA OCH METADATA

Definitionerna av metadata varierar. I den snävaste bemärkelsen beskriver metadata var, när och hur data har producerats. Denna definition kan behöva utvidgas då även motiven för själva datainsamlingen (varför behövs givaren), givarnas installation, underhåll och skick är nyttiga metadata. Att ha tillgång till dessa typer av metadata är avgörande för att skapa förtroende för själva data som givaren producerar och de analyser som i sin tur baseras på dem. Kontextuell information om t.ex. väder eller belastning är också ofta värdefullt för att avgöra om en viss delmängd data är relevant för modellering, övervakning eller styrning. Att jämföra givarsignaler med referensmätningar eller kontrollerade standardlösningar kan vara avgörande för att validera nya dataintensiva tekniker för modellering och beslutsstöd. Slutligen kan databaserade tekniker såsom maskininlärning avslöja mönster eller trender som är värdefulla för efterföljande tolkning eller användning av de analyserade

© 2025 IWAP. Detta är ett Open Access-bokkapitel som distribueras enligt villkoren i Creative Commons Attribution License (CC BY-NC-ND 4.0) som tillåter kopiering och vidaredistribution för icke-kommersiella ändamål utan derivat, förutsatt att arbetet citeras korrekt (https://creativecommons.org/licenses/by-nc-nd/4.0/). Kapitlet är från boken *Bästa möjliga data från givare i avloppsreningsprocesser*, Kris Villez, Daniel Aguado, Janelcy Alferes, Queralt Plana, Maria Victoria Ruano, Oscar Samuelsson (redaktörer).

sensorsignalerna. Alla dessa aspekter betraktar vi i denna rapport som metadata, vilka i den bredaste bemärkelsen alltså är information som beskriver något av följande:

- Metoden för att generera en signal, signalproducerande hårdvara och signalbehandling.
- Givarens skick och tillstånd, samt signalens kvalitet.
- Kontextuell information som beskriver de förhållanden under vilka data producerades.
- Information som givarsignaler som fås genom visuell inspektion eller datorstödd analys.

Eftersom detta leder till ett brett spektrum av metadata, avgränsas rapporten till (a) tillämpningar för avloppsvattenrening och (b) metadata för givare (*online-sensors*). Detta innebär att tillämpningar för ledningsnät och dricksvattenproduktion inte berörs, även om de står inför liknande utmaningar och likväl kan ha nytta av rapportens innehåll. En ytterligare avgränsning är att laboratoriedata och ställdon (t.ex. ventilbörvärdessignaler) inte ingår om det inte är motiverat utifrån pedagogiska skäl. Slutligen avgränsas rapporten från att beröra cybersäkerhet och IT-infrastruktur, vilka är viktiga men inte unika för tillämpningen avloppsvattenrening. Definitioner och rekommendationer kan naturligtvis tillämpas på andra ställen inom ett avloppsreningsverk (t.ex. digitala tvillingar och modeller, laboratoriedata och börvärden). Rapporten kan förhoppningsvis även tjäna som inspiration till andra tekniska discipliner.

1.2 INCITAMENT FÖR ATT SAMLA IN OCH ORGANISERA METADATA OCH BEHOVET AV DENNA RAPPORT

Givarutveckling, datorkraft och lagringskapacitet har drivit på utvecklingen av Big data i avloppsvattensektorn. Big data kan delas in i fyra V:n som beskriver data: datamängd (*Volume*), hastighet (*Velocity*), variation (*Variability*) och noggrannhet och tillförlitlighet (*Veracity*). Hittills har en stor del av utvecklingen av datainfrastrukturen på reningsverken fokuserat på att hantera större datamängder och snabbare insamlingshastigheter. Givare har ökat samplingshastigheten för mätningar så att den överstiger de praktiska behoven, och datainsamling i realtid är nu normen. Effekten av billig datalagringskapacitet har ytterligare lett till stora databaser, som effektivt kan logga processhistoriken för reningsverk i alla storleksklasser.

De många olika datatyperna är dock en utmaning för standardiserade rutiner för insamling och lagring av data. Data produceras av nya givare, både innanför och utanför reningsverket (*inside and outside the fence*). Behovet av att integrera externa databaser och system med de befintliga databaserna för processdata är uppenbart. Annars blir data snart föråldrade, vilket leder till datastuprör (*data silos*) med begränsad eller ingen nytta för verksamheten. Även när datakällor integreras inom reningsverket kan det fortfarande vara svårt att maximera nyttan av tillgängliga data. Metadata, inklusive en kvantifiering av datakvaliteten, är ofta bristfällig, vilket gör att de stora mängderna insamlade data snabbt blir inaktuella.

Parallellt med ny och befintlig givarteknologi används mjukvarusensorer (*softsensors*) och modeller (digitala tvillingar) i allt större utsträckning för att producera ytterligare signaler och data för drift och styrning av anläggningen. Dessa kan hanteras internt eller genom en underleverantör. Mjukvarusensorer och digitala tvillingar är beroende av korrekta och relevanta data. Att beskriva data på ett meningsfullt sätt är ofta det första steget för att göra detta genomförbart.

Det har alltid varit viktigt att säkerställa korrekta data, särskilt när de används för processreglering. Hittills har den vanligaste regleringen varit den proportionella och integrerande (PI) regulatorn, som är enkel men tillräckligt avancerad för att hantera de flesta reglertekniska problemen på ett reningsverk. I en PI-regulator är effekten av en drivande givare förutsägbar, och konsekvenserna kan förstås och hanteras inför en implementering. Det blir dock mer komplicerat när flera signaler och beräkningar används samtidigt för regleringen. Den nuvarande trenden mot användning av maskininlärning, mjukvarusensorer och digitala tvillingar gör det svårare att

redogöra för konsekvenserna av givardata med dålig noggrannhet. Den pågående digitaliseringen av reningsverken drivs inte bara av möjligheter, utan också av strävan att omvandla reningsverk (*Wastewater treatment plant – WWTP*) till en anläggning för resursåterföring (*Water Resource Recovery Facitlties – WRRF*). Denna övergång kräver mer avancerad teknik som samverkar med befintlig teknik och rutiner. Dessutom blir reningskraven mer strikta, samtidigt som utsläppen av växthusgaser också måste minimeras. Nya sensorer, data och smarta beslutsstödsystem behövs för att balansera dessa nya mål med mer traditionella kostnads- och energieffektivitetsmål. Därför är det digitala reningsverket troligen här för att stanna.

De flesta reningsverk samlar redan idag in stora mängder data. System, rutiner och kunskap för att strategiskt hantera dessa data är dock ofta begränsade. En av de bidragande orsakerna är att det ofta saknas en systematisk beskrivning av givarsignaler, inklusive kvalitetsbedömning av deras noggrannhet och tillförlitlighet. Sådana metadata är nyckeln till att omvandla lösryckta dataströmmar till data som är lämpliga för ett specifikt syfte. Sådana metadata kan också vara ovärderliga när man går igenom volymer av historiska data för att försöka identifiera vissa processförhållanden. Anteckningar i operatörens loggbok är ofta användbara som källa till metadata, men de finns ofta inte tillgängliga i digital form.

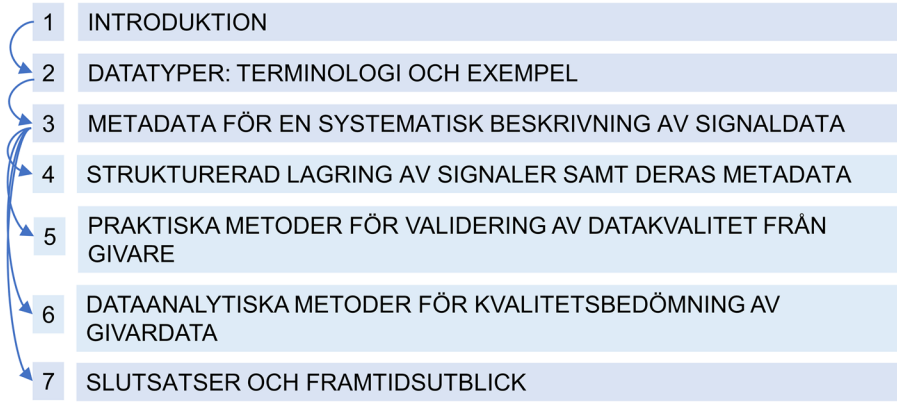
Målet med denna rapport är att klargöra vad metadata är och hur de kan vara en hjälp för att tackla de nämnda utmaningarna. Rapporten ger råd om praktiska implementering av metadatarutiner på reningsverk, som i förlängningen förväntas leda till bättre beslut och nyttjande av det stora antalet givarsignaler som finns tillgängliga på ett reningsverk.

1.3 LÄSANVISNINGAR

De flesta kapitlen i denna rapport kan läsas som fristående kapitel. Ett undantag är Kapitel 2, som innehåller definitioner av signaldata som är relevanta för alla andra kapitel. Figur 1.1 visar samband mellan de olika kapitlen och kan användas som guide för att välja ut de kapitel som är känns relevanta. Tabell 1.1 beskriver även en föreslagen läsordning utifrån roll och arbetsuppgift. Kapitlen 2 och 3 rekommenderas för samtliga läsare, medan Kapitel 4, 5 och 6 är mer detaljerade och stundtals nördiga.

Tabell 1.1 Förslag på vilka kapitel av intresse utifrån roll och arbetsuppgifter.

Kapitel	Kapiteltitel	Instrument-tekniker	Process-ingenjör	Specialist på Datahantering	Analytiker Modellerare	Driftchef
2	Datatyper: terminologi och exempel	X	X	X	X	X
3	Metadata för en systematisk beskrivning av signaldata	X	X	X	X	X
4	Strukturerad lagring av signaler samt deras metadata			X		
5	Praktiska metoder för validering av datakvalitet från givare	X	X		X	
6	Dataanalytiska metoder för kvalitetsbedömning av givardata		X	X	X	
7	Slutsatser och framtidsutsikter				X	X



Figur 1.1 Samband mellan de olika kapitlen.

1.4 TACKORD OCH AVSLUTANDE KOMMENTARER

Vi hoppas att denna rapport kommer vara en bra utgångspunkt för att förbättra datahanteringen på avloppsreningsverk. Rapporten hade inte varit möjlig att skriva utan hjälp från både akademi och personal från reningsverk. Så ett stort tack alla er som bidrog till att denna omfattande rapport om metadata på avloppsreningsverk kunde slutföras:

Åmand, Linda[1]; Amerlink, Youri[2]; Andersson, Sofia[3]; Arteaga, Francisco[4]; Bach, Peter[5]; Baeza, Juan[6]; Belia, Lina[7]; Bellandi, Giacomo[8]; Blumensaat, Frank[5]; Bott, Charles[9]; Brown, Elizabeth[10]; Copp, John[7]; Corominas, Lluís[11,12]; Corona, Francesco[13]; De Gussem, Kris[14]; Feiyi, Li[15]; Ferrer, Alberto[16]; Filali, Ahlem[17]; Fletcher, Laura[10]; Gagnon, Ali[9]; Goldman, Joshua[18]; Grievson, Oliver[19,20]; Gruber, Günter[21]; Grundestam, Jonas[1]; Gubser, Nina[22]; Haimi, Henri[23,24]; Hauduc, Helene[25]; Heffernan, Conor[26]; Jeppsson, Ulf[27]; Juahaianen, Kai[28]; Juan Garcia, Pau[19]; Juszkiewicz, Agnieszka[29]; Kazemi, Pezhman[30]; Kroll, Stefan[31]; Ledung, Erika[32]; Lumley, Doug[33,28]; Massiri, Tofik[1]; Mazeghrane, Sofiane[34]; Menniti, Adrienne[35]; Miletic, Ivan[36]; Mulas, Michela[37]; Nicolaï, Niels[15]; Nilsson, Josefine[38]; Nilsson, Sara[3]; Norling, Katrin[39]; Pahlsson, Anders[3]; Pena-Tijerina, Ana[40]; Persson, Gabriel[3]; Pierce, Alexander[41]; Quan, H Le[42]; Reynolds, Laurie[43]; Rieger, Leiv[44]; Robles, Angel[45]; Rappmann,Tove[28]; Russo, Stefania[46]; Santos, Jorge[47]; Sehlén, Robert[38]; Seib, Matt[48]; Sivchenko, Nataly[49]; Smith, Robert[50]; Sprouse, George[10]; Steyer, Jean-Philippe[51]; Therrien, Jean-David[15]; Thompson, Ken[44]; Tik, Sovanna[15]; Torfs, Elena[15]; Van Daele, Timothy[52]; Van Hoey, Stijn[53]; Vanrolleghem, Peter A. Vargas[15], Jose Leornardo$_{54}$; Volcke, Eveline I. P.[42]; Wade, Matthew[55,56]; Weijers, Stefan[57]; Yanchen, Liu[58]; Yang, Fenghua[59]; Ye, Liu[60]; Zambrano, Jesus[61]; Zegers, Emily[62].

[1]Käppala Association; [2]Aquafin; [3]Stockholm vatten och avfall; [4]Universidad Católica de Valéncia; [5]Eawag Swiss Federal Institute of Aquatic science and technology; [6]Universitat Autònoma de Barcelona; [7]Primodal, Inc; [8]AM Team advanced modelling for process optimisation; [9]Hampton Roads Sanitation District (HRSD); [10]Metropolitan Council of the Twin Cities; [11]Catalan Institute for Water Research (ICRA-CERCA); [12]University of Girona; [13]School of Chemical Engineering, Aalto University; [14]Former Aquafin; [15]modelEAU, Université Laval; [16]Multivariate Statistical Engineering Research Group, Universitat Politècnica de València; [17]Irstea, National Research Institute of Science and Technology for Agriculture and Environment; [18]Metro Water Recovery; [19]AtkinsRéalis, Birmingham, Storbritannien; [20]Visiting Professor vid University of Exeter; [21]TU Graz, Institute of Urban Water Management and Landscape Water Engineering; [22]City of zurich, Zuerich; [23]FCG Finnish Consulting Group Ltd; [24]Högskolan för teknik, Aalto-universitetet; [25]Dynamita; [26]Metropolitan Water Reclamation District of Greater Chicago, Chicago, IL, USA; [27]IEA, Lunds universitet; [28]Gryaab

AB; [29]Mälarenergi AB;[30]Departament d'Enginyeria Química, Universitat Rovira i Virgili; [31]R&D, Aquafin NV; [32]Esem; [33]Formerly DHI Group; [34]Suez; [35]Clean Water Services; [36]inCTRL Solutions Inc., Dundas; [37]Department of Teleinformatics Engineering, Federal University of Ceará; [38]Tekniska verken; [39]Mjölby kommun; [40]Plummer Associates, Inc; [41]Fort Worth, Texas; [42]BioCo Research Group, Department of Green Chemistry and Technology, Ghent University; [43]Aquamatix; [44]Jacobs Engineering Group Incorporation; [45]CALAGUA - Unidad Mixta UV-UPV, Departament d'Enginyeria Química, Universitat de València; [46]EcoVision Lab, Photogrammetry and Remote Sensing, ETH Zurich; [47]DMI, Nova Universtity Lisbon; [48]Madison Metropolitan Sewerage District; [49]DOSCON AS; [50]Black & Veatch; [51]INRAE, LBE, Univ Montpellier, Narbonne; [52]The Janssen Pharmaceutical Companies of Johnson & Johnson; [53]Fluves; [54]ABB; [55]UK Health Security Agency, Data Analytics & Surveillance, Canary Wharf; [56]School of Engineering, Newcastle University, Cassie Building; [57]Waterschap De Dommel; [58]School of Environment, Tsinghua University; [59]MMSD – Milwaukee Metropolitan Sewerage District; [60]University of Queensland, School of chemical Engineering; [61]The MathWorks AB; [62]City of Toronto, Toronto.

Denna rapport är en Open Access-publikation tack vare finansiering från följande organisationer:

- SIVL-stiftelsen för IVL (SIVL)
- Svenskt vatten utvecklling, Svenskt vatten
- Stockholm vatten och avfall AB
- Mälarenergi vatten AB
- Eskilstuna Strängnäs energi och miljö AB
- Mjölby kommun

Vi hoppas att ni kommer att ha nytta av denna rapport och kommer tycka att den är lika intressant som vi tyckte när vi skrev den. Alla förbättringsförslag och synpunkter välkomnas.

doi: 10.2166/9781789065107_0007

Kapitel 2

Datatyper: terminologi och exempel

Q. Plana[1,2] och K. Villez[3]

[1]*modelEAU, Université Laval, Québec, QC, Kanada*
[2]*Eurecat – Kataloniens teknikcentrum, Barcelona, Spanien*
[3]*Oak Ridge National Laboratory, Oak Ridge, TN, USA*

Nyckelord: bakgrundsinformation, datavariation, definitioner, dataheterogenitet, metadata, signal

Data som produceras på reningsverk erhålls genom en mängd olika enheter, inklusive ställdon, styrenheter och sensorer. Detta resulterar i data som kan vara mycket varierande i sin struktur. Att hantera denna blandning och heterogeniteten i dataformat kan vara en utmaning när data ska lagras eller tolkas. Av denna anledning beskriver detta kapitel de strukturella aspekterna av data på ett vanligt reningsverk. Syftet med detta kapitel är att:

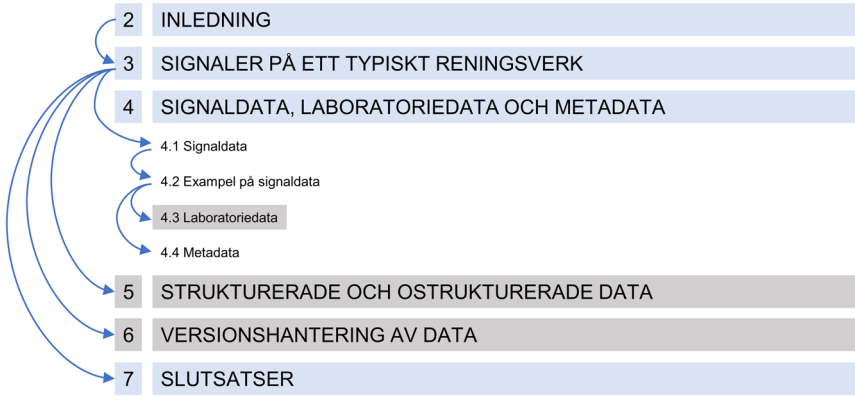
- Introducera grundläggande begrepp för att beskriva, förstå och hantera data som produceras av onlineinstrument såsom givare och ställdon (t.ex. ventiler och pumpar).
- Definiera de vanligaste termerna som används i rapporten och som rör givare och andra datakällor.
- Illustrerar definitionerna med praktiska exempel.

Där så är möjligt har vi nyttjat befintliga standarder och referenser för definitionerna även om flertalet definitioner har tagits fram specifikt för denna rapport.

2.1 LÄSANVISNINGAR

Huvudsyftet med detta kapitel är att definiera de begrepp som är nödvändiga för att förstå resten av denna rapport. Schemat nedan anger beroenden inom detta kapitel och kan användas för välja vilka avsnitt som man vill läsa. Blå- och vitmarkerade rubriker är grundläggande för att förstå innehållet i återstående kapitel. De delar som är gråtonade beskriver detaljer inom databasteknik och hantering av informationsflöden.

© 2025 IWAP. Detta är ett Open Access-bokkapitel som distribueras enligt villkoren i Creative Commons Attribution License (CC BY-NC-ND 4.0) som tillåter kopiering och vidaredistribution för icke-kommersiella ändamål utan derivat, förutsatt att arbetet citeras korrekt (https://creativecommons.org/licenses/by-nc-nd/4.0/). Kapitlet är från boken *Bästa möjliga data från givare i avloppsreningsprocesser*, Kris Villez, Daniel Aguado, Janelcy Alferes, Queralt Plana, Maria Victoria Ruano, Oscar Samuelsson (redaktörer).

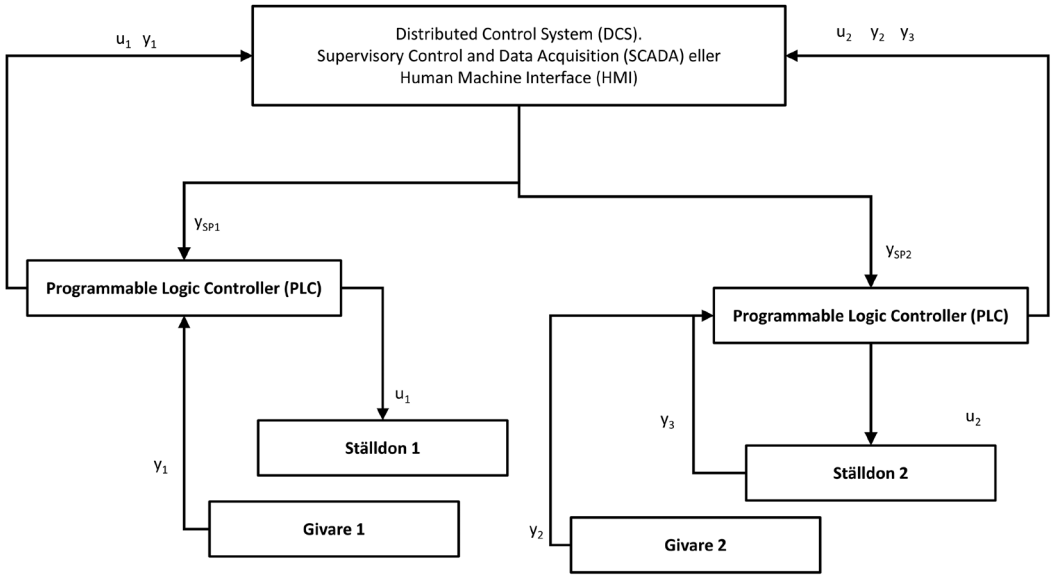


2.2 INLEDNING

Detta kapitel rekommenderas som inledningen till rapporten och bör läsas i sin helhet inför övriga kapitel. Kapitlet kan även nyttjas som referenslista för olika begrepp genom att söka i listan över definitioner (se avsnittet Ordlista). Observera att alla exempel i detta kapitel visas utan några metadata. Detta innebär att data visas utan sammanhang och därmed utan betydelse. Metadata som ger dessa data ett sammanhang kommer att diskuteras i Kapitel 3.

2.3 SIGNALER PÅ ETT TYPISKT RENINGSVERK

Figur 2.1 visar en schematisk bild av hårdvarukomponenterna i ett litet styrsystem på ett vanligt reningsverk. Styrsystemet innehåller två reglerkretsar, var och en implementerad i en programmerbar



Figur 2.1 Grafisk framställning av ett litet styrsystem. Ett styrsystem med rimlig komplexitet innehåller mätningar (y), manöver- eller styrsignaler (u) och börvärden (y_{SP}), som generar signaler.

logisk styrenhet (PLC). Den första styrenheten gör en mätning, y_1, från givare 1. Vanliga exempel på givare är mätningar av flöde, tryck och koncentrationer. Ett börvärde för y_1, benämns med $y_{1,SP}$, skickas från det överordnade styrsystemet *(Supervisory Control and Data Acquisition system – SCADA*, eller *Distributed Control System – DCS*) till styrenheten som sedan tolkar mätningen och börvärdet och genererar en styrsignal u_1. Denna styrsignal skickas till en manöverenhet, ställdon 1. Vanliga exempel på ställdon är ventiler och frekvensomriktare (VFD). Den andra styrenheten beter sig på liknande sätt med en sensorsignal y_2, ett börvärde $y_{2,SP}$ och en manöversignal u_2. Dessutom skickar ställdonet också en signal, y_3, tillbaka till styrenheten. Exempel på sådana signaler kan vara en positions- eller vridmomentmätning i en ventil eller en temperatur-, tryck- eller hastighetsmätning i en pump. Ställdon som producerar mätsignaler kallas ibland för smarta enheter (t.ex. smart ventil, smart pump). Styrenheterna rapporterar alla signaler till det överordnade styrsystemet. Sammanfattningsvis så kan alltså samtliga komponenter i detta reglersystem, bortsett från ställdon 1 producera signaler som samlas in och tolkas av andra enheter inom samma arkitektur.

I den här rapporten betraktas alla styrsignaler, börvärden och mätningar som skickas mellan hårdvarukomponenter som signaler. Observera att detta även inkluderar börvärden som matas in manuellt av en operatör och som vidarebefordras till andra komponenter i styrsystemet. Begreppet *signalenhet* används för alla hårdvarukomponenter i ett sådant system där alla hårdvarukomponenter som producerar en signal, oavsett om denna signal är en mätning, ett börvärde eller en aktivering benämns som *instrument*.

2.4 SIGNALDATA, LABORATORIEDATA OCH METADATA

I det här avsnittet beskriver vi den grundläggande terminologin kring givarsignaler, laboratoriedata och metadata. Detta omfattar definitioner och allmänna anmärkningar i ett första underavsnitt, följt av exempel som visar hur definitionerna tillämpas i praktiken.

2.4.1 Signaldata

I de följande avsnitten ges (a) definitioner, (b) en grafisk representation av datapunkter, (c) kommentarer om utvärderingen av datadimensionalitet och (d) ytterligare kommentarer.

2.4.1.1 Datapunktens dimensionalitet

De vanligaste instrumenten producerar datapunkter som skalärer, det vill säga de består av ett enda värde (t.ex. ett värde för konduktivitet). Sådana datapunkter beskrivs som skalärer eller nolldimensionella (0D) värden. Den resulterande datapunkten består därför av en tidsstämpel (tidpunkten för mätvärdet) och ett enda värde (se Avsnitt 2.4.1.2 för definitioner). Om man kombinerar datapunkter som samlats in vid flera olika tidpunkter får man en signal som betraktas som en tidsserie (tidsserien kallas univariat om den enbart innehåller ett mätvärde vid varje tidpunkt).

Det finns mer avancerade instrument som producerar datapunkter som innehåller fler än ett mätvärde per tidpunkt, t.ex. spektrum som produceras av en spektrofotometrisk sensor består av flera mätningar av ljusabsorption, där varje mätning är associerad med en enda våglängd. Detta är exempel på endimensionella (1D) datapunkter. De motsvarande signalerna kallas ofta multivariata tidsserier.

Det finns också flerdimensionella datapunkter. Exempelvis betraktas gråskalebilder som tvådimensionella (2D) medan färgbilder och hyperspektrala bilder betraktas som tredimensionella (3D). Tomografiska instrument finns för att mäta ett objekts färgintensitet i ett fysiskt 3D-utrymme. Även om en enda tomografisk bild uppvisar fyra dimensioner (t.ex. bredd, höjd, djup och våglängd) kallas den fortfarande för en enda datapunkt i den här rapporten.

2.4.1.2 Definitioner

Definition D2.1: **Signaldata** (även: signaler) består av värden som produceras av en givare för kvaliteter och kvantiteter i ett fysiskt system av intresse. Signaldata består av serier av dataposter.

Definition D2.2: **En datapost** (‘*data record*’) är ett par eller en triplett av (i) en tidsstämpel (obligatorisk), (ii) en datapunkt (obligatorisk) och (iii) en datakvalitetsstämpel (‘*qualifier*’) (valfri).

Definition D2.3: **En datapunkt** består av ett eller flera värden som uppfyller följande tre krav:

(1) Alla värden inom datapunkten har producerats av samma instrument.
(2) Alla värden är kopplade till samma tidpunkt eller, i förekommande fall, samma tidpunkt för insamling av vattenprover.
(3) Alla värden beskriver samma typ av variabel och produceras med samma mätprincip.

Definition D2.4: **En datakvalitetsstämpel** är ett enskilt tal eller en text som produceras av ett instrument och som anger en nivå av förtroende för tidsstämpeln och/eller datapunkten.

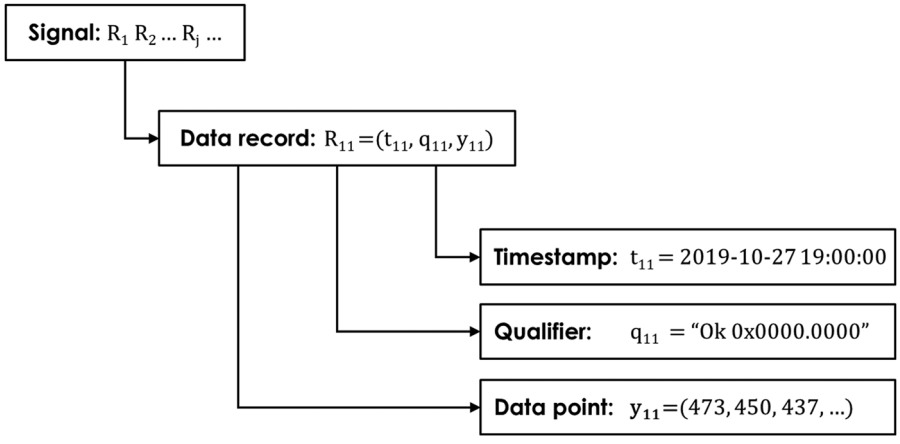
En datakvalitetsstämpel kan produceras av ett instrument för att indikera för dataanvändaren att data saknas, är avvikande, felaktiga eller misstrodda av någon anledning. Vi rekommenderar att man följer befintliga standarder för att lagra datakvalitetsstämplar. OPC Foundation (2003) specificerar t.ex. en 8-bitars beskrivning för att ange datakvalitet som en del av OPC Data Access-standarden, som har brett stöd av kommersiella DCS-, SCADA- och HMI-system.

Ovanstående definitioner illustreras i Figur 2.2. En signal består av dataposter, som var och en består av en tidsstämpel, en datapunkt och en datakvalitetsstämpel. Där så krävs i denna rapport kommer vi att tillämpa följande symboliska representation av samma koncept för signaler med 0D-värden:

$$R_j = (t_j, y_j, q_j)$$

där R_j är den *j:te* dataposten i en signal, t_j är tidsstämpeln i den *j:te* dataposten och y_j är värdet för den uppmätta variabeln i den *j:te* dataposten. När datapunkterna är 1D, 2D eller högdimensionella understryker vi texten och/eller använder ett fetstilt typsnitt i enlighet med följande konventioner:

- 1D (skalär): $R_j = (t_j, y_j, q_j)$
- 2D (matris): $R_j = (t_j, Y_j, q_j)$
- 3D och högre (tensor): $R_j = (t_j, \underline{Y}_j, q_j)$



Figur 2.2 Grafisk framställning av definitionerna D2.1, D2.2, D2.3 och D2.4. De angivna värdena motsvarar den 11:e dataposten i Tabell 2.2.

2.4.1.3 Dataförluster

Bortfall och förlorade datapunkter är ett vanligt problem i praktiken. Det finns många möjliga orsaker till detta såsom (a) inkomplett teknisk rutin för datainsamling, (b) mätvärden som överskrider instrumentets mätområde, (c) kommunikationsavbrott eller (d) felaktig samplingsfrekvens (hastighet för datainsamling). I denna rapport skiljer vi på frånvarande data och förlorade data utifrån följande definitioner.

Definition D2.5: En datapost, tidsstämpel eller datapunkt anses vara **frånvarande** (*absent data*) om tillgång till dessa data är värdefull vid tidpunkten för dataanalysen, men inget instrument har installerats för att samla dessa data.

Definition D2.6: En datapost, tidsstämpel eller datapunkt anses vara **förlorad** (*missing data*) om ett instrument installerats för att samla in data, men misslyckats med att producera förväntade mätvärden.

Den största skillnaden mellan frånvarande data och förlorade data är avsikten med datainsamlingen. Om man hade för avsikt att samla in information, men misslyckades, betraktas informationen som förlorad. Om det inte gjordes något försök att samla in informationen och nyttan av denna information blir uppenbar först senare, betraktas den istället som frånvarande.

Det finns flera skäl till att lagra eller arkivera information om förlorade data, det vill säga (a) att inte utelämna information om avsikten att samla in uppgifter och (b) att inte ersätta förlorade data med ersättningsvärden (*imputed data*) för långtidslagring. Lagring av information om frånvarande och förlorade data bör istället lagras på ett robust men ändå tillgängligt sätt. Exemplen 2.1, 2.2 och 2.3 nedan illustrerar frånvarande och förlorade data. Vetskapen om att det finns förlorade data kan vara värdefull i sig. Detta kan t.ex. inträffa när värdet för den uppmätta variabeln ligger utanför mätområdet. I detta fall kan det vara bra att veta att det uppmätta värdet låg under eller över en viss gräns för att uppskatta och ersätta det senare. Detta är ett exempel på ett MNAR-mönster (*missing-not-at-random*) (t.ex. Dong & Peng, 2013; Kang, 2013; Osman *et al.*, 2018). Däremot kommer data som förlorats på grund av ett avbrott i strömförsörjningen till ett instrument i allmänhet inte att korrelera med värdet på den uppmätta variabeln. Detta benämns då som ett MAR-mönster (*missing-at-random*) eller MCAR-mönster *(missing-completely-at-random*) där olika ersättningsmetoder för dessa fel föreslås t.ex. i Bhaskaran & Smeeth, 2014).

För att möjliggöra optimal diagnostik och efterföljande ersättning av förlorade data rekommenderas att olika indikatorer används vid dataförluster och långtidslagring. Detta kräver dock att alla delar av den digitala infrastrukturen som påverkar datainsamling, kommunikation och lagring utformas för att acceptera värden i form av text eller en specialiserad kod. Många moderna instrument, inklusive givare och ställdon, erbjuder automatisk generering av metadata, som ofta beskriver grundorsakerna till att data saknas. Insamling och lagring av sådana metadata kan användas för att underlätta både datahantering och givarhantering.

2.4.1.4 Hur lagrar man information som beskriver förlorade data?

Med tanke på att frånvarande data motsvarar förhållanden där det inte fanns någon avsikt att samla in data finns det inget särskilt behov av att lagra information om frånvarande data. Förutsatt att information om förlorade data lagras på ett bra sätt kan man kategorisera data som inte är tillgängliga som frånvarande om ingen motsvarande information om förlorade data finns tillgänglig. Förlorade data bör däremot dokumenteras för att visa att data inte har registrerats när detta var avsikten.

Ett sämre sätt att beskriva förlorade data: Ett vanligt sätt att beskriva förlorade data är att använda en specialiserad numerisk kod som anger en del av den information som saknas. Till exempel så används ett negativt tal på ett instrument som normalt ger positiva mätvärden såsom en koncentration. Detta kan vara användbart när en enda analog anslutning används för kommunikation. Det bör dock undvikas om det är möjligt av två skäl: (1) det är mycket sannolikt att ett numeriskt värde misstolkas som ett verkligt mätvärde, särskilt om dataanvändaren inte är bekant med tillämpningen, och (2) konsekvenserna av att använda konventionella databehandlingssteg och därmed ignorera saknade

värden blir sannolikt synliga först i de sena stadierna av projekt som är beroende av data, vilket i sin tur leder till förseningar, fel och oförutsedda kostnader. I vilket fall som helst kan den potentiella förekomsten av specialiserade värden i dataset som indikerar att data saknas bör diskuteras tidigt i ett projekt för att undvika uppenbara fel eller övertro på databaserade resultat.

Ett bra sätt att beskriva förlorade data: Om det är möjligt bör en datapunkt som saknas representeras med en specifik icke-numerisk kod, t.ex. *not-a-number*, *NA*, *NaN*, *below-detection-limit* och *out-of-range*. Man kan lägga till en datakvalitetsstämpel (se ovan) eller ett ytterligare värde i datapunkten för att göra detta (t.ex. 0 för ett accepterat värde och 1 för ett avvisat värde). Detta är det rekommenderade tillvägagångssättet eftersom det är mycket svårare att missta den här typen av kod för ett accepterat numeriskt värde eftersom beräknade resultat ger samma textbeskrivning och inte fortplantar sig genom beräkningar som ger användbar information. Detta är kommersiellt tillgängligt i de flesta DCS- och SCADA-system och ingår i OPC-standarden (Open Platform Communications).

Ett annat bra sätt att beskriva förlorade data: Ett annat sätt som underlättar en entydig beskrivning av förlorade data i en signal (signal A) är att lagra en separat signal (signal B). Man kan t.ex. lagra separata par som består av en tidsstämpel och en datapunkt, där datapunkten är en numerisk indikator som beskriver förlorad datapunkt. En nollvärderad datapunkt i signal B kan t.ex. ange att motsvarande datapunkt i signal A saknas, medan ett värde lika med 1 kan användas för att ange att motsvarande datapunkt i signal A samlades in på avsett sätt. Detta koncept ingår redan i OPC Data Access-standarderna (OPC Foundation, 2003) och stöds av de flesta kommersiellt tillgängliga styrsystem, inklusive DCS- och SCADA-system.

En kommentar om lagring av frånvarande data: Eftersom frånvarande data är data som aldrig var avsedda att samlas in, finns det förstås inget behov av att lagra information om detta. Dessutom kan det vara opraktiskt att definiera en unik tidsstämpel för dataposter som aldrig var avsedda att samlas in. Detta förutsätter dock att förlorade data registreras på rätt sätt så att det senare är lätt att skilja på frånvarande och förlorade data.

En kommentar om lagring av ersättningsvärden: Många metoder som används inom dataanalys kan inte hantera frånvarande eller förlorade data som standard. Detta hanteras vanligen istället genom att uppskatta och ersätta de förlorade värdena (se Kapitel 6, Avsnitt 6.6.2). Processen för att lägga till information för att ta hänsyn till frånvarande och förlorade data kallas *data imputation* på engelska. Ett dataset med ursprunglig rådata utan några tillägg eller ändringar bör alltid finnas tillgänglig. Det främsta skälet till detta är att dataersättningsmetodiken kan förändras över tid. Dessutom används en ersättningsmetod oftast för ett visst projekt eller en viss uppgift som kanske inte är generellt tillämpbar. För att kunna ersätta data på ett effektivt sätt bör dataset där frånvarande eller förlorade data har ersatts innehålla metadata som tydligt identifierar de ersatta värdena och själva ersättningsmetodiken. Detta diskuteras mer i detalj nedan i avsnittet om versionshantering av data (Avsnitt 6).

2.4.1.5 Ytterligare kommentarer

De numeriska värdenas ursprung: Värdena i en datapunkt kan vara kvantitativa (t.ex. en koncentration, en ventilposition), kvalitativa (t.ex. en pumps på/av-läge) eller indikera att information saknas (*t.ex. inte ett tal*, *utanför intervallet*, *under detektionsgränsen*). Signaldata kan produceras av vilket instrument som helst, inklusive ställdon (t.ex. en ventilmomentmätning), sensorer (t.ex. en sensor för upplöst syre) eller styrenheter (t.ex. en PLC).

Instrument med flera variabler eller multigivare: Många instrument kan ge flera olika mätningar samtidigt. Till exempel så mäter en syregivare både temperatur och syrehalt. I detta fall behandlas temperatursignalen och den upplösta syrgasen som separata signaler, som var och en är univariata tidsserier och innehåller skalära data (0D).

Anledningen till detta är att de metadata som beskriver dessa två signaler, även om de produceras av samma instrument, är tillräckligt olika för att motivera separat datahantering och lagring. Notera att detta är det huvudsakliga motivet för det tredje villkoret i definition D2.3.

2.4.2 Exempel på signaldata

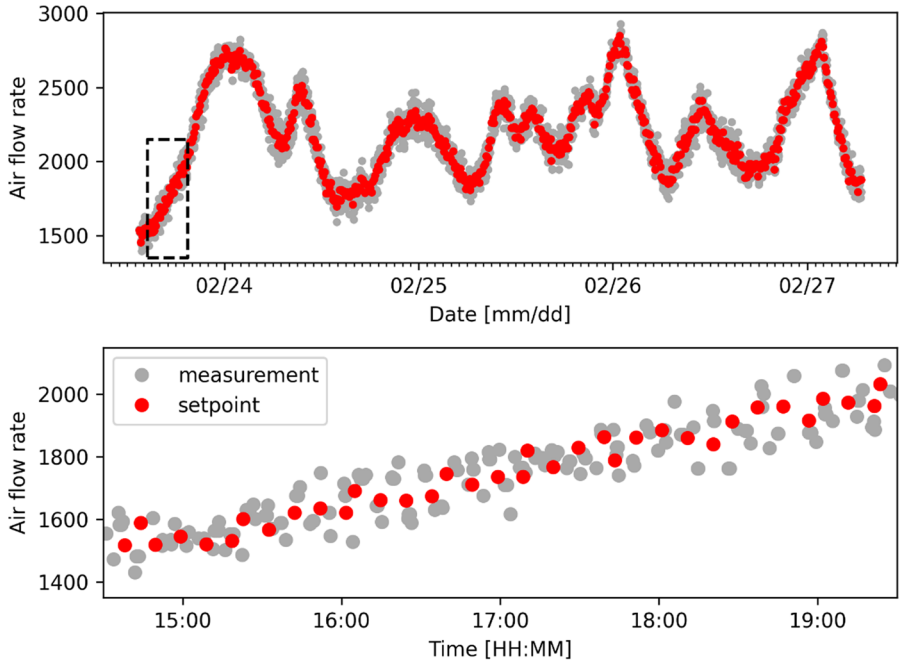
För att visa hur definitionerna ovan kan tillämpas ges fyra exempel. Det första och andra exemplet består av ett börvärde som regleras av en styrenhet och en motsvarande styrsignal. Det tredje och fjärde exemplet består av givarsignaler som har en mer komplex form.

2.4.2.1 Exempel E2.1: börvärde för luftflödesstyrning

Det första exemplet på en enkel signal är en börvärdessignal för luftflödet. En börvärdessignal visas i den övre figuren i Figur 2.3 som röda punkter. Börvärdessignalen består av dataposter med en tidsstämpel och ett börvärde. Några av dessa punkter visas i de två kolumnerna på vänster sida i Tabell 2.1. Denna typ av signal är en univariat tidsserie som är den vanligaste typen av tidsserie.

2.4.2.2 Exempel E2.2: mätning av luftflöde

Det andra exemplet på en enkel signal är en luftflödesmätning som visas i den övre figuren i Figur 2.3 som grå punkter. I det här fallet följer mätvärdet börvärdessignalen väldigt bra. Givarsignalen, liksom börvärdessignalen i denna figur, består av dataposter med en tidsstämpel och ett mätvärde. Några av dessa punkter visas i de två kolumnerna på högra sidan i Tabell 2.1. Återigen är denna typ av signal en univariat tidsserie. Den nedre figuren i Figur 2.3 illustrerar varför börvärdet och mätsignalen behandlas som olika signaler. Tidsstämplarna för de två signalerna överensstämmer inte eftersom dataposterna genereras på ett asynkront sätt. Notera också hur hastigheten för generering av dataposter är olika för de två signalerna och förändras över tiden. Dataposterna i Tabell 2.1 är



Figur 2.3 Exemplen E2.1 och E2.2: två univariata signaler. Ett börvärde (rött) för ett luftflöde och en motsvarande mätning (grå). Översta panelen: Tidsfönster på fyra dagar. Nedre panelen: En inzoomad vy över de dataposter som matchar den streckade rutan i den övre panelen. Observera att (a) dataposterna registreras på ett asynkront sätt och att (b) hastigheten för generering av dataposter skiljer mellan de två signalerna. Försök att synkronisera denna typ av data bör dokumenteras och implementeras på ett reversibelt sätt.

Tabell 2.1 Exempel E2.1 och E2.2: två univariata signaler.

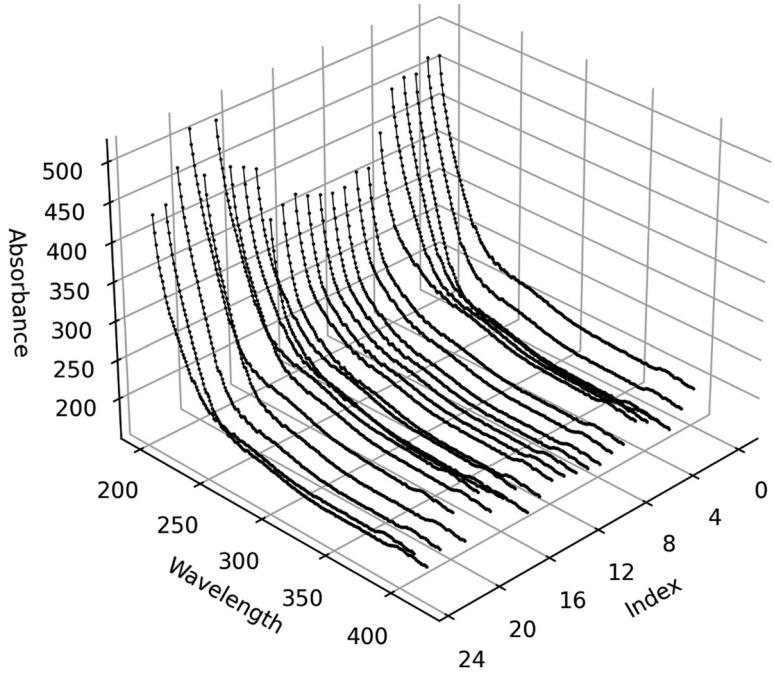
Univariat signal A (E2.1)		Univariat signal B (E2.2)	
Tidsstämpel	**Luftflödesbörvärde**	**Tidsstämpel**	**Mätning av Luftflöde**
t	***y***	***t***	***y***
n/a	n/a	2020-02-23 14:31:01	1555,17
n/a	n/a	2020-02-23 14:33:52	1473,6
n/a	n/a	2020-02-23 14:35:44	1621,91
n/a	n/a	2020-02-23 14:36:09	1583,18
n/a	n/a	2020-02-23 14:36:48	1583,18
n/a	n/a	2020-02-23 14:37:20	1593,9
2020-02-23 14:38:05	1517,25	n/a	n/a
n/a	n/a	2020-02-23 14:42:00	1430,47
n/a	n/a	2020-02-23 14:42:32	1482,39
n/a	n/a	2020-02-23 14:43:19	1482,39
2020-02-23 14:44:05	1589,78	n/a	n/a
n/a	n/a	2020-02-23 14:48:41	1604,06
n/a	n/a	2020-02-23 14:48:48	1604,06
2020-02-23 14:49:33	1519,64	n/a	n/a
n/a	n/a	2020-02-23 14:52:17	1516,99
n/a	n/a	2020-02-23 14:52:23	1534,57
n/a	n/a	2020-02-23 14:55:30	1534,57
n/a	n/a	2020-02-23 14:55:43	1521,66
n/a	n/a	2020-02-23 14:56:56	1585,66
2020-02-23 14:59:14	1546,29	n/a	n/a

Den första signalen är en styrsignal och innehåller dataposter som består av en tidsstämpel som genereras av styrenheten och ett börvärde för ett luftflöde. Den andra signalen är en givarsignal och innehåller dataposter som består av en tidsstämpel som genereras av givaren och det uppmätta värdet för luftflödet. De två signalerna (a) registreras asynkront, (b) registreras med olika frekvenser och (c) uppvisar tidsintervall som varierar över tiden. Observera att datauppsättningen är vanligt förekommande i praktiken. Det finns inga förlorade data här, dvs. mättillgängligheten är som tänkt för denna applikation. På grund av bristen på synkronisering saknas dock data eftersom det inte finns några poster i de två signalerna med matchande tidsstämplar. Dessa indikeras som *n/a* (‘*not applicable*’) vilket betyder ej tillgänglig.

organiserade så att varje rad motsvarar en unik tidsstämpel, antingen för börvärdessignalen eller för givarsignalen. Ingen speciell symbol används i tabellens celler där data saknas. Detta beror på att informationen i dessa celler anses vara frånvarande. Det finns faktiskt inga förlorade data här eftersom datainsamlingssystemet fungerade som avsett (Definition D.4).

2.4.2.3 Exempel E2.3: absorbansspektrofotometri med ultraviolett ljus

Spektrofotometriska sensorer används vanligen för att övervaka en eller flera vattenkvalitetsvariabler. De omfattar t.ex. totalt suspenderat material (TSS), kemisk syreförbrukning (COD), nitrit och nitrat. Det är viktigt att notera att de flesta spektrofotometriska sensorer är utrustade med mjukvarusensorer som översätter ljusabsorptionsmätningarna till värden för målvariablerna (se t.ex. Lepot *et al.*, 2016; Mašić *et al.*, 2015; Rieger *et al.*, 2006). De producerade dataposterna består av en tidsstämpel och en datapunkt, där datapunkten motsvarar ett spektrum. Varje spektrum, som ibland kallas ett fingeravtryck (‘*fingerprint*’), består av flera absorbansvärden, där varje absorbansvärde motsvarar en våglängd i det synliga ultravioletta mätområdet. Figur 2.4 visar 24 dataposter som produceras av en spektrofotometer. Varje rad motsvarar en enda datapost. Antalet absorbansvärden i en enda



Figur 2.4 Exempel E2.3: Spektrofotometriska sensorer levererar 1D data. Varje datapost består av en tidsstämpel (t) och en motsvarande datapunkt, som är en uppsättning absorbansmätningar (y) som tagits för en fast uppsättning ljusvåglängder (λ). Varje datapunkt är 1D (längd: 221) och visas som en prickad linje.

datapunkt är 221. Tabell 2.2 visar de 24 dataposter som visas i Figur 2.4 i tabellform. Tabellen visar tidsstämpeln (t), datakvalitetsstämpeln (q) och de multivariata 1D-datapunkterna (y). Denna tabell innehåller värden som är lika med *nan* (‘*not-a-number*’), vilket indikerar att data saknas (se definition D2.6).

2.4.2.4 Exempel E2.4: flödescytometri

Flödescytometri är en relativt ny mätteknik för övervakning av vattenkvalitet i realtid (se t.ex. Besmer *et al.*, 2014; Wang *et al.*, 2010). I ett typiskt flödescytometriskt instrument analyseras ett vattenprov genom att mäta ljusintensiteten som erhålls genom ljusspridning och fluorescenseffekter som genereras av enskilda partiklar. Detta utförs i två steg. Först färgas partiklarna med en fluorescerande färg. Sedan skickas laserljus med en specifik våglängd (t.ex. blått ljus: 488 nm) till en partikel, vilket omedelbart följs av en mätning av intensiteten hos det ljus som returneras av partikeln genom ljusspridning eller genom emission vid en annan specifik våglängd (t.ex. grönt ljus: 533 nm, rött ljus: 670 nm). Det är möjligt att registrera ljusintensiteten vid en mängd olika våglängder och i olika vinklar. Mätningen av ett enda vattenprov resulterar därför i en flerdimensionell tabell, där raderna motsvarar enskilda partiklar och kolumnerna olika ljusdetektorer. Detta innebär att varje datapost består av en tidsstämpel, som återspeglar tidpunkten för insamlingen av vattenprovet, och en 2D-matris med ljusintensitetsmätningar som är kopplade till denna tidsstämpel. I Figur 2.5 visas värdena för två ljusintensiteter i två dataposter, där varje panel i figuren motsvarar ett enstaka vattenprov.

De två ljusintensiteter som visas här motsvarar antalet spridningar framåt (FSC) och antalet spridningar åt sidan (SSC). Samma dataposter innehåller dock ytterligare mätningar, t.ex.

Tabell 2.2 Exempel E2.3: Spektrofotometriska sensorer levererar strukturerade data som illustreras med denna tabell.

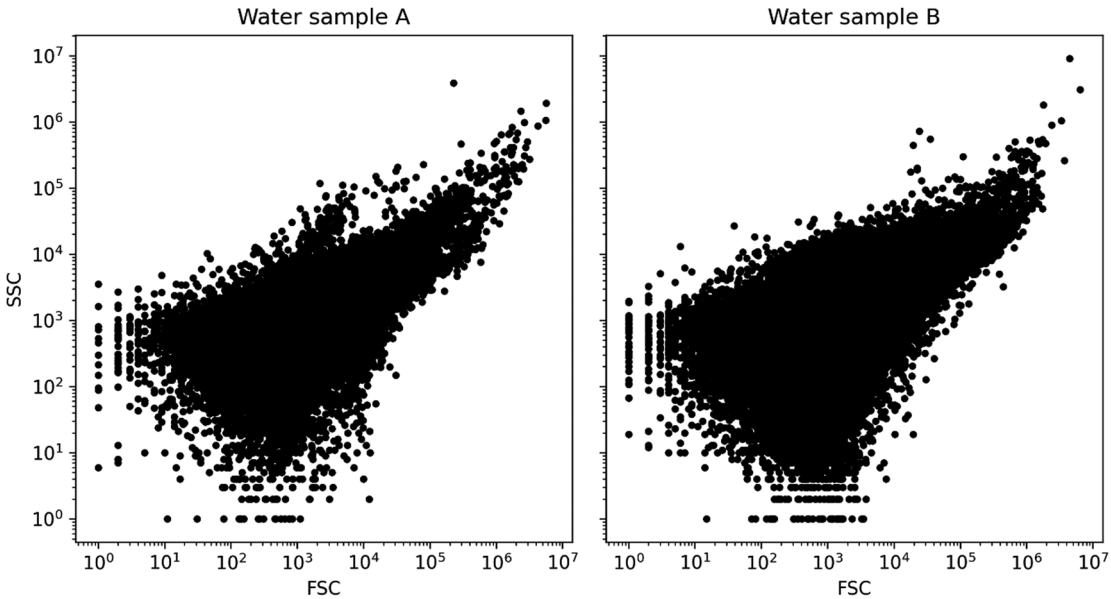
Registerindex	Tid	Kvalitet	Datapunkt							
j	t	q	y							
			0	1	2	...	201	202	203	204
0	2019-10-27 19:00:00	Ok 0 × 0000.0000	473	450	437		197	nan	nan	nan
1	2019-10-27 20:00:00	Ok 0 × 0000.0000	477	453	440		179	nan	nan	nan
2	2019-10-27 21:00:00	Ok 0 × 0000.0000	462	440	426		160	nan	nan	nan
3	2019-10-27 22:00:00	Ok 0 × 0000.0000	465	446	432		179	nan	nan	nan
4	2019-10-27 23:00:00	Ok 0 × 0000.0000	457	435	423		187	nan	nan	nan
5	2019-10-28 00:00:00	Ok 0 × 0000.0000	408	385	370		196	nan	nan	nan
6	2019-10-28 01:00:00	Ok 0 × 0000.0000	369	348	331		174	nan	nan	nan
7	2019-10-28 02:00:00	Ok 0 × 0000.0000	371	350	336		172	nan	nan	nan
8	2019-10-28 03:00:00	Ok 0 × 0000.0000	359	336	324		165	nan	nan	nan
9	2019-10-28 04:00:00	Ok 0 × 0000.0000	359	339	327		169	nan	nan	nan
10	2019-10-28 05:00:00	Ok 0 × 0000.0000	364	345	332		172	nan	nan	nan
11	2019-10-28 06:00:00	Ok 0 × 0000.0000	370	348	334		176	nan	nan	nan
12	2019-10-28 07:00:00	Ok 0 × 0000.0000	379	358	343		179	nan	nan	nan
13	2019-10-28 08:00:00	Ok 0 × 0000.0000	373	350	337		167	nan	nan	nan
14	2019-10-28 09:00:00	Ok 0 × 0000.0000	361	337	323		155	nan	nan	nan
15	2019-10-28 10:00:00	Ok 0 × 0000.0000	433	410	397		195	nan	nan	nan
16	2019-10-28 11:00:00	Ok 0 × 0000.0000	443	422	407		191	nan	nan	nan
17	2019-10-28 12:00:00	Ok 0 × 0000.0000	451	429	415		183	nan	nan	nan
18	2019-10-28 13:00:00	Ok 0 × 0000.0000	516	496	482		217	nan	nan	nan
19	2019-1028 14:00:00	Ok 0 × 0000.0000	455	431	418		163	nan	nan	nan
20	2019-10-28 15:00:00	Ok 0 × 0000.0000	520	501	490		209	nan	nan	nan
21	2019-10-28 16:00:00	Ok 0 × 0000.0000	479	459	445		171	nan	nan	nan
22	2019-10-28 17:00:00	Ok 0 × 0000.0000	440	419	406		157	nan	nan	nan
23	2019-10-28 18:00:00	Ok 0 × 0000.0000	435	414	399		183	nan	nan	nan

Varje datapunkt är 1D. Varje datapost består av en tid (*t*), en kvalifikator (*q*) och en datapunkt, som är en serie absorbansmätningar (*y*) som gjorts för en fast uppsättning ljusvåglängder. Kolumnpositionen för varje absorbansvärde i tabellen motsvarar en enda våglängd medan radpositionen motsvarar en unik mättidpunkt. Observera att endast en delmängd av absorbansvärdena visas och att saknade värden (nan) förekommer.

fluorescerande ljus som avges av en grön färg (fluoresceinisotiocyanat, FITC) och en röd färg (peridinin-klorofyllprotein, PerCP), som båda exciteras med hjälp av blått laserljus. Mer information om de tillämpade mätprinciperna och det praktiska genomförandet ges i exempelvis (Besmer *et al.*, 2014; Hammes & Egli, 2010).

2.4.3 Laboratoriedata

Även om denna rapport främst fokuserar på givardata, ges här en definition av laboratoriedata för att kunna särskilja sådana data från givardata.



Figur 2.5 Exempel E2.4: två dataposter från en flödescytometrisignal. Tidsstämpeln visas överst på figurpanelerna medan mätningarna för två ljusintensiteter visas som en funktion av varandra. De uppmätta värdena är FSC och SSC. Vänster panel: En datapost med mätvärden för 56 358 partiklar som erhölls den 9 maj 2019. Höger panel: En datapost med mätvärden för 164 809 partiklar som erhölls den 23 maj 2019.

2.4.3.1 Definition

Definition D2.7: **Laboratoriedata** (även: laboratoriemätningar) består av värden som (a) registreras genom en sekvens av steg som omfattar minst ett manuellt steg för beredning av vattenprover och (b) beskriver observerade värden för egenskaper och storheter i ett fysiskt system av intresse.

Liksom givardata består laboratoriedata av serier av dataposter, med samma definitioner som i Avsnitt 2.4.1.2 (Definitioner D2.2 och D2.3). Laboratoriedata produceras av ett analytiskt instrument eller ett analytiskt test.

2.4.3.2 Exempel E2.5: laboratoriemätningar

Tabell 2.3 illustrerar några laboratoriemätningar. Alla mätningar har utförts i ett laboratorium med manuell beredning av vattenprover, vilket innebär att definition D2.7 uppfylls. Även om värdena i varje rad är kopplade till ett enda vattenprov, behandlas koncentrationsmätningarna av varje variabel (NH_4^+-N, NO_3^--N, PO_4^+-P, TOC) som enskilda datapunkter, i enlighet med definition D2.3. Tabellen innehåller 61 datapunkter, varav 8 värden saknas. Indikatorerna för saknade värden orsakas av att den uppmätta variabeln ligger under detektionsgränsen (anges med '<LOD'). I enlighet med definition D2.6 anser vi att de uppmätta värdena är förlorade. Indikeringen '<LOD' ger diagnostisk information om grundorsaken. Detektionsgränsen, LOD, är det lägsta värdet för koncentrationen i ett vattenprov som förväntas ge ett positivt värde (dvs. högre än noll) för mätningen. Detektionsgränsen är vanligtvis unik för en mätprincip och ett tillämpat analysprotokoll (t.ex. tillämpad utspädning). För en detaljerad genomgång av detta ämne och exakta definitioner hänvisar vi till Currie (1995, 1999) och ISO (1997).

Detektionsgränsen är ett användbart koncept för att förstå vad det maximala värdet för motsvarande koncentration kan ha varit, även om en exakt mätning kanske inte finns tillgänglig. Detektionsgränsen

Tabell 2.3 Exempel E2.5: en tabell som innehåller en uppsättning laboratoriemätningar inklusive frånvarande data, förlorade data och textinformation.

Datum/Tid	NH_4^+-N	NO_3^--N	PO_4^+-P	TOC	Kommentar
2018-11-02 11:00:00		9,3	0,60	2,90	
2018-11-06 14:00:00				3,40	
2018-11-13 11:30:00	<LOD	5,9	1,30	3,12	
2018-11-15 09:30:00	<LOD	5,3	0,40	2,36	
2018-11-19 12:20:00	0,041	11,0	1,20	5,87	
2018-11-21 15:00:00	3,050	11,5	1,20	8,40	
2018-11-26 14:15:00	<LOD	5,2	0,56	3,40	
2018-11-26 16:00:00				4,70	
2018-11-28 14:30:00	<LOD	6,6	1,00	3,00	
2018-12-03 14:00:00	<LOD	5,8	1,10	6,80	
2018-12-06 15:30:00	0,038	8,2	0,74	4,50	
2018-12-06 12:00:00				2,20	Batch 1.2.a.1
2018-12-06 12:00:00				2,56	Batch 1.2.a.2
2018-12-06 12:00:00				3,91	Batch 1.2.b.1
2018-12-06 12:00:00				3,13	Batch 1.2.b.2
2018-12-06 12:00:00				2,08	Batch 1.2.c.1
2018-12-06 12:00:00				2,51	Batch 1.2.c.2
2018-12-06 12:00:00				2,59	Batch 6.1
2018-12-06 12:00:00				2,96	Batch 6.2
2018-12-08 10:00:00	<LOD	5,4	0,36	2,75	
2018-12-10 09:50:00	<LOD	5,0	0,44	3,44	
2018-12-12 11:30:00	<LOD	8,0	1,00	2,90	
2018-12-18 14:00:00	2,170	5,8	2,40	10,40	

kan variera från ett vattenprov till ett annat (t.ex. på grund av förändringar i den använda utspädningen). I en ideal värld registreras därför detektionsgränsen individuellt för varje mätning där och när det är relevant. Alternativt kan man registrera den tillämpade utspädningsfaktorn och den analysmetod som användes så att detektionsgränsen för det outspädda vattenprovet kan beräknas utifrån en känd detektionsgräns för det utspädda provet. De metadata som beskriver laboratoriedata kan naturligtvis variera och innehåller vanligtvis referenser till insamling av vattenprov, beredning av vattenprov och analysprotokoll.

Det finns också 31 tomma celler i Tabell 2.3 Notera att det inte finns någon indikation i dessa celler om att värden saknas. Följaktligen anses dessa (ouppmätta) värden vara frånvarande. När dessa datapunkter paras ihop med respektive tidsstämpel bildar de totalt 61 dataposter, inklusive dataposter där datapunkten saknas, men utan att räkna med eventuella frånvarande data.

2.4.4 Metadata

I det här kapitlet beskrivs metadata kortfattat, främst för att ge en inledande definition. Eftersom det i stort sett inte finns någon övre gräns för antalet olika typer av metadata så avsätts ett helt kapitel, Kapitel 3 till att beskriva metadata och deras tillämpningar.

Definition D2.8: **Metadata** definieras som den information som används för att beskriva signaldata eller laboratoriedata.

Metadata kan användas för att beskriva flera aspekter som är relevanta för användning eller tolkning av signaldata eller laboratoriedata, inklusive, men inte begränsat till:

(1) Hur, var, när och varför uppgifterna genererades och/eller modifierades.
(2) Hur data kan eller bör tolkas.
(3) Kvaliteten på data.
(4) All kontextuell information om syftet, ursprunget och användningen av data.

Specifika exempel på metadata ges i Kapitel 3.

2.5 STRUKTURERADE OCH OSTRUKTURERADE DATA

De flesta signaldata innehåller datapunkter som har en konstant storlek, det vill säga de har ett konstant antal dimensioner och ett konstant antal värden längs varje dimension. Datapunkter som är förutsägbara i storlek är praktiska att hantera. Deras förutsägbara form (antal dimensioner) och storlek (längder längs varje dimension) underlättar många datarelaterade uppgifter, t.ex. kommunikation, långtidslagring, databasfrågor och navigering, dataanalys, datavisualisering och modellering.

Den senaste utvecklingen inom onlineinstrumentering har också gjort det möjligt att producera signaldata som inte följer ett format med konstant storlek. Ett exempel på detta visas nedan som inkluderar data i form av text eller mätningar av partikelegenskaper. Att ta bort begränsningar för konstant storlek gör det lättare att navigera och tolka genom visuell inspektion, men kräver specialiserad behandling för att möjliggöra effektiv användning i dataanalys, modellering eller automatisering. För att skilja på strukturerade data, som uppvisar ett format med konstant storlek, och ostrukturerade data, vars format är mer flexibelt men kräver specialiserad hantering för långsiktig lagring och användning ges definitionerna nedan.

De flesta signaldata är strukturerade med få undantag. De metadata som finns tillgängliga i en typisk anläggning förekommer däremot i både strukturerad och ostrukturerad form. Nedan definierar vi strukturerade och ostrukturerade data och illustrerar definitionerna med hjälp av ett antal exempel.

2.5.1 Strukturerade data

I detta avsnitt definieras strukturerade data och konsekvenserna av att använda sådana data diskuteras med stöd av ett antal exempel.

2.5.1.1 Definition

Definition D2.9: **En signal** sägs innehålla strukturerade data när varje datapunkt i varje datapost uppfyller följande krav:

(1) Antalet dimensioner är konstant.
(2) Antalet värden i alla dimensioner är konstant.
(3) Ett värdes position längs varje dimension är meningsfull och dess betydelse förändras inte över tiden.
(4) Alla datapunktsvärden är begränsade till att tillhöra en enda taltyp, t.ex. binär (*boolean*), heltal (*integer*) eller decimaltal (*float*).

Det tredje kravet innebär att en ändring av ett värdes position inom en datapunkt sannolikt kommer att ändra datapunktens giltighet. Med andra ord har ett värdes position betydelse. Denna idé utforskas ytterligare nedan.

2.5.1.2 Ytterligare anmärkningar

Matris och tabellformat. Den stora fördelen med strukturerade data är att de enkelt kan lagras i en matris eller tabell med fasta längder längs varje dimension. Positionen längs varje dimension (t.ex. radposition, kolumnposition) har en unik betydelse så att jämförelser av värden som förekommer i samma rad eller kolumn kan göras. Dessutom kan data enkelt visualiseras i ett diagram eller på

en instrumentpanel i realtid. Den enkla bearbetningen säkerställer också att man snabbt kan gå vidare till ytterligare analys utan omständliga dataförbehandlingssteg. Denna användarvänlighet är en utmärkande egenskap hos strukturerade data.

Icke-numeriska data kan struktureras. Strukturerade data kräver inte att värdena är numeriska. Till exempel är datakvalitetsstämplar (se Avsnitt 2.4.1.2 för definitioner) ofta en text med en fördefinierad och konstant längd. På samma sätt kan strukturerad signaldata innehålla specialiserade koder för frånvarande eller förlorade data, vilket diskuteras nedan.

Förlorade data. Avsaknad av eller saknade värden i en datapunkt har inga konsekvenser för tillämpningen av definitionen ovan. Strukturerade data kräver dock att en specialiserad symbol eller text används för att ange ett värde som saknas (t.ex. *NaN*) för att säkerställa att antalet värden i alla dimensioner förblir konstant. Dessutom underlättar användningen av en accepterad symbol eller text identifiering och hantering av frånvarande och saknade värden under dataanalysen.

2.5.1.3 Exempel E2.1 och E2.2 (fortsättn.): univariata signaler

Univariata signaler eller tidsserier, dvs. data som innehåller skalära (dvs. 0D) datapunkter, är grundläggande exempel på strukturerade data, vilket visas i Exempel E2.1 och E2.2.

2.5.1.4 Exempel E2.3 (fortsättn.): absorbansspektrofotometri med ultraviolett ljus

De data som produceras av spektrofotometriska sensorer klassificeras som strukturerade eftersom (a) varje datapunkt, dvs. spektrum, alltid är 1D, (b) antalet absorbansmätningar i en enskild datapunkt är konstant för en given sensor och (c) de enskilda absorbansmätningarna i samma position motsvarar samma ljusvåglängd. Den strukturerade karaktären på data gör det lätt att ordna alla dataposter i tabellformat, vilket visas i Tabell 2.2. De spektrofotometriska mätningarnas strukturerade karaktär gör det också möjligt att konvertera värdena till en värmekarta med rektangulära rutor. Figur 2.6 visar 24 spektra i denna form. Det är samma spektra som användes för att producera Figur 2.3 och Tabell 2.2.

2.5.1.5 Exempel E2.5 (fortsättn.): koncentrationsmätningar

Koncentrationsdata från Exempel 2.5 (Tabell 2.3) och de tillhörande tidsstämplarna klassas som strukturerade data. Det är lätt att se att varje datapunkt är skalär (0D) och att alla villkor i definition D2.9 uppfylls. Observera att förekomsten av frånvarande värden (tomma celler) och saknade värden (<LOD) inte har påverkar klassificering som strukturerade data.

2.5.2 Ostrukturerade data

I detta avsnitt definieras ostrukturerade data där även konsekvenserna av att använda ostrukturerade data beskrivs med ett antal exempel.

2.5.2.1 Definition

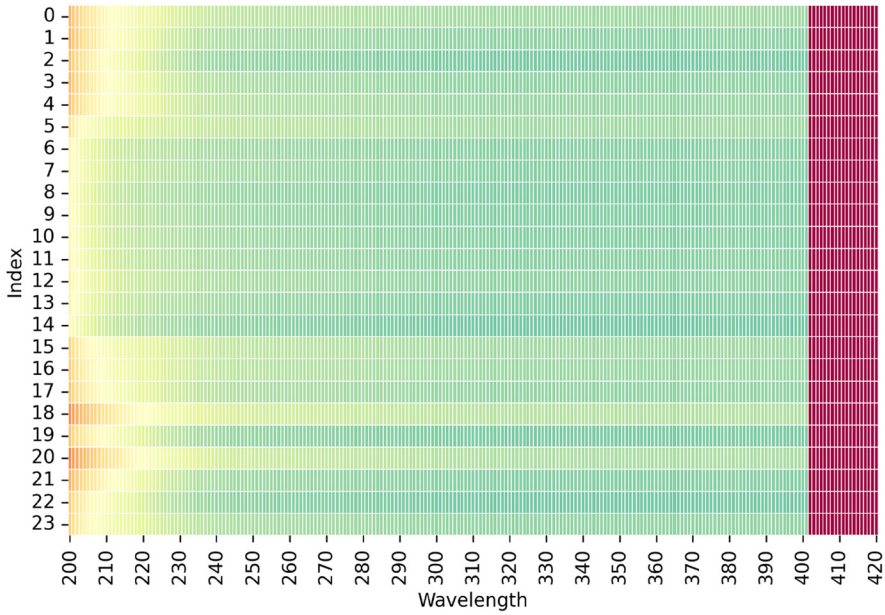
Definition D2.10: **Data sägs vara ostrukturerade** så snart ett av dessa villkor är uppfyllt:

(1) Antalet dimensioner för datapunkterna kan variera.
(2) Antalet värden längs en eller flera dimensioner kan variera.
(3) Ett värdes position längs en eller flera dimensioner kan ändras utan att det får konsekvenser för den fortsatta analysen.
(4) Datapunktsvärden kan tillhöra flera olika taltyper, t.ex. binärt, heltal eller decimaltal.

2.5.2.2 Hantering av ostrukturerade data

Att samla in ostrukturerade data har två viktiga konsekvenser:

- Lagring av ostrukturerade data kräver en specialiserad metod för datahantering, eftersom data inte kan struktureras på ett vettigt sätt i en tabell struktur. Några typer av datalagringstekniker som kan hantera ostrukturerade data diskuteras i Kapitel 4.



Figur 2.6 Exempel E2.3: Spektrofotometriska sensorer levererar data som är 1D med en fast längd (här: 221). Som ett resultat kan man enkelt visualisera de uppmätta värdena i en värmekarta (*'heatmap'*) med ett regelbundet rutnät. Cellerna är markerade med vita linjer. Färgen anger det uppmätta absorbansvärdet som en funktion av tid (t) och våglängd (λ). Förlorade data indikeras med rött.

- För att kunna använda vanliga modeller och verktyg för dataanalys krävs att ostrukturerade data först omvandlas till ett strukturerat format. Denna typ av förbearbetning av data diskuteras i detalj i Kapitel 6 (Avsnitt 6.6.7).

2.5.2.3 Ytterligare kommentarer

Tabellformat. Även om ostrukturerade data kan visas i tabellformat är det osannolikt att det är meningsfullt att jämföra värden som visas på samma rad eller i samma kolumn.

Typiska fall. De flesta ostrukturerade datakällor uppfyller villkor 2 och/eller 3 och det finns systematiska metoder för att hantera detta. Ostrukturerade data för vilka villkor 1 gäller är sällsynta. Som ett hypotetiskt exempel kan man tänka sig att en kamera som producerar bilder med inställningar som ändras över tid (vissa gråskalebilder och vissa färgbilder) kommer att leda till bilddata som uppfyller det första villkoret. Det finns inga metoder som vi känner till för att hantera detta fall på ett systematiskt sätt.

Frånvarande och förlorade data. Om data saknas eller är frånvarande påverkar det inte huruvida data anses vara strukturerade eller inte. I allmänhet gör användningen av ostrukturerade dataformat det möjligt att lagra data genom att helt enkelt ta bort poster som motsvarar frånvarande eller förlorade data, och därmed undvika behovet av ett särskilt värde eller en kod som anger frånvaro eller förlorad datapunkt. Detta kan spara en del lagringsutrymme. Detta innebär dock att antalet värden kan ändras längs en eller flera dimensioner. Eftersom de flesta dataanalysverktyg och modeller kräver strukturerade data med fasta dimensioner och antal värden i varje dimension, rekommenderar vi inte användning av ostrukturerade data i syfte att minimera lagringsutrymmet eller för att undvika användning av specialiserade symboler för att ange att data saknas. Som tidigare nämnts kan det dessutom vara skadligt att ta bort information om förlorade data.

Tabell 2.4 Exempel E2.4: Flödescytometridata.

Vattenprov A					Vattenprov B				
Tidsstämpel: 2019-05-09T11:35:32Z					Tidsstämpel: 2019-05-23T14:34:24Z				
Partikel nr.	FSC	SSC	FITC	PerCP	Partikel nr.	FSC	SSC	FITC	PerCP
0	968	384	4 802	270	0	509	299	7 542	837
1	3 428	770	32 189	2 459	1	0	283	1 368	162
2	0	181	28 152	2 421	2	6 895	3 691	86 812	9 478
3	259	0	1 353	373	3	0	161	1 273	183
4	7 984	5 331	89 288	12 484	4	3 990	3 487	101 834	6 708
…	…	…	…	…	…	…	…	…	…
56 353	3 203	1 440	32 005	2 495	164 805	302	1 083	26 686	3 764
56 354	275	18	656	253	164 806	651	0	32 965	3 659
56 355	55 324	4 066	1 009	2 743	164 807	953	359	32 600	2 995
56 356	129	0	4 703	766	164 808	9 741	1 557	1 583	192
56 357	507	2 103	40 058	5 055	164 809	7 353	4 327	41 676	5 346

Mätvärden för de första och sista fem partiklarna för den datapost som visualiseras i den vänstra respektive högra figuren i Figur 2.5 för vattenproverna A och B. Dessa data klassificeras som ostrukturerade data eftersom antalet rader i dataposterna ändras för varje vattenprov.

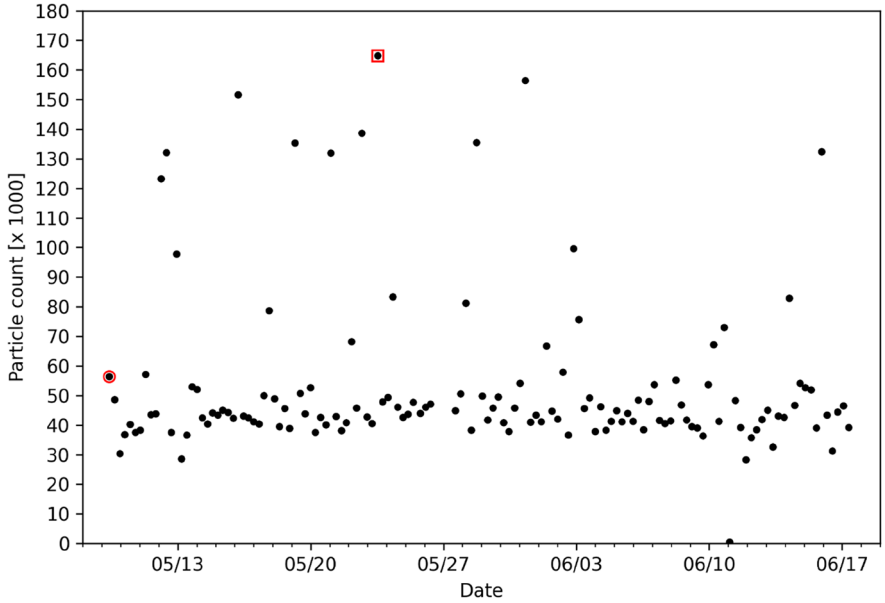
2.5.2.4 Exempel E2.4 (fortsättn.): flödescytometri

Tabell 2.4 visar värden för fyra ljusintensiteter som uppmätts i två vattenprover. Varje vattenprov motsvarar en enda datapost. Varje datapost består av en tidsstämpel och mätningar av partiklar i vattenprovet. Värdena visas för de fem första och de fem sista partiklarna i motsvarande dataposter. Notera att antalet partiklar i de båda proven skiljer sig åt (56 358 och 164 810). Figur 2.7 visar antalet partiklar som en funktion av tiden för en signal som innehåller totalt 137 dataposter. Det är tydligt att längden längs dataposternas första dimension (partikelantalet) varierar från en datapost till nästa. Detta innebär att villkor 2 i definition D2.10 är tillämplig och att data klassificeras som ostrukturerade data. Observera att längden längs den andra dimensionen (emissionsvåglängden) är konstant. Den ordning i vilken partiklarna utsätts för excitation och emission har ingen betydelse vid tolkningen av data. Man kan faktiskt blanda om radpositionerna utan att det får konsekvenser för den fortsatta analysen av dessa data. Detta innebär att villkor 3 i definition D2.10 också gäller. I Kapitel 6 (Avsnitt 6.6.7) används samma dataset för att illustrera tre metoder för förbehandling av data som syftar till att omvandla denna typ av ostrukturerade data till en strukturerad form.

2.5.2.5 Exempel E2.5 (fortsättn.): text

Text är ett extremt effektivt och mångsidigt verktyg för kommunikation och lagring av information. Mycket beskrivande information uttrycks i form av text, särskilt när det behövs en form som kan tolkas av människor. Exempel från reningsverk är fristående dokument som manualer, protokoll och rapporter samt anteckningar som läggs till för att beskriva och isolera delmängder av större datamängder (t.ex. text som beskriver en avvikelse eller tillämpade signalkorrigeringar). Inom området språkbearbetning (*natural language processing – NLP*), ett delområde inom maskininlärning, betraktas text vanligtvis som 1D, det vill säga som en lista med ord åtskilda av mellanslag och andra avgränsare. I denna form har textinformation inget fast antal tecken eller ord, så att villkor 2 i definition D2.10 är tillämpligt.

Kommentarerna i den sista kolumnen i Exempel 2.5 (Tabell 2.3) klassificeras som ostrukturerade data. Fritextkommentarerna har varierande längd, från 0 tecken (ingen kommentar) till 13 tecken. Denna variation i längd är anledningen till att dessa kommentarer betraktas som ostrukturerade data. Observera att textdata inte automatiskt betraktas som ostrukturerade data.



Figur 2.7 Exempel E2.4: Flödescytometridata. Antal partiklar för varje datapost i en flödescytometrisignal som en funktion av tiden. Cirkeln och kvadraten motsvarar vattenproverna A och B i Figur 2.5.

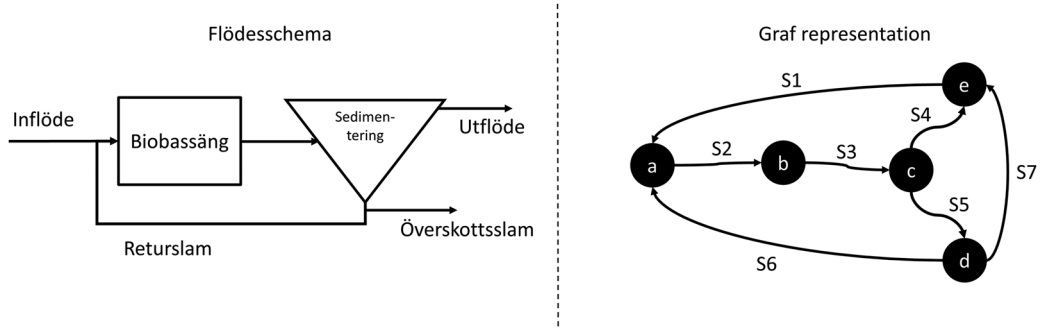
När värdena har en fast längd, som i Exempel 2.3 (se Tabell 2.2), betraktas de som strukturerade data. På samma sätt kan etiketter som används för att kommentera data ha en beskrivning med fast längd och anses då vara strukturerade data.

2.5.2.6 Exempel E2.6: balansmatris för att representera anläggningens layout och flödesschema

Ett reningsverks processkonfiguration och driftsätt förändras över tid. Till exempel så kan ytterligare reningssteg krävas eller förbättringar av styrsystemet. Information om sådana förändringar lagras i olika dokument såsom rör- och instrumentdiagram (*Piping and Instrumentation Diagram – P&ID*), konstruktions- och/eller produktspecifikationer, manualer, driftrutiner *(Standard Operating Procedure – SOP*) och andra dokument. De flesta anläggningar har också flera komponenter som manövreras manuellt (t.ex. ventiler och omkopplare). De flesta styrsystem gör det dessutom möjligt för operatörerna att aktivera eller avaktivera delar av reglerkretsarna (t.ex. för vinterdrift, vid underhållsarbeten eller givarunderhåll). Dessa manuella och halvautomatiska förändringar i systemet registreras ofta tillfälligt (*ad hoc*) eller inte alls.

Metadata som beskriver förändringar i anläggningens konstruktion och konfiguration, styrsystemets konstruktion och konfiguration samt manuella och halvautomatiska förändringar i driften är vanligtvis ostrukturerade och svåra att nyttja för dataanalys i sitt standardformat. Samtidigt är det ofta viktigt att ha enkel och fullständig tillgång till dessa metadata för att ta rätt beslut vid en dataanalys. Ett tillvägagångssätt som skulle kunna bidra med en lösning är om det gick att beskriva anläggningens eller styrsystemets konfiguration med hjälp av en abstrakt, matematisk beskrivning av, som även håller reda på förändringar i denna beskrivning allteftersom tiden går. Detta illustreras härnäst med en graf som är utformad för att representera flödesbalanser.

En graf är ett matematiskt objekt som består av noder och kanter (*nodes and edges*) och kan användas för att sammanfatta nätverksstrukturen i ett system som består av separata komponenter. Figur 2.8 visar ett flödesschema för ett litet reningsverk till vänster och motsvarande graf till höger.



Figur 2.8 Exempel 2.6: Användning av grafer för att beskriva flödesmönster och massbalanser i en matematiskt exakt form. Till vänster: en enkel processkonfiguration – till höger: ett diagram över anläggningslayouten, som ger all information som behövs för att konstruera massbalanser.

Grafen erhålls genom att identifiera dess noder. Dessa motsvarar de platser i anläggningens flödesschema där två eller flera strömmar möts. I exemplet är dessa (a) inblandningszonen av returslamflöde med inkommande flöde (b) det biologiska reningssteget, (c) sedimenteringsbassängen, och (d) avsättning av överskottsslam och returslam. En omgivningsnod (e) läggs till för att koppla samman alla flöden som korsar systemgränsen, i det här fallet de strömmar som lämnar/kommer in till reningsverket. Noderna visas som cirklar i den högra figuren i Figur 2.8. Sedan läggs kanter till för att representera varje flöde. Dessa visas som pilar i den högra panelen i Figur 2.8. Riktningen på dessa pilar motsvarar positiva värden i flödesriktningen. Positiva värden för inflödet innebär t.ex. ett positivt nettoflöde från omgivningen (e) till inblandningszonen (a). Det finns sju strömmar i detta enkla fall (S1-S7). För att använda denna graf på ett effektivt sätt i datorsystem kan den lagras i form av en balansmatris (*incidence matrix*), $\mathbf{M}$. Detta illustreras i Tabell 2.5. Balansmatrisen innehåller en rad för varje nod och en kolumn för varje ström. För en given nod (rad i matrisen) placeras värdet +1 för kanter som är riktade mot noden och värdet –1 för kanter som är riktade bort från noden. Flöden som inte är kopplade till den aktuella noden får värdet noll i motsvarande rad. Om alla element fylls i korrekt blir summan av alla element i varje kolumn noll.

En praktisk fördel med den grafbaserade metoden är att flödesbalanser kan definieras på ett enkelt sätt. Till exempel kan det hydrauliska balansfelet över sedimenteringsbassängen skrivas enligt följande:

$$e_3 = \sum_{j=1}^{7} \mathbf{M}_{3,j} \cdot q_j$$

Tabell 2.5 Exempel E2.6: Balansmatrisen är en kompakt representation av anläggningens flödesschema och de ekvationer som representerar flödes- och massbalanser.

		Kant/Ström						
Fysiskt element	**Nod**	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**
Inblandningszon	a	+1	–1	0	0	0	+1	0
Biologiskt reningssteg	b	0	+1	–1	0	0	0	0
Sedimenteringsbassäng	c	0	0	+1	–1	–1	0	0
Avsättning överskottsslam	d	0	0	0	0	+1	–1	–1
Omgivning	e	–1	0	0	+1	0	0	+1

där q_j är den hydrauliska flödeshastigheten i ström j. Balansmatrisen **M** ger således en kompakt men ändå fullständig representation av alla flödesbalanser. Detta kan enkelt utvidgas till att även specificera och använda mass- och energibalanser (se t.ex. Spindler, 2014). Grafer är mycket användbara för att utvärdera vilka variabler (flödeshastigheter och koncentrationer) som kan uppskattas utifrån tillgängliga mätningar. Detta kan i sin tur användas för att placera flödes- och koncentrationssensorer på ett optimalt sätt givet bedömning och kontroll av datakvalitet (Villez *et al.*, 2016, 2020). I sin matrisform (**M**) är grafer lämpliga för databalansering (*data reconciliation*) och feldetektion (se t.ex. Le *et al.*, 2018, 2022; Puig *et al.*, 2008).

Eftersom flödesschemat för ett typiskt reningsverk vanligtvis förändras, kommer därmed även grafen och balansmatrisen att förändras på motsvarande sätt. Exempelvis kan antalet noder ändras när processenheter tas i eller ur drift. Antalet kanter ändras när recirkulationsflöden tas i eller ur drift. Dessa förändringar i dimensionerna innebär att grafer normalt bör betraktas som ostrukturerade data (villkor 2 i definition D2.10).

En systematisk användning av grafer kan också förenkla en dataanalys. Genom ett register och lagring i en databas av sådana grafer går det att automatiskt identifiera intressanta driftperioder då anläggningens konfigurationen är konstant, vilket möjliggör dataanalys av jämförbara perioder, trots förändringar i anläggningens layout eller drift. Detta kan i sin tur förenkla en kostnads- och nyttoanalys mellan olika driftsätt och styrstrategier. På lång sikt bör metadata i form av grafer vara användbara för att underhålla anläggningsmodeller online med begränsad mänsklig övervakning (såsom en digital tvilling).

Är grafer metadata? Grafer och balansmatriser används och lagras ofta separat från vanliga signaldata. Som en separat informationskälla nämns grafer sällan i samband med metadata. Som exemplet ovan illustrerar kan grafer användas för att dokumentera hur flera variabler på en anläggning, och i förlängningen deras mätningar, är relaterade till varandra. Av denna anledning bör information om anläggningens- och styrsystemets layout inkluderas som potentiella metadata för att beskriva ställdon, sensorer, och börvärdesdata.

I framtiden är det troligt att det blir allt vanligare med förändringar av processen och dess styrsystem. Eftersom historiken för modifieringar av både anläggningen och styrsystemen kanske inte är uppenbar i efterhand och eftersom anläggningsmodeller kan kräva historiska data som sträcker sig över flera konfigurationer, kan det värdefullt att samla in sådan information på ett automatiskt och systematiskt sätt.

2.6 VERSIONSHANTERING AV DATA

Det är vanligt att signaldata modifieras efter insamling för att förbättra kvaliteten (t.ex. genom brusreducering) och göra dem mer användbara (t.ex. ändra enheter eller medelvärdesbilda data för ett visst tidsintervall). Under idealiska förhållanden är alla steg som används för att modifiera signaldata spårbara på ett sätt som möjliggör granskning och, där så är möjligt, återställa modifieringarna. Detta kan uppnås med s.k. versionshantering, vilket innebär att man sparar och spårar alla ändringar av insamlade signaldata. Detta kan jämföras med funktionen *Spåra ändringar* (*track-changes*) i textredigeringsprogram. Vanligtvis bygger versionshantering av data på samma principer som versionshantering av programvara. När detta utförs möjliggör det navigering framåt och bakåt i datasetens historik.

2.6.1 Hantering av dynamiska datamängder

Ett dataset kommer att utvecklas med tiden och/eller under ett projekts gång när mer data eller information blir tillgänglig. Därför bör versionshantering av data tillämpas på dataset som modifieras genom något av följande steg:

(1) Data läggs till rutinmässigt (t.ex. genom registrering av givarsignaler).
(2) Data tas bort rutinmässigt (t.ex. genom borttagning av avvikande värden).
(3) Data modifieras rutinmässigt (t.ex. genom brusfiltrering).

Därför är det bäst att identifiera en central lagringsplats för alla uppgifter som omfattas av något av dessa steg. Ett sätt att göra detta beskrivs nedan.

2.6.2 Definitioner

I detta avsnitt definieras olika versioner av data som bör spåras och versionshanteras, vilket illustreras med ett exempel.

2.6.2.1 Originaldata eller rådata

Definition D2.11: **Rådata** är data som produceras av en signalkälla, t.ex. en givare.

Många instrument genomför viss modifiering av rådata, t.ex. brusreduktion, innan de blir tillgängliga för loggning, vilket förhindrar insamling av data i obearbetad form. I dessa fall bör modifieringsmetodiken dokumenteras så bra som möjligt i form av metadata vilket förklaras i Kapitel 3 (Avsnitt 3.2.4).

2.6.2.2 Primärdata

Definition D2.12: **Primärdata** (*primary data*) är data som består av den mest exakta, omfattande och uppdaterade versionen av en given uppsättning data. Denna datamängd används som den *sanna datakällan*, det vill säga en datamängd som varje dataanvändare accepterar som den bästa tillgängliga informationskällan.

Genom att tydligt etablera en källa till primärdata (t.ex. en central databas) får alla dataanvändare tillgång till samma information. En god hantering av primärdata syftar till att uppnå följande:

(1) Primärdata omfattar en fullständig historik över varje datapost i databasen, inklusive (a) alla modifieringar som utförts för att omvandla de ursprungliga rådata till lagrade data och (b) alla modifieringar som utförts efter registrering i denna databas. Typiska modifieringar som görs före lagring omfattar, men är inte begränsade till, formatering, ändring av enhet, ersättning av förlorade värden och annotering. Kapitel 3 ger en mer utförlig översikt av detta och Kapitel 6 beskriver vanliga metoder för att modifiera data efter insamling.

(2) Minimera mängden värdefull information som går förlorad. För att klara detta bör vissa modifieringar av data undvikas i primärdata. Dessa inkluderar alla steg som uppnår (a) brusreducering, (b) borttagning av data eller (c) aggregering av data. Om dessa steg ändå tillämpas bör primärdata innehålla en kopia av samma data där data inte har modifierats. Det finns två anledningar till detta. För det första kan många modifieringsförfaranden ändras men kanske inte kan återkallas när de väl har tillämpats. För det andra kommer troligen olika syften och tillämpningar av data att kräva olika modifieringar av primärdata.

(3) Kontinuerlig uppdatering av data genom tillägg av nya dataposter och noggrann uppdatering för att förbättrad datakvaliteten.

Alla exempel ovan (Exempel 2.1 till 2.6) är tänkta att lagras som primärdata. För vissa projekt är detta inte den mest praktiska formen av data. Exempelvis har luftflödesdata i Exempel 2.1 och 2.2 asynkrona tidsstämplar, vilket försvårar vidare analys. På samma sätt kan flödescytometridata (Exempel 2.4) inte hanteras direkt med konventionella dataanalysmetoder på grund av det varierande antalet mätvärden per vattenprov. Denna typ av problem med primärdata bör hanteras genom att producera s.k. sekundärdata, och inte genom att irreversibelt modifiera primärdata. Detta finns två skäl till detta. För det första kan förbehandlingsprocedurerna för att ta fram strukturerade varianter av ostrukturerade data förändras. Man bör därför kunna tillämpa en annan eller ny förbättrad procedur senare. För det andra kommer förbehandlingssteg som syftar till att skapa struktur (t.ex. synkronisering, binning) att leda till informationsförlust.

2.6.2.3 Sekundärdata

Definition D2.13: **Sekundärdata** är data som har genererats genom urval och modifiering av primärdata för ett specifikt sammanhang eller syfte.

Sekundärdata genereras vanligtvis för ett specifikt projekt (t.ex. uppgradering av en anläggning) eller en specifik uppgift (t.ex. beräkna massflöden vid ett modelleringsprojekt). För att göra primärdata praktiskt användbara kan flera databehandlingssteg krävas, som vart och ett påverkar informationsinnehållet. Man kan t.ex. vilja utvärdera syrehaltsregulatorn genom att jämföra styrsignalen med luftflöde från Exempel 2.1 och 2.2. Eftersom tidsstämplarna inte stämmer överens i dessa två signaler, kan man få synkroniserade tidsstämplar genom interpolering eller medelvärdesbildning. De steg som används för att ta fram sekundärdata är ofta unika för en viss uppgift, studie eller ett visst projekt. De kan omfatta aggregering, filtrering, ersättning, utjämning, borttagning av datapunkter etc. Kapitel 6 i denna rapport innehåller en omfattande förteckning över databehandlingssteg som kan användas vid generering av sekundärdata.

I motsats till primärdata uppdateras sekundärdata vanligtvis inte kontinuerligt. Istället är ett sekundärdataset sannolikt statiskt under ett projekt. Det innebär också att sekundärdata snabbt kan bli föråldrade och osynkroniserade med befintliga primärdata. Av denna anledning rekommenderas att:

(1) Aldrig lägga till data i en sekundär datauppsättning om den inte först har gjorts till en del av den primära datauppsättningen.
(2) Tilldela ett 'bäst-före-datum' till varje sekundär datauppsättning.
(3) Dokumentera processen för att generera sekundärdata så att sekundärdata enkelt kan regenereras från primärdata. Se Kapitel 6 för typiska databehandlingssteg som bör finnas tillgängliga i form av väldokumenterad kod eller annan typ av dokumentation.

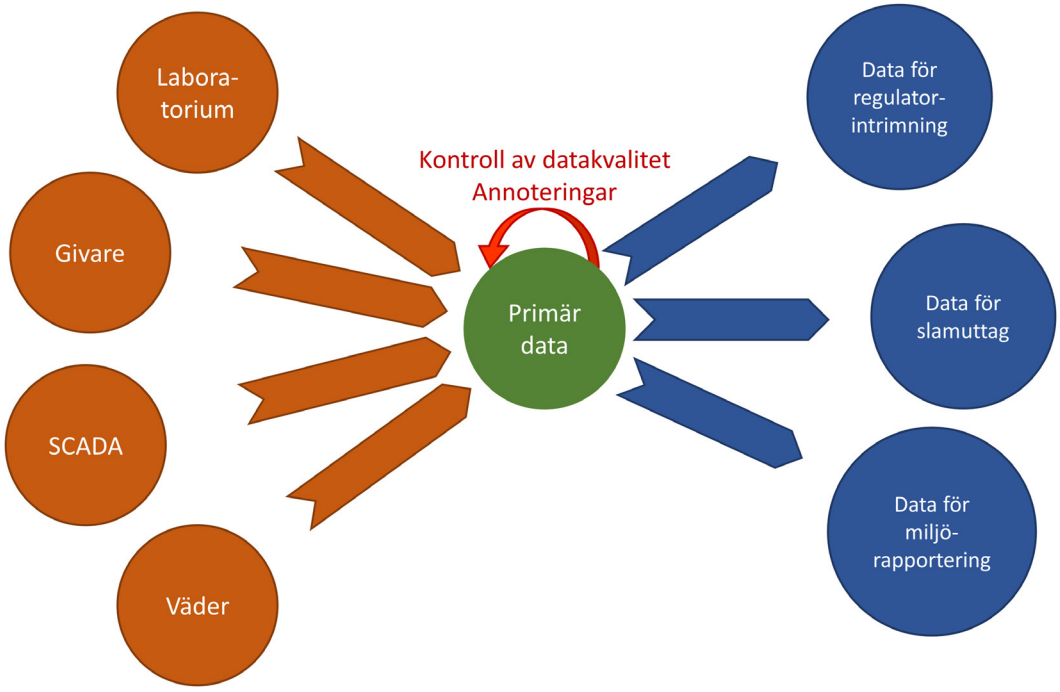
2.6.2.4 Spårbarhet av data: dokumenterad historik från rådata till sekundärdata

Definition D2.14: **Data är spårbara** om det finns system som möjliggör övervakning av hela processen för insamling, lagring och bearbetning av data för en korrekt och effektiv användning.

Att dokumentera och spåra alla ändringar av rådata fram till sekundärdata kan bidra till att rätt beslut tas i de fall sekundärdata utgör beslutsunderlag. Dessutom gör detta det möjligt att reproducera dataset utan osäkerhet om hur data har genererats.

I Figur 2.9 ges ett exempel på hur spårbarhet av data kan uppnås med hjälp av definitionen ovan. Data som produceras i laboratoriet, givarsignaler, styrsignaler som produceras av ett SCADA- eller DCS-system (*Supervisory Control and Data Acquisition*) och väderrelaterade data samlas i ett enda dataset som innehåller en omfattande datamängd som kallas primärdata och som fungerar som en enda, allmänt accepterad informationskälla. Dessa primära data används som utgångspunkt för att ta fram andra dataset för specifika uppgifter eller projekt. Ett sekundärt dataset byggs t.ex. upp specifikt för att användas för intrimning av regulatorer. Detta dataset kommer sannolikt att ha en hög tidsupplösning och synkroniserade tidsstämplar. Ett andra dataset skapas för att analysera slamhanteringen på reningsverket. Detta kräver istället ett dataset med låg tidsupplösning som endast innehåller ett fåtal variabler. Det kan t.ex. innehålla vissa variabler som baseras på andra mätningar, såsom uppskattning av slamåldern (SRT). Slutligen kan ett tredje dataset sammanställas med aggregerade data för årlig miljörapportering. Detta dataset omfattar data med mycket låg tidsupplösning.

I idealfallet är procedurerna för att generera olika dataset tillgängliga som tydligt dokumenterade koder eller procedurer, vilket gör det möjligt att producera härledda dataset från primärdata med hjälp av information som endast finns i koden eller proceduren. En väldokumenterad kod eller procedur kommer således att inkludera och fullständigt specificera varje steg i genereringen av dessa sekundära dataset. Detta innebär att en regulatorintrimning kan kontrolleras senare genom enkla modifieringar av motsvarande skript. På samma sätt kan en uppdatering av proceduren för att beräkna SRT genomföras genom att ett uppdaterat skript exekveras. Slutligen kan skriptet för att producera data för årets årsrapport enkelt återanvändas för nästa års rapport.



Figur 2.9 Dataflödesdiagram med olika versioner av primärdata (rött) och modifierade data till sekundärdata (blått).

2.6.2.5 Versionshantering av data och mjukvara: likheter och skillnader

Versionshantering av data och mjukvara gör det möjligt att spåra alla ändringar i en datamängd eller programkod. Båda typerna av versionskontroll används för att säkerställa tillförlitligheten och reproducerbarheten i datorstödda beslutsprocesser. Idag är versionshantering av mjukvara en utbredd metod, medan versionshantering av data inte är det. I praktiken kan versionshantering av mjukvara utföras effektivt med relativt små krav på datalagring. Däremot är lagringskraven för versionshantering av data idag stora. Med tiden förväntas mer avancerade lösningar för versionshantering av data minska dessa krav samtidigt som användarvänligheten förbättras.

2.7 SLUTSATER

- I denna rapport skiljer vi på signaldata, laboratoriedata och metadata.
- De flesta givarsignaler producerar strukturerade data, vilket sannolikt kommer bestå även i framtiden.
- Många typer av metadata och specialiserade givarsignaler är ostrukturerade och därför svårare att hantera och analysera med traditionella metoder.
- En god datahanteringsmetodik inkluderar att:
  - Rådata lagras utan modifieringar.
  - En unik datamängd med accepterad kvalitet definieras som den bästa datakällan på ett öppet och tydligt sätt (primärdata).
  - Processer och metoder som används för att rutinmässigt modifiera data med dokumenteras.

TACKORD

Exemplen 2.1, 2.2 och 2.3 tillhandahölls av Ali Gagnon (HRSD). Exempel 2.4 tillhandahölls av Michael Besmer (onCyt Microbiology AG, Dübendorf, Schweiz). Exempel 2.5 tillhandahölls av Angelika Hess och Eberhard Morgenroth (Eawag, Dübendorf, Schweiz). Exempel 2.6 är inspirerat av en diskussion med Adrienne Menniti (Clean Water Services, Hillsboro, OR, USA). Detta kapitel har granskats av Ali Gagnon (HRSD) och Matthew Wade (Newcastle University).

REFERENSER

Besmer M. D., Weissbrodt D. G., Kratochvil B. E., Sigrist J. A., Weyland M. S. and Hammes F. (2014). The feasibility of automated online flow cytometry for in-situ monitoring of microbial dynamics in aquatic ecosystems. *Frontiers in Microbiology*, **5**, 265. https://doi.org/10.3389/fmicb.2014.00265

Bhaskaran K. and Smeeth L. (2014). What is the difference between missing completely at random and missing at random? *International Journal of Epidemiology*, **43**(4), 1336–1339. https://doi.org/10.1093/ije/dyu080

Currie L. A. (1995). Nomenclature in evaluation of analytical methods including detection and quantification capabilities (IUPAC recommendations 1995). *Pure and Applied Chemistry*, **67**(10), 1699–1723. https://doi.org/10.1351/pac199567101699

Currie L. A. (1999). Detection and quantification limits: origins and historical overview. *Analytica Chimica Acta*, **391**(2), 127–134. https://doi.org/10.1016/S0003-2670(99)00105-1

Dong Y. and Peng C. Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, **2**(1), 222. https://doi.org/10.1186/2193-1801-2-222

Hammes F. and Egli T. (2010). Cytometric methods for measuring bacteria in water: advantages, pitfalls and applications. *Analytical and Bioanalytical Chemistry*, **397**, 1083–1095. https://doi.org/10.1007/s00216-010-3646-3

ISO (1997). ISO 11843-1:1997, Capability of detection – Part 1: Terms and definitions. 10pp.

Kang N (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, **64**(5), 402–406. 10.4097/kjae.2013.64.5.402

Le Q. H., Verheijen P. J., van Loosdrecht M. C. and Volcke E. I. (2018). Experimental design for evaluating WWTP data by linear mass balances. *Water Research*, **142**, 415–425. https://doi.org/10.1016/j.watres.2018.05.026

Le Q. H., Verheijen P. J., van Loosdrecht M. C. and Volcke E. I. (2022). Application of data reconciliation to a dynamically operated wastewater treatment process with off-gas measurements. *Environmental Science: Water Research & Technology*, **8**, 2114–2125. https://doi.org/10.1039/D2EW00006G

Lepot M., Torres A., Hofer T., Caradot N., Gruber G., Aubin J. B. and Bertrand-Krajewski J. L. (2016). Calibration of UV/Vis spectrophotometers: a review and comparison of different methods to estimate TSS and total and dissolved COD concentrations in sewers, WWTPs and rivers. *Water Research*, **101**, 519–534. https://doi.org/10.1016/j.watres.2016.05.070

Mašić A., Santos A. T., Etter B., Udert K. M. and Villez K. (2015). Estimation of nitrite in source-separated nitrified urine with UV spectrophotometry. *Water Research*, **85**, 244–254. https://doi.org/10.1016/j.watres.2015.08.031

OPC Foundation (2003). OPC Data Access Custom Interface Specification. Version 3.0. 190pp.

Osman M. S., Abu-Mahfouz A. M. and Page P. R. (2018). A survey on data imputation techniques: water distribution system as a use case. *IEEE Access*, **6**, 63279–63291. https://doi.org/10.1109/ACCESS.2018.2877269

Puig S., Van Loosdrecht M. C. M., Colprim J. and Meijer S. C. F. (2008). Data evaluation of full-scale wastewater treatment plants by mass balance. *Water Research*, **42**(18), 4645–4655. https://doi.org/10.1016/j.watres.2008.08.009

Rieger L., Langergraber G. and Siegrist H. (2006). Uncertainties of spectral in situ measurements in wastewater using different calibration approaches. *Water Science and Technology*, **53**(12), 187–197. https://doi.org/10.2166/wst.2006.421

Spindler A. (2014). Structural redundancy of data from wastewater treatment systems. Determination of individual balance equations. *Water Research*, **57**, 193–201. https://doi.org/10.1016/j.watres.2014.03.042

Villez K., Vanrolleghem P. A. and Corominas L. (2016). Optimal flow sensor placement on wastewater treatment plants. *Water Research*, **101**, 75–83. https://doi.org/10.1016/j.watres.2016.05.068

Villez K., Vanrolleghem P. A. and Corominas L. (2020). A general-purpose method for Pareto optimal placement of flow rate and concentration sensors in networked systems – With application to wastewater treatment plants. *Computers & Chemical Engineering*, **139**, 106880. https://doi.org/10.1016/j.compchemeng.2020.106880

Wang Y., Hammes F., De Roy K., Verstraete W. and Boon N. (2010). Past, present and future applications of flow cytometry in aquatic microbiology. *Trends in Biotechnology*, **28**(8), 416–424. https://doi.org/10.1016/j.tibtech.2010.04.006

doi: 10.2166/9781789065107_0031

Kapitel 3

Metadata för en systematisk beskrivning av signaldata

J. Alferes[1], J. A. Baeza[2], N. Nicolaï[3], J-D. Therrien[3], Q. Plana[3,4] och K. Villez[5]

[1]*VITO – WaterClimateHub, Ostende, Belgien*
[2]*Universitat Autònoma de Barcelona, Cerdanyola del Vallès (Barcelona), Spanien*
[3]*modelEAU, Université Laval, Québec, QC, Kanada*
[4]*Eurecat – Kataloniens teknikcentrum, Barcelona, Spanien*
[5]*Oak Ridge National Laboratory, Oak Ridge, TN, USA*

Nyckelord: annotering, kontextuell information, datakvalitet, metadata, givarhistorik

Detta kapitel syftar till att ge en grundlig inblick i de olika typer av metadata som kan vara användbara vid design, optimering och reglering av reningsverk. Metadata kan grupperas i tre huvudkategorier: (a) metadata som beskriver hur signalen genererats, (b) metadata som beskriver signalens kvalitet och (c) kontextuell information om signalen och dess tillämpning. Var och en av dessa kategorier introduceras och förklaras i tre separata avsnitt. Detta kapitel ger svar på *vad* som anses vara metadata. I mindre utsträckning ges rekommendationer om urvalet av metadata för långtidslagring. I Kapitel 4 förklaras istället *var* och *hur metadata bör lagras*. Kapitel 5 och 6 förklarar *hur* metadata kan samlas in genom särskilda metoder för givarvalidering (Kapitel 5) eller med hjälp av dataanalytiska metoder (Kapitel 6).

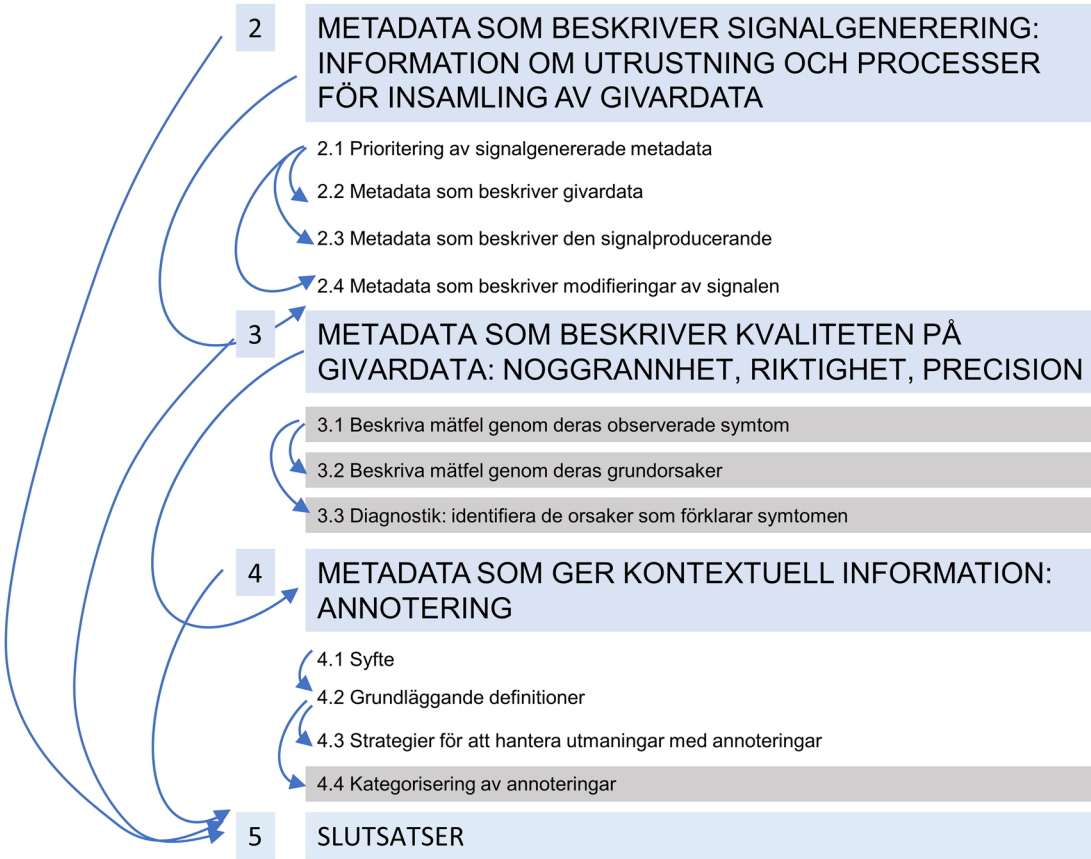
3.1 LÄSANVISNINGAR

Huvudsyftet med detta kapitel är att ge en översikt och definitioner av olika typer av metadata och hur de kan användas vid ett typiskt dataflöde (*data pipeline*). Tre kategorier av metadata beskrivs i kapitlet:

- Metadata som beskriver signalgenerering, inklusive datainsamling och lagring.
- Metadata för att bedöma kvaliteteten på givarsignaler.
- Metadata som ger kontextuell information som är relevant för senare användning.

Illustrationen nedan visar kapitlets struktur. Avsnitt som är markerade med blått eller vitt är viktiga för att kunna följa med i resten av rapporten, särskilt när det gäller (a) strukturering och lagring av data (Kapitel 4) och (b) generering av metadata från omfattande givarvalidering (Kapitel 5) eller dataanalytisk signalvalidering (Kapitel 6). De avsnitt som är markerade med grått ger

© 2025 IWAP. Detta är ett Open Access-bokkapitel som distribueras enligt villkoren i Creative Commons Attribution License (CC BY-NC-ND 4.0) som tillåter kopiering och vidaredistribution för icke-kommersiella ändamål utan derivat, förutsatt att arbetet citeras korrekt (https://creativecommons.org/licenses/by-nc-nd/4.0/). Kapitlet är från boken *Bästa möjliga data från givare i avloppsreningsprocesser*, Kris Villez, Daniel Aguado, Janelcy Alferes, Queralt Plana, Maria Victoria Ruano, Oscar Samuelsson (redaktörer).



bakgrundskunskaper för att förstå detaljerna i rutiner för givarvalidering (Kapitel 5) och dataanalytiska metoder för att bedömning av datakvalitet (Kapitel 6).

3.2 METADATA SOM BESKRIVER SIGNALGENERERING: INFORMATION OM UTRUSTNING OCH PROCESSER FÖR INSAMLING AV GIVARDATA

I detta avsnitt beskrivs olika typer av metadata som beskriver de rutiner som används för att generera data och de resultat som produceras genom dessa rutiner. Sådana rutiner kallas på engelska för “data acquisition, data collection, data pipelines, data-generating processes”. Detta inkluderar dokumentation av de rutiner som används (t.ex. dataflöden) samt de metadata som produceras genom sådana rutiner (t.ex. en kalibreringskurva). I princip kan de metadata som behandlas i detta avsnitt beskrivas innan några signaldata har samlats in. Detta är den största skillnaden mellan de metadata som beskrivs i detta avsnitt jämfört med de metadata som beskrivs i Avsnitten 3.2 och 3.3. I praktiken kan informationen ändras under inköp, installation och driftsättning av givare. Regelbunden verifiering av dessa metadata kan därför bidra till att säkerställa att de är aktuella. Även om denna typ av metadata ofta är begränsade är de enkla att samla in och lagra.

Nästa avsnitt inleds med en översikt över alla metadata som hör till kategorin signalgenerering, vilket illustreras med en syregivare och en pH-givare. I samma exempel grupperas de olika typerna av metadata utifrån prioritet. Slutligen beskrivs var och en av de olika metadatatyperna i detalj.

3.2.1 Prioritering av signalgenererade metadata

Tabell 3.1 ger en översikt över typiska metadata som kan samlas in för att beskriva hur en signal genereras hos en syregivare och en pH-givare. Tabellen indikerar hur ofta denna typ av metadata bör granskas och uppdateras, samt en prioritetsnivå utifrån följande beskrivningar:

- Nivå 1 – Grundläggande: En signal kan inte förstås utan denna information. Att samla in denna typ av metadata efter datainsamlingen är svårt. Därför bör denna nivå av metadata ha högsta prioritet.
- Nivå 2 – Avancerad: Dessa metadata är mycket användbara för validering av givare, särskilt när man använder automatiserade verktyg för kvalitetssäkring och kontroll av data såsom beskrivs i Kapitel 6.
- Nivå 3 – Framtidssäkrad: Dessa metadata är särskilt användbara i fall där den framtida användningen av de insamlade uppgifterna kan skilja sig väsentligt från de nuvarande ändamålen och syftena.

De olika nivåernas olika prioritet baseras på kapitelförfattarnas egen uppfattning. Beroende på syftet med datainsamlingen kan andra prioriteringar vara mer lämpliga på olika reningsverk. En rekommendation är att diskutera prioriteringar inom organisationen och med potentiella intressenter innan slutlig metod väljs för att samla in och hantera metadata.

3.2.2 Metadata som beskriver givardata

De metadata som beskrivs här ger den grundläggande information som krävs för att förstå sammanhanget och ge mening åt de värden som produceras av en givare. I följande avsnitt listas metadata som tillhör denna kategori. Varje typ av metadata definieras och motiveras separat med exempel i de fall det är relevant.

3.2.2.1 Tidsstämpelns ursprung

Definition D3.1: **Tidsstämpelns ursprung** är den tekniska enhet som producerar tidsstämpeln i en datapost (se Kapitel 2, Avsnitt 2.4.1.2).

Motivering: Kunskap om tidsstämpelns ursprung kan vara till stor hjälp vid diagnostisering av situationer där olika instruments eller styrenheters klockor inte överensstämmer. Till exempel kan vissa enheter aktivera sommartid medan andra inte gör det. Även om det finns standarder för tidsstämplar för att förhindra feltolkningar (ISO 2019a, *b*), är tillämpningen ofta begränsad i praktiken, vilket minskar förtroendet för de data som produceras eller ökar kostnaderna för deras användning.

Ytterligare kommentar: I instrument som är utrustade med digital kommunikation produceras tidsstämpeln ofta från dess interna klocka. I instrument som är anslutna via en analog anslutning skapas tidsstämpeln istället vanligtvis av det system som är konfigurerat för dataloggning (t.ex. SCADA eller DCS).

3.2.2.2 Tidsupplösning eller samplingsfrekvens

Definition D3.2: **Tidsupplösningen** är tidsintervallet mellan efterföljande dataposter som produceras av en givare.

Motivering: Tidsupplösningen är ofta avgörande för om givaren kommer att fånga upp vissa fenomen. Om ett intressant fenomen inträffar inom en kortare tidsram än givarens tidsupplösning, kanske det inte registreras eller mäts på ett dåligt sätt.

Ytterligare kommentar: Tidsupplösningen är normalt konstant för givare som drivs med permanent strömförsörjning. Givare som drivs med batterier eller intermittent strömförsörjning (t.ex. solenergi) kan dock ha varierande tidsupplösningar. I detta fall kan det vara användbart att känna till den lägsta tidsupplösning (dvs. det kortaste tidsintervallet) som är acceptabel med tanke på batteriet eller de data som givaren producerar.

Tabell 3.1 Typiska metadata för signaldokumentation.

Avsnitt	Typ av Metadata	Exempel		Granskningsfrekvens	Prioritering
3.2.2	**Metadata som beskriver genering av en signal**				
		Syregivare	pH-givare		
3.2.2.1	*Ursprung för tidsstämpel*	Analog-till-digital-omvandlare integrerad i PLC-enheten	SCADA-system	År	Grundläggande
3.2.2.2	*Tidsupplösning*	15s	1s	År	Avancerad
3.2.2.3	*Numerisk upplösning*	0.3 × 10-3(16-bit)	Ej tillgänglig	År	Framtidssäkrad
3.2.2.4	*Givarens placering*	Luftningsbassäng 3 i zon 2	Utlopp försedimentering	Månader - År	Grundläggande
3.2.2.5	*Mätenhet*	g/m^3	Ej tillämpligt (dimensionslöst)	År	Grundläggande
3.2.2.6	*Mätområde*	0–20	2–12	År	Grundläggande
3.2.2.7	*Variablens intervall*	2–6	5–7	Månader - År	Avancerad
3.2.2.8	*Signalöverföring (analog/ digital/manuell)*	Digital: Modbus	Analogt: 4-20 mA	År	Avancerad
3.2.2.9	*Mätprincip*	Luminescensbaserad givare	Jon-selektiv elektrod	År	Framtidssäkrad
3.2.2.10	*Syfte med mätningen*	Automatisk reglering av koncentration av syrehalt.	Varning för industriellt/ toxiskt inflöde	Månader - År	Framtidssäkrad
3.2.3	**Metadata som beskriver den signalproducerande enheten (givaren)**				
3.2.3.1	*Produktspecifikationer och hänvisning/länk till lagrad manual, datablad och/eller installationsguide*			År	Avancerad
3.2.3.2	*Tillverkare, modell namn och serie*	Hach, 9020000, 9FT3K45G	E + H, CPS11D, R2EKGEQT	År	Grundläggande
3.2.3.3	*Unik identifierare*	DO20210001	PH18065	År	Grundläggande
3.2.3.4	*Enhetens historik*	*Referens/länk till inköpsorder, installationsprotokoll, historisk givare kalibreringskurvor och valideringsresultat*		Veckor - Månad	Avancerad
3.2.4	**Metadata som beskriver modifiering av en signal**				
3.2.4.1	*Ändring av enheter*	Ej tillämpat	Ej tillämpat	År	Avancerad
3.2.4.2	*Modifieringar med förlust (aggregering, medelvärdesbildning, filtrering, komprimering)*	Ej tillämpat	Medianfilter applicerat på fem konsekutiva mätningar för att undvika falsklarm *(lägg till en hänvisning till kod)*	Månader - År	Framtidssäkrad
3.2.4.3	*Förbättring av data interpolation,* känt värde för att säkerställa kontinuerlig *synkronisering)*	Saknade data ersätts med det senaste mätvärdet None *(imputation,* drift av styrenheten	Internt: Nernst-ekvationen används för att beräkna pH från	Månader - År	Framtidssäkrad
3.2.4.4	*Fusion av data*	Ej tillämpat	råa potential- och temperaturmätningar	Månader - År	Framtidssäkrad

Anmärkning om multivariata givare: Multivariabla instrument eller multigivare kan innehålla flera mätelement i ett enda hölje, som vart och ett registrerar en annan fysisk storhet. Beroende på enheten kommer vissa metadata att vara desamma för alla signaler som produceras av givaren (t.ex. tidsupplösning, plats), medan andra kan vara olika (t.ex. mätenhet, mätområde). Se även de relaterade kommentarer i Avsnitt 2.4.1.5 i Kapitel 2.

3.2.2.3 Numerisk upplösning (kvantisering)

Definition D3.3: **Den numeriska upplösningen** är den minsta ökning av den uppmätta variabeln som kan registreras digitalt (dvs. analog-till-digital-omvandling). Inom signalbehandlingsområdet benämns detta som kvantisering.

Motivering: Kvantiseringens effekter är ofta försumbara. Att känna till den numeriska upplösningen kan dock hjälpa till att bedöma datakvaliteten och att definiera lämpliga prestandamått för modellbaserade eller algoritmiska förutsägelser. I allmänhet kan man inte förvänta sig att en förutsägelse ska vara mer exakt än själva mätningens upplösningen.

Ytterligare noteringar: På grund av den binära karaktären hos moderna dator- och datalagringssystem lagras datapunkter i digitalt. Digitalisering av signaler innebär att man använder en diskret (d.v.s. ej kontinuerlig) representation för både tidsstämpeln (digitaliserad) och datapunkten (kvantiserad). Den numeriska upplösningen är den parameter som beskriver denna process. Effekten av avrundning kallas även kvantiseringsfel (se Avsnitt 3.3.1.2).

3.2.2.4 Givarens placering

Definition D3.4: **Givarens placering** är den fysiska positionen av givaren (givarens mätelement).

Motivering: Att känna till en givares exakta position är nyckeln till signaltolkning. Dessutom har det flera fördelar vid bedömning och kontroll av datakvalitet.

- Det kan hjälpa till att upptäcka, isolera och diagnostisera avvikande datapunkter som kan uppstå till följd av underhållsåtgärder, t.ex. rengöring eller kalibrering av givare.
- Det kan ge ett sammanhang till de registrerade datapunkterna. De datapunkter som produceras av en ammoniumgivare i en luftningsbassäng täcker t.ex. vanligtvis ett annat mätintervall än en givare i avloppsvattenreningssystemets utflöde.
- Det kan underlätta att övervaka givarens omgivning, särskilt genom att beskriva och kvantifiera exponering vid tuffa mätmiljöer såsom partiklar och biofilmspåväxt.

Ytterligare noteringar: En givares position kan beskrivas på minst tre olika sätt:

(1) **Position i förhållande till en fysisk referenspunkt.** I form av en text som beskriver den relativa positionen i förhållande till anläggningen. Till exempel: *i zon 5 i luftningsblock 2* eller *10 m från returslamröret*. Sådana beskrivningar kan innehålla relativa positionsangivelser, t.ex. *mitt i bassängen*, *djup* $= 2\,m$, eller *vid utloppet*. Sådan här information är enkel att formulera, diskutera och tolka för en mänsklig operatör. Den kan dock kräva ytterligare metadata (t.ex. ett flödesschema eller andra processdiagram) för att underlätta en effektiv automatiserad dataanalys.

(2) **Position i förhållande till konstruktionsscheman.** Som en del av ett P&ID-diagram ("Piping and Instrumentation Diagram") eller liknande ritningar. Modern programvara för datorstödd design av P&ID-diagram gör det möjligt att utöka eller länka element i ritningen med metadata som beskriver de avbildade enheterna för att indikera givarnas lokalisering. Sådana filer kan underlätta automatisk tolkning vid dataanalys (Udugama *et al.*, 2021).

(3) **Geografisk koordinater.** Som en uppsättning geografiska koordinater. Detta format kan vara särskilt användbart för mobila givare, t.ex. för givare på drönare (nedsänkta, flytande eller flygande) eller robotarmar. Koordinaterna kan produceras genom ett globalt refererat system (t.ex. GPS) eller ett lokalt refererat system, t.ex. avståndsmätningar baserade på högupplösta kameror, laserljus eller radiosignaler (t.ex. Bluetooth, RFID). Ett lokalt refererat system för positionsmätning kommer naturligtvis att kräva sina egna metadata för att säkerställa noggrannhet och tillit till de geografiska koordinater som det producerar.

Även om de flesta givare har en fast position under långa perioder, kräver underhållsåtgärder ibland en ändring av givarens exakta position. Ett system för positionsmätning kan möjliggöra automatisk loggning av tidsperioder då data inte var representativa för den avsedda mätplatsen. Dessutom har ny

forskning undersökt användningen av mobila givare, oavsett om de är nedsänkta, flytande, flygande eller kopplade till robotsystem. Automatisk spårning av givarens position bör erbjuda tydliga fördelar jämfört med system som kräver mänsklig beskrivning av dess placering.

Vissa nya givarsystem, t.ex. de drönare som nämns ovan förflyttar, av naturliga skäl, sin mätposition kontinuerligt. För sådana givare kan metadata lagras som tidsserier, vilket kan göras med ett liknande tillvägagångssätt som för signaldata (se Kapitel 2).

Slutligen så kräver placeringen av redundanta instrument särskild omsorg. Den första formen av positionsmetadata som beskrivs ovan (position i förhållande till anläggningskomponenter) är känslig för att flera givares position rapporteras som densamma. För vissa uppgifter eller projekt kan en korrekt differentiering mellan den exakta placeringen av redundanta givare vara av stor betydelse för en korrekt tolkning av data.

3.2.2.5 Mätenhet

Definition D3.5: **Mätenheten** är det valda standardmåttet med vilket storleken på givaravläsningen uttrycks.

Motivering: Mätenheten ger en fysisk innebörd åt det numeriska värde som genereras av givaren. Genom att lagra enheten som metadata säkerställs att data kan tolkas korrekt.

3.2.2.6 Mätområde

Definition D3.6: **Mätområdet** är det intervall som definieras av de minimi- och maximivärden som kan produceras av en givare som är korrekt konfigurerad och används på rätt sätt.

Motivering: Att känna till mätområdet hjälper till att upptäcka och isolera felaktigt konfigurerade givare. Ett värde utanför mätområdet kan bland annat bero på skadade delar (t.ex. givarmembran), felaktig programmering av förhållandet mellan en analog signal och mätvariabelns mätområde i DCS eller SCADA, eller felaktig kalibrering.

Ytterligare noteringar: Observera att mätområdet kan skilja sig från den analoga signalens mätområde (t.ex. kan en pH-givare konfigureras att mata ut en 4–20 mA-signal för ett linjärt motsvarande pH-område mellan 4 och 10, även om mätområdet är 0–15). Mätområdet beror endast på givaren och inte på dess placering eller mättidpunkten. Mätområdet bör vara större än det värdeintervall som förväntas för en viss plats på anläggningen, tid på dygnet eller årstid (variabelt område, se nedan).

3.2.2.7 Variabelns fysiska intervall

Definition D3.7: **Variabelområdet** är det intervall som definieras av det minimi- och maximivärde som är acceptabelt eller typiskt observerat under normal drift, antingen i allmänhet eller i ett visst sammanhang (t.ex. plats och/eller tid).

Motivering: Att känna till variabelns fysiska intervall kan underlätta detektion, isolering och diagnosticering av funktionsfel. En datapunkt utanför det variabelns normalområde kan bero på ett fel i utrustningen, processen (t.ex. toxisk belastning) eller givarna (t.ex. igensättning).

Ytterligare noteringar: Observera att variabelområdet beskriver den övervakade processen snarare än själva givaren och kan definieras för en specifik processtyp (t.ex. försedimentering) eller specificeras för en viss anläggning, processenhet, tid på dygnet och/eller säsong.

3.2.2.8 Signalöverföring (analogt/digitalt/manuellt)

Definition D3.8: **Signalöverföringssättet** är det sätt på vilket en signal som produceras av en givare överförs till den fysiska lagringsenheten (t.ex. PLC, SCADA).

Motivering: Att förstå aktuella kommunikationsprotokoll och deras konsekvenser för datakvalitet och datahantering är värdefullt med tanke på genereringen av tolkningsbara och framtidssäkrade dataset. Att känna till vilket sätt signalen överförs på underlättar också felsökning av kommunikationsfel.

Ytterligare noteringar: Automatiserad, högfrekvent signalkommunikation kan grovt sett delas in i två kategorier:

(1) *Analog överföring:* Det uppmätta värdet överförs via en kontinuerlig elektrisk ström (ofta varierande mellan 4 och 20 mA) som kan tolkas av system nedströms givaren (t.ex. PLC, SCADA). Den här typen av överföring är enkel, vilket underlättar underhållet. Klassiska analoga signaler tillåter dock inte överföring av metadata, t.ex. givarens interna status. Särskilda försiktighetsåtgärder måste vidtas för att undvika interferens med andra signaler (t.ex. radio) vid överföring över långa avstånd. Protokoll som HART (*highway addressable remote transducer*) gör det möjligt att lägga till kompletterande digital information genom frekvensmodulering (HART Communication Foundation, 2013).
(2) *Digital överföring:* Mätvärdet digitaliseras först (A/D, analog-till-digital-omvandling) och överförs sedan digitalt via digital nätverkskabel, fiberoptisk kabel eller trådlöst nätverk med hjälp av ett definierat kommunikationsprotokoll. Digital överföringsmetodik kan kräva vidareutbildning av personal om traditionell analog överföring använts tidigare. Digital överföring erbjuder en rad fördelar såsom automatisk loggning av ytterligare information, t.ex. givarens interna status (t.ex. kalibreringsläge, batteristatus) eller parametrar (t.ex. kalibreringsdata, kalibreringskurva). Modbus-, Profibus- och Profinet-kommunikation är alla digitala kommunikationsprotokoll. En utmaning som är unik för digital överföring är att tidsstämplarna på givarenheterna kan vara osynkroniserade med PLC:ns tidsstämplar (se Avsnitt 3.2.2.1).

Utöver de automatiska överföringsmetoderna ovan är det också möjligt att manuellt mata in data:

(1) *Manuell överföring:* Det uppmätta värdet kopieras från en enhet (t.ex. en handhållen enhet) eller ett skriftligt dokument (t.ex. ett kalibreringsblad på papper eller kalkylark) och skrivs in i en enhet eller ett system för långtidslagring, vanligtvis via ett särskilt grafiskt användargränssnitt (GUI). Åtminstone ett steg i överföringen kräver manuell hantering.

När flera olika överföringssätt används rekommenderas att antingen:

(1) Beskriva det kompletta dataflödet utifrån de olika kommunikationsvägarna som används.
(2) Beskriv det mest tidskrävande och minst informativa kommunikationssättet i dataflödet (i följande ordning): manuell överföring, analogt, digitalt.

3.2.2.9 Mätprincip

Definition D3.9: En givares **mätprincip** är en beskrivning av det grundläggande fysikalisk-kemiska förhållandet mellan den intressanta variabeln och den råsignal som givaren producerar.

Motivering: Kunskap om den tillämpade mätprincipen och dess konsekvenser för signalkvalitet och underhållskostnader är nyckeln till adekvat kvalitetssäkring och kvalitetskontroll av data.

Ytterligare noteringar: Alla mätprinciper är inte lika korrekta eller exakta. Dessutom sammanfaller skillnader i mätprincip ofta med skillnader i installationskrav (t.ex. djup, vinkel), krav på givarunderhåll (t.ex. frekvens för byte av delar) och mätområden. Dagens ammoniumgivare baseras på jonselektiva elektroder och mäter t.ex. inte exakt under 1 mg NH_3-N/L. Däremot anses en ammoniumanalysator som använder den kolorimetriska principen vara noggrann under 1 mg NH_3 -N/L.

3.2.2.10 Syfte med mätningen

Definition D3.10: **Syftet** med mätningen beskriver hur signalen är tänkt att användas.

Motivering: En beskrivning av en givares ursprungliga och senare syften kan underlätta prioriteringen av underhållsåtgärder och hjälper en dataanalytiker att avgöra om dataposter som produceras av en givare är användbara för en viss uppgift.

Ytterligare noteringar: Signaldata samlas ofta in för ett specifikt syfte (t.ex. för övervakning av en process) och används senare för ett annat (t.ex. modellering, automatisering). Olika syften kräver olika metoder för kvalitetsbedömning och kvalitetskontroll, vilket leder till betydande skillnader i den resulterande datakvaliteten. Syftet med mätningen kan också indikera om en viss mätning är avsedd att vara tillfällig, t.ex. som en del av en mätkampanj, eller om den är permanent.

3.2.3 Metadata som beskriver den signalproducerande enheten (instrumentet)

De metadata som rör instrumentet innehåller information om själva enheten, dess mätprincip, tekniska egenskaper och historik.

3.2.3.1 Produktspecifikationer och guider

Definition D3.11: **Produktspecifikationer** och guider består av dokumentation som beskriver installation, användning och underhåll av ett instrument.

Motivering: Produktspecifikationer och guider hjälper till att säkerställa att korrekt datakvalitetsbedömning och kvalitetskontroller finns på plats och att givaren installeras och används på avsett sätt.

Ytterligare noteringar: Information som beskriver korrekt installation, användning och underhåll av givaren tillhandahålls vanligtvis av tillverkaren. Denna information bör samlas in vid inköpstillfället för att undvika inkonsekvenser på grund av senare ändringar som återspeglar produktuppdateringar. Information kan finnas i ett eller flera dokument eller elektroniska filer, inklusive:

- Specifikationsblad
- Installationsguide
- Handbok för operatören
- Datablad
- Guide för snabbstart
- 3D-konstruktionsfil (t.ex. CAD)

Viktiga uppgifter att hålla utkik efter är bland annat

- Erforderlig strömförsörjning
- Enhetens IP-klassning (*Ingress Protection*) eller krav på damm- eller vattentät installation
- Tillgängliga kommunikationsprotokoll (analoga, digitala)
- Antal inkommande och utgående kanaler
- Inspektions- och underhållskrav, inklusive byte av delar (lock, filter, skarv, membran) och förbrukningsvaror (t.ex. kemiska reagens)

3.2.3.2 Tillverkare, modellnamn och serie

Definition D3.12: **Tillverkningsinformation** består av instrumentets märke eller fabrikat, modellnummer och serienummer.

Motivering: Genom att ha tillgång till märke, modell och serienummer som tillverkaren har producerat kan man söka efter manualer och specifikationer hos tillverkaren. Att ha tillgång till den referens som tillverkaren använder underlättar dessutom snabb beställning av nya delar eller reservdelar med samma eller motsvarande egenskaper.

3.2.3.3 Unik identifierare

Definition D3.13: **En unik identifierare** (*unique identifer*) är ett namn eller nummer som är unikt för ett visst instrument eller en viss komponent i ett instrument, vilket möjliggör en detaljerad beskrivning av signalkällan.

Motivering: Det blir lättare att hålla reda på alla olika givare om man registrerar den information som unikt identifierar en givare och alla delar som har modifierats eller bytts ut. Det gör det möjligt

att knyta samman flera olika delar av informationen, inklusive tillverkarinformation, händelser under givarens livscykel och de data som produceras av instrumentet. Slutligen möjliggör det en sökbar översikt över alla instrument som finns tillgängliga i en anläggning.

Ytterligare noteringar: En unik identifierare kan ofta skapas genom att kombinera fabrikat (tillverkare), modellnummer och serienummer (se ovan). Enskilda delar kan också identifieras på samma sätt.

3.2.3.4 Instrumentets historia

Definition D3.14: **Instrumentets historik** består av all dokumentation som beskriver viktiga händelser i den signalproducerande enhetens historia.

Motivering: En väldokumenterad givarhistorik underlättar uppföljning av givarhanteringen vilket kan underlätta framtida val av givare för inköp, bidra till optimerad kalibrering, validering och verifiering av givare. Det kan också vara till stor hjälp för att bedöma kvaliteten på givarsignalerna, eftersom man kan få information om problem som är svåra att upptäcka, t.ex. kalibreringsfel, slitage på givaren och symtom på drift.

Ytterligare noteringar: Enhetens historik omfattar dokumentation av viktiga delar av en givares livscykel, inklusive men inte begränsat till:

- Tidpunkten för inköp, leverans, driftsättning och avveckling.
- Tidsperioder när givaren ur drift.
- Alla underhållsåtgärder som utförs av tillverkarens eller egen personal, t.ex. rengöring, kalibrering på plats eller i fabrik, validering, testning och byte av givare eller delar.
- En historik över kalibreringskurvor och/eller de mätningar som används för att ta fram sådana kurvor.

I Kapitel 5 beskrivs hur metadata inom denna kategori kan genereras, särskilt produktionen av valideringsdata genom referensmätningar.

Notering om kalibreringskurvor: I det vanligaste scenariot resulterar givarkalibrering i en ändring av kalibreringskurvan som är lagrad på en elektronisk enhet, t.ex. en mätförstärkare, transformator eller transmitter. I många fall är endast den aktuella kalibreringskurvan tillgänglig på enheten. Många moderna instrument har dock numera automatisk loggning av kalibreringshistoriken. Dessutom är det ofta möjligt att exportera de aktuella och historiska kalibreringskurvorna från enheten till ett annat system, t.ex. till en historikdatabas.

3.2.4 Metadata som beskriver modifieringar av signalen

De data som produceras av givaren och överförs nedströms till en PLC, SCADA, historikdatabas eller någon annan databas kan modifieras avsevärt under processen. Orsakerna till att den ursprungliga signalen modifieras före lagring kan vara flera, såsom:

- Ändring av den ursprungliga måttenheten till en SI-enhet eller en accepterad standard.
- Rensning av oönskade egenskaper i signalen, t.ex. extremvärden, toppar eller interpolering, för att minska förekomsten av frånvarande och/eller förlorade data, antingen genom manuella eller automatiserade metoder (se Avsnitt 3.2.4.2 och 3.2.4.3).
- Synkronisering av tidsstämpeln med en referensklocka. Tidssynkronisering efter datagenerering bör undvikas i så stor utsträckning som möjligt. I en ideal värld nyttjar alla klockor i ett datainsamlingssystem samma referensklocka.

För att göra både ursprungliga och modifierade dataposter framtidssäkra bör irreversibla förändringar (t.ex. brusreduktion (*denoising*), borttagning av avvikande extremvärden (*outliers*)) hanteras genom att lagra den ursprungliga signalen utöver den modifierade signalen. På så sätt kan en dataanalytiker använda omodifierade data vid behov. Reversibla förändringar (t.ex. byte av enhet, ändring av tidsstämpel) kräver inte sådan duplicering så länge som själva modifieringen registreras på lämpligt sätt. Genom att använda de typer av metadata som beskrivs nedan kan man se till att en

framtida användare kan förstå signalmodifieringens inverkan på den slutliga dataströmmen och, om det är möjligt och användbart, återställa de genomförda ändringarna. I Avsnitt 2.6 i Kapitel 2 beskrevs strategier och verktyg för att lagra olika versioner av samma data.

Noteringar om tidpunkt för tillgänglighet: Observera att detta avsnitt handlar om de *procedurer* som används för att producera en modifierad version av de ursprungliga dataposterna. I princip kan dessa dokumenteras innan de används på data.

Detta gäller oavsett (a) om förfarandet är automatiserat eller manuellt och (b) om det utförs i realtid, online eller offline.

3.2.4.1 Förändring av enheter

Definition D3.15: En **ändring av enheter** är en avsiktlig ändring av enheter från den måttenhet som tillhandahålls av instrumentet till en enhet som anses lämplig för långtidsförvaring och användning.

Motivering: Dokumentation av alla enhetsändringar mellan givare och datalagring undviker problem med inkonsekventa enheter och kan spara avsevärd tid i datakrävande projekt.

Ytterligare noteringar: En vanlig transformation är omvandlingen av en mätning till en annan enhet än den som används av givaren. Denna omvandling kan ske genom att applicera en enkel omvandlingsfaktor (t.ex. omvandling av gram per liter till milligram per liter). Omvandlingen kan dock ibland innebära ytterligare antaganden om mätsammanhanget. Ett exempel är omvandlingen av en TSS-mätning som givaren gör i nefelometriska turbiditetsenheter (NTU) till den bedömda ekvivalenta koncentrationen av fasta ämnen uttryckt i milligram per liter. I ett sådant fall ändras den fysiska innebörden av det uttryckta värdet med omvandlingen, vilket innebär att omvandlingen kräver en korrelationsmodell. I sådana fall bör man åtminstone lagra den tillämpade modellen och historiska versioner av den. Om det är möjligt bör man också lagra den omodifierade signalen.

3.2.4.2 Modifiering med informationsförlust (aggregering, medelvärdesbildning, filtrering, komprimering)

Definition D3.16: En **modifiering** av en signal som innebär förlust av information är en procedur som omvandlar originalsignalen till en derivatsignal så att originalsignalen inte kan rekonstrueras perfekt från derivatsignalen, dvs. viss information går förlorad i processen.

Motivering: Dokumentation av förlustmodifieringar av en signal kan hjälpa en dataanalytiker att avgöra om en viss signal är lämplig för ändamålet.

Ytterligare noteringar: För att förstå modifieringar med förlust kan man först tänka sig en ändring av enheter enligt ovan. En sådan ändring är reversibel eftersom man alltid kan ändra tillbaka till den ursprungliga enheten, förutsatt att proceduren är väl dokumenterad. Vi kallar det därför en förlustfri modifiering. Modifieringar med informationsförlust leder däremot till förändringar som inte kan göras ogjorda. Att till exempel beräkna ett timmedelvärde från en signal som mäts varje minut är irreversibelt. Det går inte att beräkna den ursprungliga signalen med hjälp av timmedelvärdena. Av denna anledning betraktas detta som en irreversibel eller förlustbringande modifiering. Om inte originalsignalen och den modifierade signalen lagras separat kommer informationen i originalsignalen att gå förlorad utan möjlighet till återskapande. Modifiering med förlust kan också medföra oönskade artefakter i den modifierade signalen. Icke-linjära komprimerings- eller filtreringsmetoder kan t.ex. medföra svängningar vid frekvenser som inte ursprungligen fanns i en signal.

Typer av irreversibla databehandlingssteg inkluderar filtrering (oftast för brusreduktion), nedsampling (t.ex. att behålla var tionde datapost i en originalsignal) eller aggregering (t.ex. att beräkna dagliga eller veckovisa medelvärden). Modifieringar med förlust kan ske på många ställen i ett dataflöde, t.ex. i instrumentet, vid inmatning i en databas, vid sökning från en databas eller vid dataanalys. Se även Avsnitten 4.4.4 och 4.4.5 i Kapitel 4 som beskriver *lossless*- och *lossy*-modifieringar som ofta används i historikdatabaser. Lossy-modifieringar har ofta ett antal inställningsparametrar (t.ex. fönster för medelvärdesbildning). En fullständig beskrivning av lossy-modifieringar inkluderar historiken för de tillämpade värdena för inställningsparametrarna.

Notering om smarta givare: Många givare har egen beräkningskapacitet och kan tolka den råa signalen på en mängd olika sätt. Databehandling på enheten kan användas för kalibrering, validering, brusreduktion och borttagning av avvikande värden. Moderna givare kan också producera ny information baserat på en modell (*soft-sensing*). Till exempel kan UV-VIS-spektrometrar vara utrustade med en modell som beskriver förhållandet mellan den registrerade signalen (absorbansspektrum) och en variabel av intresse (t.ex. COD, nitrat). Tyvärr är information om hur databehandlingen i enheten fungerar ofta inte tillgänglig för användaren. I vilket fall som helst bör behovet av att få tillgång till information på enheten helst diskuteras med tillverkaren innan givaren köps.

3.2.4.3 Ersättning av data (imputation, interpolering, synkronisering)

Dataimputation innebär att nya, datorgenererade dataposter ersätter frånvarande eller saknade data i en signal. Detta omfattar tekniker för uppsampling (t.ex. interpolering) och imputering (t.ex. ersättning av frånvarande eller förlorade data). De flesta imputeringstekniker använder ett interpolerat värde för att skatta frånvarande eller förlorade data (se Kapitel 6). Mer avancerade metoder består av modellbaserad imputering, t.ex. tekniker som bygger på Kalman filter eller specialiserade tidsseriemodeller (se t.ex. Alavi *m.fl.*, 2006; Ba-Alawi *m.fl.*, 2022; De Mulder *m.fl.*, 2018; Osman *m.fl.*, 2018). På samma sätt kan två signaler med asynkrona tidsstämplar synkroniseras genom enkel interpolering eller modellbaserade metoder.

Observera att imputerings- och interpoleringssteg kan vara irreversibla om de är dåligt dokumenterade. Det kan vara svårt att avgöra vilka dataposter är originaldata och vilka som lagts till senare, vilket gör det omöjligt att återskapa originalsignalen. Det är i allmänhet att föredra att den modifierade datasignalen lagras separat medan originalsignalen hålls orörd. Det är önskvärt att metadata för den modifierade signalen innehåller de eventuella inställningsparametrar som användes vid modifieringen.

Motivering: Eftersom modifiering innebär att artificiella dataposter läggs till i datasetet kommer sannolikt detta att påverka datasetets statistiska egenskaper. Slutanvändaren av data behöver därför vara medveten om att datasetet producerats med visas modifieringar, främst för att kunna lita på data i lagom omfattning. Om så inte är fallet kan det leda till övertro på de databaserade resultaten eller besluten.

3.2.4.4 Datafusion och softsensorer

Många intressanta signaler erhålls genom en kombination av flera signaler. Ta en pH-mätning som ett exempel. Vanligtvis mäts inte pH-värdet direkt. Istället registrerar givaren en temperatursignal och en elektrodpotential. Med hjälp av Nernst ekvation, en enkel modell som relaterar pH, potential och temperatur till varandra, kan man beräkna pH-värdet. När det gäller pH-mätningar sker denna datafusion inuti mätutrustningen. Parametrarna i Nernstekvationen lagras normalt i själva instrumentet och kräver vanligtvis en särskild insats för att bli tillgängliga för dataanalys. I andra fall kan en modell byggas av slutanvändaren. I så fall kallas metoden för att beräkna den nya signalen ofta för en *softsensor* (*software sensor* till skillnad från en *hardware sensor*). Ett exempel är beräkningen av ett massflöde (massflöde för fasta ämnen) baserat på en flödesmätning och en koncentrationsmätning (t.ex. TSS). Man bör överväga att spåra historiken för softsensorer, särskilt om den utplacerade modellen uppdateras regelbundet (se även Avsnitt 3.2.3.4).

Motivering: Att förstå hur flera signaler kombineras för att producera en signal av intresse kan hjälpa till att diagnostisera en felaktig softsensor. Dokumentationen beskriver då vilka av de relevanta givarna som behöver prioriteras för visuell inspektion eller underhåll.

3.3 METADATA SOM BESKRIVER KVALITETEN PÅ GIVARDATA: NOGGRANNHET, RIKTIGHET, PRECISION

För att bedöma hur användbara data är för ett givet ändamål är det viktigt att signalkvaliteten bedöms och att mätfel identifieras och, där så är lämpligt, kvantifieras och korrigeras. Mätfel har flera orsaker, bland annat fel i hårdvara, den mänskliga faktorn och förändringar i den mätmiljön.

Ett tydligt exempel på de problem som orsakas av en signal med otillräcklig kvalitet är när mätningen används för att styra en processvariabel. Vid t.ex. pH-reglering kommer en pH-givare som reagerar långsamt på en förändring av pH-värdet i bulk att innebära långsamma förändringar i den manipulerade variabeln, vilket leder till ett fördröjt svar på förändringar i pH-värdet. Detta kan leda till försämrad processprestanda och längre tid för att nå nya börvärden.

Mätfel kan ha flera orsaker. Ett kalibreringsfel kan t.ex. orsaka en känslighet (*sensitivity error*) (se Kapitel 5) som är lägre än den verkliga givarens känslighet, så att de uppmätta variationerna i pH-värdet är mindre än de som finns i det verkliga systemet. Det kan också bero på att givaren har skadats. Ett kalibreringsfel kan påverka noggrannhet och precision, men sällan responstiden. Däremot påverkar nedsmutsning och skalning vanligtvis både precisionen och svarstiden. Det är därför viktigt att beskriva mätfelen i detalj för att kunna isolera grundorsaken.

Antalet tillgängliga tekniker för bedömning av datakvalitet, inklusive detektion och identifiering av fel, diagnos och kontroll av data, är många, se t.ex. (Venkatasubramanian, *2003a*, *b*, *c*). En utmaning i detta forskningsområde är dock att rotorsak och symptom ofta blandas ihop. För att kunna dra nytta av tillgängliga feldetektionsteknikerna är det bäst att tydligt skilja mellan identifierade orsaker till en avvikelse eller ett fel och de symtom som ledde till att de upptäcktes. Detta avsnitt syftar till att bidra till att undvika denna begreppsförvirring genom att tillhandahålla en lista över begrepp som kan användas för detta ändamål, med en skarp gränsdragning mellan

(a) en beskrivning av observerade mätfel och (b) potentiella grundorsaker. I detalj så beskrivs:

- Typiska åtgärder för att kvantifiera symptomen på mätfel (Avsnitt 3.3.1).
- Kända grundorsaker till mätfel för givare som används på reningsverk (Avsnitt 3.3.2).
- Den viktiga utmaningen med att använda kvantifierade symtom på mätfel för att identifiera grundorsaken till mätfel (Avsnitt 3.3.3).

3.3.1 Beskriva mätfel genom deras observerade symtom

I det här avsnittet introduceras terminologi som används för att kvantifiera symtom på förbättrad eller försämrad datakvalitet. Dessa används ofta för att utvärdera analytiska mätmetoder och tillämpas regelbundet på givarsignaler.

I avsnittet skiljer vi på beskrivningar av datakvalitet som utgår från statiska förhållanden och sådana som specifikt tar hänsyn till givardynamik eller varierande datakvalitet. I det första avsnittet definieras termen mätfel explicit. I de följande två avsnitten beskrivs grundläggande och avancerade beskrivningar av datakvalitet under statiska förhållanden. Det tredje avsnittet fokuserar på dynamiska symptom.

3.3.1.1 Fel som avvikelser från referensvärden

Ett fel definieras utifrån definitionen i Isermann och Ballé (1997):

Definition D3.17: **Ett fel** är en avvikelse mellan ett uppmätt eller beräknat värde och det sanna, specificerade eller teoretiskt korrekta värdet.

Mätfelet kvantifieras som skillnaden mellan det uppmätta värdet y och det sanna värdet y_{ref}

$$y - y_{\mathrm{ref}} \tag{3.1}$$

Det referensvärde som används för att utvärdera felet är ett noggrant mätvärde som erhållits genom en procedur som är grundad på vetenskaplig teori eller principer. En vanlig procedur är att placera givaren i ett medium för vilket det sanna värdet för den uppmätta variabeln är känt. En pH-givare kan t.ex. utvärderas genom att placera den i en lösning med känt pH-värde (t.ex. en buffrad lösning eller ett kalibreringsmedium) vid standardtryck och standardtemperatur. En annan procedur för att få en redundant, högkvalitativ referensmätning är att nyttja en betrodd mätmetod. Till exempel utvärderas tillförlitligheten hos en ammoniumjonselektiv givare vanligtvis genom jämförelse med en analytisk mätning av koncentrationen. De flesta standarder tar inte hänsyn till att referensmätningarna själva

kan vara behäftade med fel. I allmänhet förutsätter tillämpliga standarder att referensmätningarnas noggrannhet är flera storleksordningar bättre än den givarsignal som kontrolleras. I Kapitel 5, Avsnitt 5.5.2 beskrivs metoder för att genomföra referensmätningar av hög kvalitet som kan uppfylla detta krav.

3.3.1.2 Statisk beskrivning av datakvalitet - grunderna: noggrannhet, riktighet och precision

Grunden för att beskriva kvaliteten på givardata är noggrannhet (*accuracy*), bias (*bias*) och precision (*precision*). Dessa grundläggande begrepp beskrivs nedan och några exempel illustrerar hur de kan kvantifieras. För att bedöma noggrannhet, bias och precision krävs statiska (konstanta) förhållanden. I följande definitioner och metoder antas därför att flera felvärden finns tillgängliga inom en tidsperiod som är tillräckligt kort för att givarmätningarnas kvalitet ska vara oförändrad, dvs. man förutsätter att givaren förblir i samma tillstånd.

Noggrannhet, riktighet och precision utvärderas vanligen genom att jämföra flera mätningar som gjorts under samma förhållanden, y, med ett enda referensvärde, y_{ref}. Med N mätningar, y_i, fås N mätfel:

$$e_i = y_i - y_{\mathrm{ref}}, \quad i = 1,2,\ldots N \tag{3.2}$$

Genom att skapa förutsättningar för att ett enda referensvärde ska kunna antas ges också förutsättningarna för en bra kvantifiering av noggrannhet, riktighet och precision. Det innebär ofta att en givare måste isoleras från mätstörande faktorer (*confounding factors)*. I Kapitel 5 beskrivs flera standardrutiner för att göra detta.

Det är inte alltid praktiskt möjligt att erhålla ett enda referensvärde. Det kan därmed även vara svårt att få flera mätningar under konstanta omständigheter. För att ta hänsyn till denna svårighet kan man välja att kvantifiera noggrannheten genom att para ihop varje mätning som görs av den aktuella givaren (y_i) med ett motsvarande referensvärde ($y_{\mathrm{ref},i}$). Mätfelen beräknas sedan som de parvisa avvikelserna:

$$e_i = y_i - y_{\mathrm{ref},i}, \quad i = 1,2,\ldots N \tag{3.3}$$

Följande avsnitt definierar och diskuterar begreppen noggrannhet, riktighet och precision.

Noggrannhet (”accuracy”)

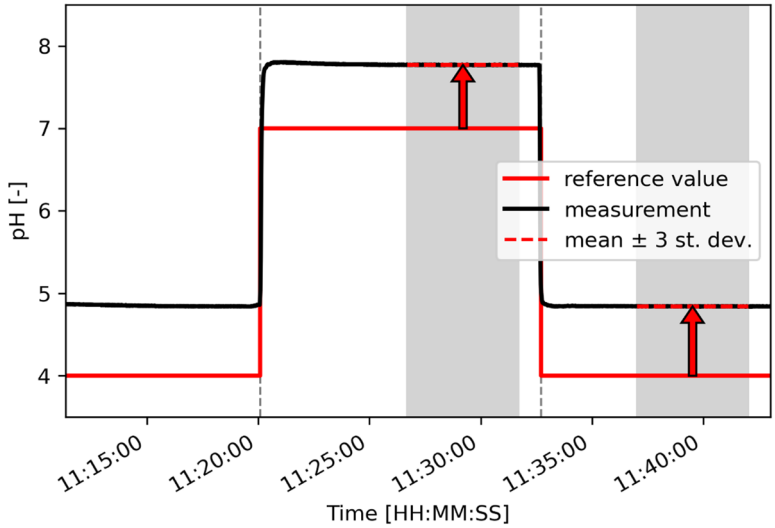
Definition D3.18: **Noggrannheten** beskriver hur god överensstämmelsen är mellan en mätning och det accepterade referensvärdet (ISO, 2006).

Motivering: Att bedöma noggrannheten är vanligtvis det första steget för att ta evidensbaserade beslut om hur ofta givare ska underhållas och bytas ut. Det är en god grundläggande metodik eftersom slumpmässiga och systematiska fel (definitioner nedan) alla bidrar till en försämrad noggrannhet. Noggrannheten är också ett stöd för att avgöra om datakvaliteten är tillräckligt bra för ett visst syfte. Enbart uppskattning av noggrannheten är dock otillräcklig för att välja korrigerande åtgärder för att förbättrad givarens mätning. För att kunna göra det krävs att noggrannheten delas upp i riktighet (*trueness*) och precision (*precision*), se definitionerna nedan.

Kvantifiering: Noggrannheten utvärderas vanligen genom att jämföra flera mätningar som gjorts under samma förhållanden (y) med ett referensvärde (y_{ref}). Med N mätningar erhåller man N mätfel och kvantifierar sedan noggrannheten som medelvärdet av mätfelet i kvadrat (*Root-mean-squared error* – RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} e_i^{\,2}} \tag{3.4}$$

Ett högt värde på RMSE indikerar dålig noggrannhet. Figur 3.1 illustrerar en utvärdering av noggrannheten hos en pH-givare. En pH-elektrod placeras i kalibreringslösningar med kända pH-värden på 4, 7 och 4. Noggrannheten beräknas sedan genom att välja perioder med en stabil



Figur 3.1 Exempel E3.1 – utvärdering av en pH-givares noggrannhet, riktighet och precision. Givaren exponeras för lösningar med kända pH-värden (4 och 7) och givarmätningarna registreras. Noggrannhet, riktighet och precision utvärderas under stationära förhållanden genom att beräkna mätfelen som avvikelser mellan det uppmätta värdet (svart linje) och referensvärdet (röd linje). De mätningar som används för att göra detta erhålls genom att välja ett tidsfönster med en stabil signal. I det här fallet mäter givaren ganska dåligt, eftersom RMSE är 0,84 vid pH 4 och 0,77 vid pH 7. Mätfelets standardavvikelse visualiseras med röda streckade linjer som anger intervallet som omfattar sex gånger standardavvikelsen runt medelvärdet. De övre och nedre värdena som avgränsar detta intervall går inte att urskilja. Givarens precision är alltså utmärkt i det här fallet, med värden för mätfelets standardavvikelse som ligger på låga 0,0018 vid pH 4 och 7. Mer information finns i Ohmura *et al.* (*2019a*). Som ett resultat av detta är noggrannheten (RMSE) praktiskt taget lika stor som korrektheten (bias). Den dåliga noggrannheten beror alltså främst på bristande riktighet i det här fallet.

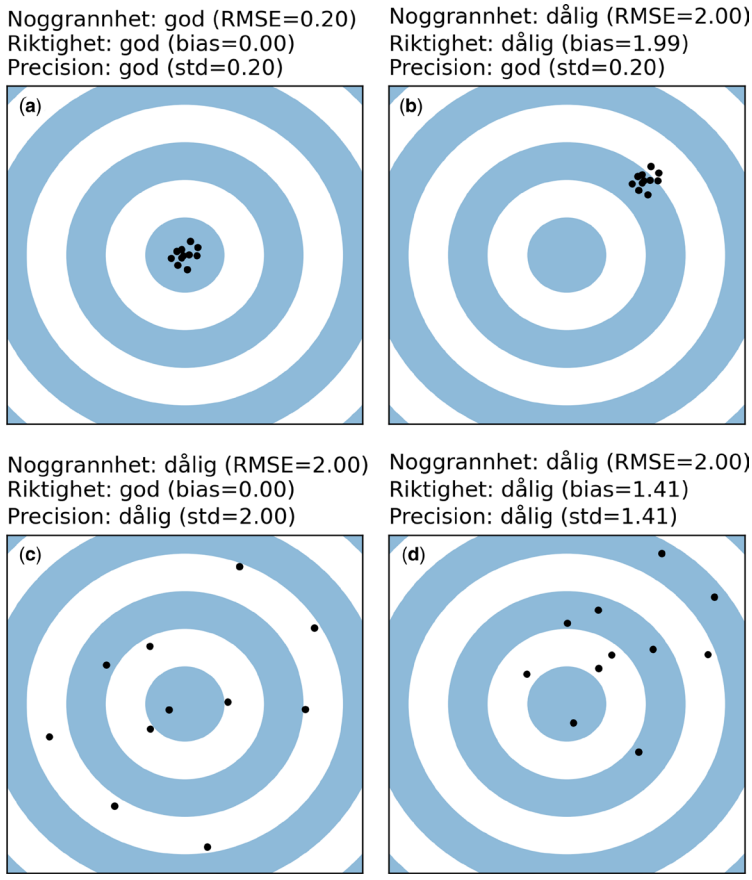
signal och beräkna noggrannheten enligt (3.4). Noggrannheten är mycket dålig i detta exempel, med RMSE-värden på 0,84 vid pH 4 och 0,77 vid pH 7.

Ytterligare noteringar: Observera att noggrannhet är ett övergripande mått på datakvalitet som kombinerar de två måtten riktighet och precision. Som förklaras nedan så kvantifieras riktighet och precision vanligen som bias och standardavvikelse. I dessa fall så gäller att RMSE, bias och standardavvikelse följer sambandet:

$$\mathrm{RMSE}^2 = \mathrm{bias}^2 + \mathrm{standardavvikelse}^2 \tag{3.5}$$

Figur 3.2 illustrerar detta genom en analogi med bågskytte. I (a) kan man se perfekt skjutna pilar. Träffpunkterna ligger nära målets mitt samtidigt som de ligger nära varandra. Detta leder till hög precision och noggrannhet som återspeglas i ett lågt RMSE på 0,2. I figurerna (b), (c) och (d) kan man se resultaten från tre skyttar med sämre träffsäkerhet. Dessa tre skyttar har samma RMSE-värde på 2,0 och skjuter därmed lika dåligt. Den genomsnittliga landningspunkten (riktighet) och spridningen (precision) av landningspunkterna skiljer sig dock mycket åt mellan dessa tre skyttar. De kan behöva förbättra olika saker i sin bågskytteteknik. På samma sätt kan det hända att en givares noggrannhet inte ger tillräcklig information för att förbättra givarens noggrannhet. För att göra detta kan det vara till hjälp att uppskatta riktigheten och precisionen genom att dela upp RMSE i bias och varians (se nedan).

Om man även antar att mätfelen är normalfördelade kan bias och standardavvikelse användas som estimat av medelvärdet och standardavvikelsen för denna fördelning. Detta illustreras i Figur 3.3 med



Figur 3.2 Exempel E3.2 – illustration av noggrannhet, riktighet och precision genom analogi med bågskytte. Tavlans mittpunkt är analogt med referensvärdet för en givare eller mätmetod: (a) en utmärkt bågskytt med god riktighet och precision, vilket leder till god noggrannhet; (b) en bågskytt med dålig riktighet och god precision, vilket leder till dålig noggrannhet; (c) en bågskytt med god riktighet och dålig precision, vilket leder till dålig noggrannhet; (c) en bågskytt med dålig riktighet och dålig precision, vilket leder till dålig noggrannhet. Observera att den övergripande noggrannheten mätt med RMSE i panelerna (b), (c) och (d) är densamma.

hjälp av de datapunkter som visas i Figur 3.2. Denna figur visar att mätfelsfördelningar med olika parametervärden kan ge samma noggrannhet.

Observera också att RMSE bara är ett sätt att utvärdera noggrannheten. Andra mått, som medianabsolutavvikelse (MAD), föredras ofta när datasetet innehåller många extremvärden (*outliers*). Se Daszykowski *et al.* (2007) och Zoubir *et al.* (2018) för en introduktion till robust statistik (*robust statistics*). Andra mått som baseras på s.k. *bootstrap*-teknik finns i Efron och Tibshirani (1986). Slutligen kan datakvalitet också mätas Med befintliga mått för att mäta modellnoggrannhet (se t.ex. Hauduc *m.fl.*, 2015) om modellprediktioner ersätts med referensvärden.

Riktighet och bias

I följande definitioner diskuteras riktighet och det vanligaste måttet för kvantifiering av detta, det vill säga bias.

Definition D3.19: **Riktighet** är överensstämmelsen mellan det medelvärde som erhålls från en stor uppsättning datapunkter och ett accepterat referensvärde (ISO, 2006).

Definition D3.20: **Bias** är den förväntade skillnaden mellan en uppsättning datapunkter och det accepterade referensvärdet (ISO, 2006).

Motivering: Att bedöma riktighet, eventuellt genom att kvantifiera bias, gör det möjligt att upptäcka och kvantifiera systematiska fel. Genom att fokusera på en genomsnittlig avvikelse mellan mätningarna och referensen, elimineras inverkan av mätfel.

Kvantifiering: Bias är det vanligaste måttet för att kvantifiera riktighet. Det uppskattas ofta som medelvärdet av en uppsättning mätfel. Givet en uppsättning mätfelen enligt (3.2)–(3.3) kan bias kvantifieras som mätfelens medelvärde:

$$\text{bias} = \frac{1}{N}\sum_{i=1}^{N} e_i \tag{3.6}$$

Precis som när det gäller noggrannhet är det viktigt att det finns ett korrekt referensvärde. Se kommentarerna ovan om val av metod för att erhålla sådana referensvärden. Figur 3.1 visar data som erhållits genom att utsätta en pH-givare för lösningar med kända pH-värden på 4 och 7. Givaren avviker tydligt från referensvärdena med cirka 0,75 pH-enheter vid de aktuella pH-värdena (avvikelsen indikeras med röda pilot i Figur 3.1).

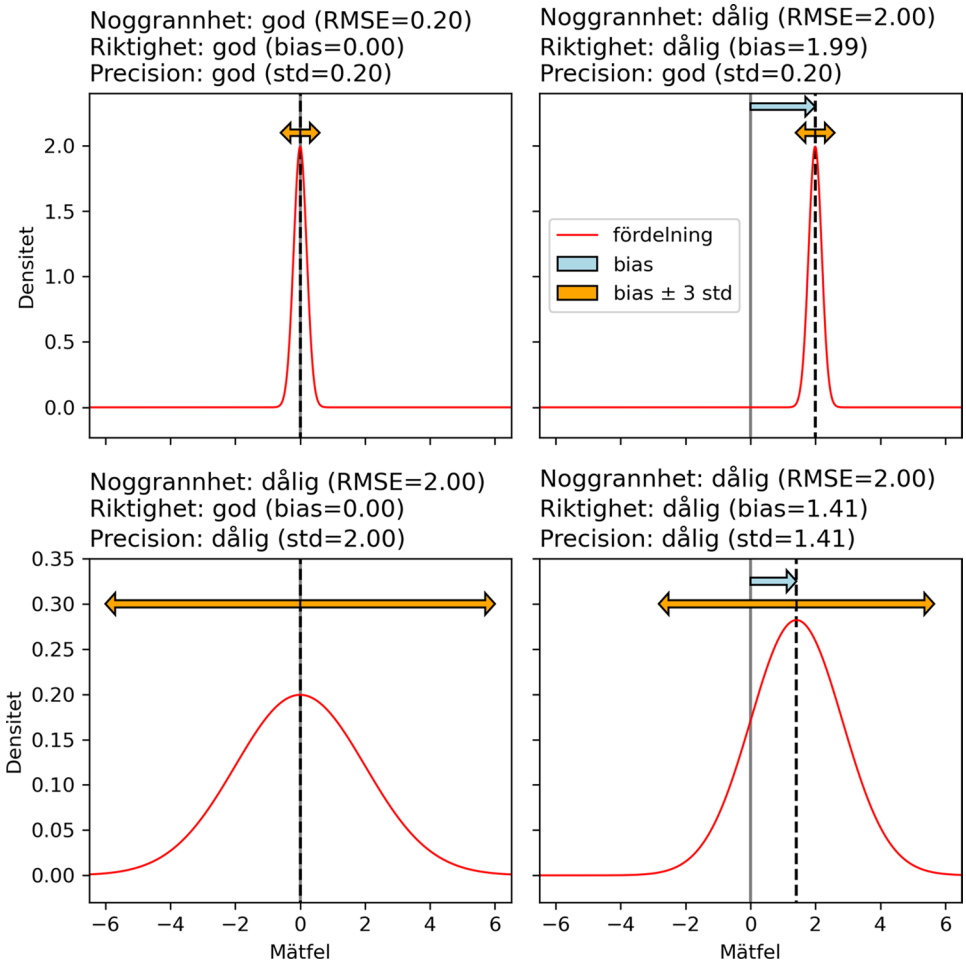
Ytterligare noteringar: Figur 3.2b illustrerar hur en bristande riktighet leder till dålig noggrannhet. Figur 3.3b illustrerar hur en normalfördelning som beskriver mätfelsfördelningen nu har ett medelvärde som inte är noll på grund av bristen på riktighet.

Givarens bias kan påverkas av flera faktorer, bland annat installationssätt och placering, rengöring, kalibrering och slitage. Ibland är givarens bias konstant och påverkas inte av externa faktorer. Ett konstant bias kan t.ex. orsakas av ett fel i kalibreringskurvans offset (se Kapitel 5). I det fallet kan givaren fortfarande fånga upp trender i den uppmätta variabeln. En givares bias kan också förändras över tiden vilket kallas för drift (se nedan).

Ett annat exempel är felaktig kalibrering av en givares känslighet eller skalning (*sensitivity*) (se Kapitel 5), vilket uttrycks som ett bias. Storleken på bias kommer dock att förändras som en funktion av det verkliga värdet på den uppmätta variabeln. När en sådan givare används i ett dynamiskt system leder den systematiska avvikelsen till ett mätfel vars storlek förändras med förändringar i det uppmätta värdet. Detta kan göra det svårt att skilja på ett kalibreringsfel och drift på grund av slitage. Detta illustrera vikten av att undersöka korrektheten under väl kontrollerade förhållanden och inte enbart under drift.

Bias kan också orsakas vid omvandling av analog signal till digital (A/D)-omvandling. A/D-omvandling innebär att varje datapunkt avrundas uppåt eller nedåt till närmaste multipel av den numeriska upplösningen innan den skrivs till en fysisk datalagringsenhet (t.ex. hårddisk, dataserver). Detta kallas även kvantisering och dess konsekvenser och åtgärder har studerats i detalj (se t.ex. Oppenheim och Schafer, 1975; Wang *et al.*, 2010; Widrow och Kollár, 2008). Mätområdet dividerat med den numeriska upplösningen är lika med antalet unika värden (N) som kan lagras för datapunkten (se Kapitel 2, Avsnitt 2.4.1 för definition). Detta antal (N) beskrivs ofta i bitar, det vill säga $\log_2(N)$. Antalet bitar kan variera och kan anta värden mellan 1 bit (enkel på/av-signal) och 64 bitar (som i de flesta moderna stationära och bärbara datorer). För kontinuerliga variabler såsom syrehalt, flödeshastighet och pH är det vanligt att använda 16 bitar eller mer, vilket innebär att minst 2 216 = 65 536 unika värden kan registreras inom mätområdet. Om det sanna värdet för den uppmätta variabeln är konstant kommer kvantiseringen att ge en konstant bias. Förskjutningens tecken och storlek beror dock på det sanna värdet. Eftersom det sanna värdet för den uppmätta variabeln sannolikt kommer att förändras över tiden, kommer även storleken på bias att vara dynamisk.

Bias kan ofta elimineras genom kalibrering av dess offset och skalning (se Kapitel 5). Man bör dock vara noga med att först eliminera de faktorer som orsakar bias i den mån det är möjligt. När orsaken



Figur 3.3 Exempel E3.2 – uppdelning av noggrannhet i riktighet och precision gör det möjligt att beskriva mätfelen med en normalfördelning. Denna fördelning har två parametrar, som är medelvärde och standardavvikelse. Dessa parametrar är vanligtvis inställda på bias och mätfelets standardavvikelse. Figurerna (a)–(d) visar sannolikhetsfördelningar som motsvarar bågskytteexemplet i Figur 3.2. Notera att en dålig noggrannhet kan vara en konsekvens av bristande riktighet (b), bristande precision (c) eller en kombination av båda (d).

till bias inte kan elimineras (t.ex. kvantisering, vanligt slitage), kan detta leda till att bias förändras över tiden (dvs. drift, se nedan), vilket gör att regelbunden kalibrering blir nödvändig.

Precision, varians och standardavvikelse

Medan uppskattning av riktighet kräver att slumpmässiga fel elimineras, kräver precision det motsatta. Systematiska fel behöver tas bort för att beskriva de slumpmässiga felen. Följande definitioner baseras på ISO (2006).

Definition D3.21: **Precision** beskriver hur nära enskilda datapunkter ligger varandra.

Definition D3.22: **Mätfelets varians** är det förväntade värdet för den kvadratiska avvikelsen mellan en enskild mätning och det förväntade värdet (t.ex. medelvärdet) för mätningen.

Motivering: Genom att följa upp precisionen, eventuellt genom att uppskatta variansen i mätfelet, får man ett mått på den spridning som slumpmässiga fel ger upphov till. Precisionen påverkas av tillfälliga och slumpmässiga mätfel, men inte av systematiska fel, vilket gör det lättare att skilja på olika felkällor.

Kvantifiering: Oftast kvantifieras precisionen som mätfelets varians. Givet en uppsättning mätningar såsom i (3.2) eller (3.3) blir mätfelets varians

$$\text{variance} = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \text{average})^2 \tag{3.7}$$

där medelvärdet är det genomsnittliga uppmätta värdet:

$$\text{average} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{3.8}$$

Standardavvikelsen är ett relaterat spridningsmått som beräknas som kvadratroten av variansen:

$$\text{standard deviation} = \sqrt{\text{variance}} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \text{average})^2} \tag{3.9}$$

En ökande variansen i mätningen indikerar en försämrad precision. Till skillnad från noggrannhet och riktighet krävs det inget accepterat referensvärde för att utvärdera precisionen. Alla datapunkter ska dock vara oberoende mätningar av samma variabel under samma förhållanden. Av denna anledning kräver (3.7)–(3.9) att givaren är isolerad från yttre variationer. Hinkmetoden som beskrivs i Kapitel 5 syftar till att göra just detta.

Figur 3.1 illustrerar hur standardavvikelsen kan utvärderas genom att utsätta en pH-givare för lösningar med kända pH-värden 4 och 7. De horisontella röda streckade linjerna anger det intervall som motsvarar +/-3 gånger standardavvikelsen vid pH 4 och 7. De streckade linjerna är omöjliga att skilja åt i båda fallen, vilket tyder på att givaren är mycket exakt. De beräknade standardavvikelserna är både lägre än 0,002.

Som nämnts ovan förutsätter ekvationerna att det verkliga värdet av det uppmätta värdet är konstant, dvs. i ett stabilt tillstånd (*steady-state*). Detta går inte alltid att uppnå. Det kan t.ex. vara opraktiskt att flytta givaren från mätplatsen för att kontrollera mot ett referensvärde eller potentiella störningsfaktorer. I så fall bör man ta hänsyn till effekten av ett förändrat referensvärde på y_i. Variationen i skillnaden mot medelvärdet (y_i – medelvärde) beror nu inte enbart på slumpmässiga mätfel, och ekvationerna ovan ger en dålig kvantifiering av givarens precision. För att lösa detta problem kan man istället beräkna variansen eller standardavvikelserna utifrån (3.3). I praktiken beräknar man först bias (3.6). Därefter kvantifieras precisionen som följande varians:

$$\text{variance} = \frac{1}{N}\sum_{i=1}^{N}(e_i - \text{bias})^2 \tag{3.10}$$

eller standardavvikelsen enligt:

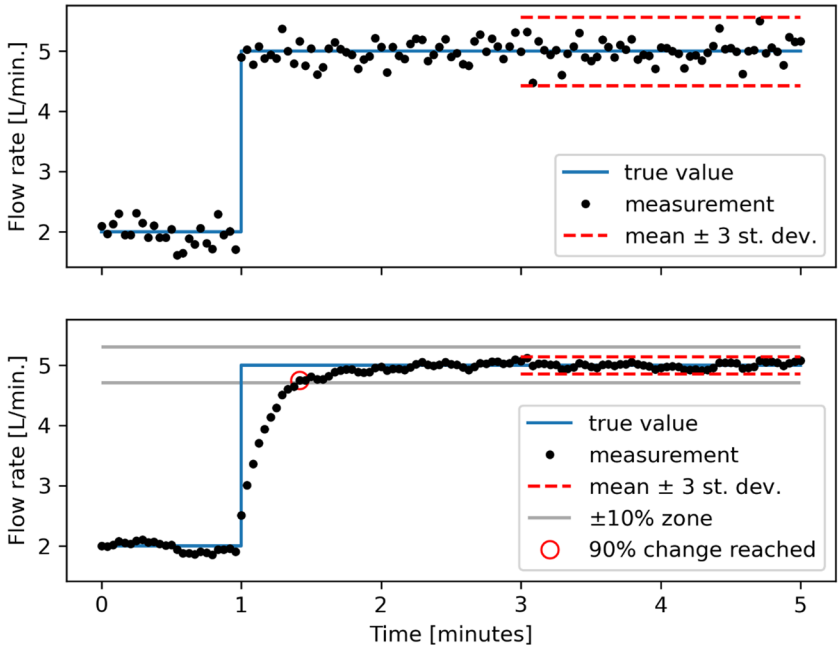
$$\text{standard deviation} = \sqrt{\text{variance}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(e_i - \text{bias})^2} \tag{3.11}$$

Observera att precisionens uppskattning nu i hög grad är beroende av korrekta referensvärden. Detta är inte fallet om (3.8) och (3.9) kan uppskattas från data som tagits fram under konstanta och kontrollerade förhållanden.

Ytterligare noteringar: Figur 3.2c illustrerar hur bristande precision leder till dålig noggrannhet. I Figur 3.3c kan man se hur en normalfördelning som mätfelet med en stor spridning på grund av bristande precision.

Till skillnad från variansen uttrycks standardavvikelsen i samma enheter som den ursprungliga givarsignalen, vilket underlättar tolkningen. Som nämnts ovan så är dock standardavvikelsen en icke väntevärdesriktig (*biased*) uppskattning av den sanna standardavvikelsen. Variansen, enligt beräkningen ovan, är därför ofta att föredra ur ett statistiskt perspektiv, eftersom den är en väntevärdesriktig (*a consistent and unbiased estimator*) av den sanna variansen. Det finns andra sätt att beskriva precision med s.k. robusta skattningar, se Chen *m.fl.* (1997), Daszykowski *m.fl.* (2007) och Zoubir *m.fl.* (2018).

Till viss del kan en låg precision mildras genom medelvärdesbildning eller filtrering. För att illustrera detta visar Figur 3.4 (överst) ett simulerat sant värde och en brusig givarsignal för ett flöde. Givaren reagerar omedelbart men har dålig precision med en uppskattad standardavvikelse under stabila förhållanden på 0,2 L/min. Figur 3.4 (nederst) visar det sanna värdet och en filtrerad version av givarsignalen. Efter att filtrering har standardavvikelsen (vid stationärt tillstånd) reducerats till 0,05 L/min, vilket är betydligt lägre. Baksidan med att filtrera bort bruset är en förlängd responstid. Det innebär att snabba förändringar i mätvärdet inte fångas upp på ett korrekt sätt. I detta nås 90 % av förändringen på 34 s.



Figur 3.4 Exempel E3.3 – datafiltrering kan förbättra precisionen. Överst: En förändring av det sanna värdet för flödeshastigheten simuleras vid 1 min. Givaren reagerar omedelbart men är mycket brusig. Standardavvikelsen uppskattas till 0,19 L/min mellan 3 och 5 minuter. Botten: Givarsignalens precision kan förbättras genom filtrering. Standardavvikelse mellan 3 och 5 minuter är nu 0,05 L/min. Givaren reagerar dock långsammare på den applicerade förändringen och det tar 34 sekunder att registrera 90 % av förändringen.

3.3.1.3 Statisk beskrivning av datakvalitet – avancerad nivå: repeterbarhet och reproducerbarhet

Ovan begrepp och metoder för kontroll av en enda givare förutsätts ske inom ett kort tidsspann. Andra faktorer kan påverka noggrannheten vilket väldefinierade protokoll kan bidra till att undvika.

Repeterbarhet

Definition D3.23: **Repeterbarhet** är precisionen när mätproceduren (vanligtvis givarvalideringen) upprepas under exakt identiska förhållanden. Det vill säga att oberoende testresultat erhålls med samma metod i identiska tester i samma laboratorium, av samma operatör, med samma utrustning, inom ett kort tidsintervall.

Motivering: Utvärderingen av repeterbarhet kräver att alla faktorer som kan påverka mätresultatet elimineras. En jämförelse med den precision som utvärderas utan att eliminera sådana faktorer kan potentiellt fastställa den exakta orsaken till mätfelet.

Kvantifiering: Repeterbarhet är den precision som erhålls efter att alla potentiella orsaker till variation noggrant har eliminerats från mätprocessen. Se ovan för att se hur precisionen kan kvantifieras.

Reproducerbarhet

Definition D3.24: **Reproducerbarhet** är den precision som erhålls från testresultat där samma metod används på identiska testobjekt med olika laboratorier, operatörer och utrustning (ISO, 2006).

Motivering: Medan uppskattning av repeterbarhet kräver att man eliminerar så mycket variation i mättekniken som möjligt, kräver uppskattning av reproducerbarhet att man introducerar flera potentiella felkällor i mätprocessen på ett avsiktligt och systematiskt sätt. Detta kan bidra till att identifiera systematiska och slumpmässiga fel som beror på variationer i olika provtagnings- och mätprocedurer. Eftersom tillfälliga och systematiska fel kan påverka enskilda laboratorier, operatörer eller utrustning kan kvantifiering av precisionen bidra till att upptäcka, diagnostisera och minska mätfelen.

Kvantifiering: Reproducerbarheten bedöms vanligen genom ett s.k. *round-robin*-test. I detta fall delas ett enda vattenprov upp i flera vattenprover. Varje vattenprov analyseras sedan av ett enda laboratorium, en enda operatör eller med ett enstaka mätinstrument. De rapporterade mätningarna används för att kvantifiera den resulterande precisionen eller variansen. Se ovan för att se hur precisionen kan kvantifieras.

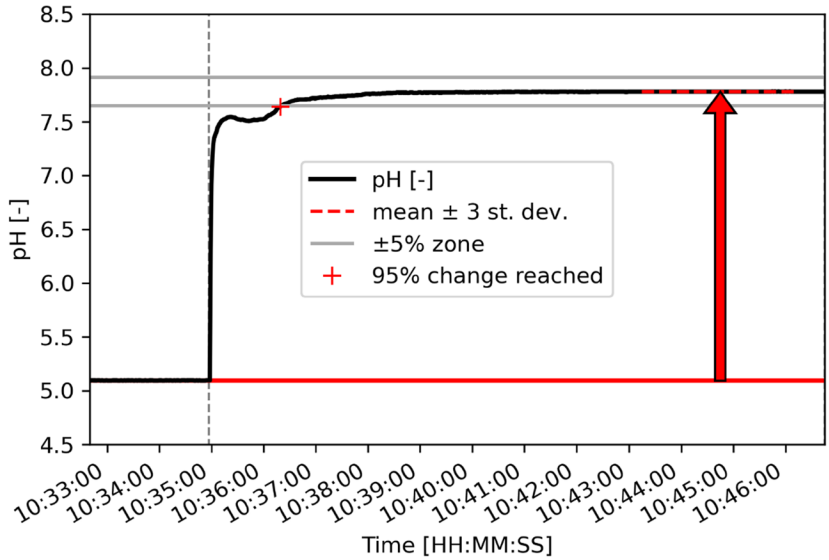
3.3.1.4 Redovisning av dynamik: responstid, chock, drift och extremvärden

I detta avsnitt introduceras ett antal begrepp beskriver olika dynamiska fenomen. Responstiden beskriver en givares dynamiska respons på en förändring av den omgivande halten. Typiskt har de flesta givare någon form av fördröjning och responstid. De återstående termerna – chock och drift – kan beskrivas som symtom på givarfel.

Svarstid

Definition D3.25: **Responstiden** (*response time*) är den tid det tar för ett instrument att visa ett nytt värde när det utsätts för en stegvis förändring. Den kan också definieras som den tid som krävs för att uppnå en procentandel av en stegvis förändring av förändringen av det sanna värdet (typiskt 90 % eller 95 %).

Motivering: Medan noggrannhet, bias och varians kvantifierar givarens prestanda under statiska förhållanden (d.v.s. *steady-state*), kvantifierar de inte hur väl givaren kan hantera dynamik i den uppmätta variabeln. Vanligtvis kvantifieras en givares förmåga att följa trender i den uppmätta variabeln som responstid. En tillräckligt snabb responstid krävs för att använda mätningen för processreglering, där “tillräckligt” beror på hur snabb styrsignal som krävs för att reglera processen.



Figur 3.5 Exempel E3.1 – utvärdering av svarstiden för en pH-givare. Givaren flyttas från en lösning med pH 4 till en lösning med pH 7. Detta gör att givarens värde ändras från 5,1 till 7,8. Svarstiden beräknas som den tid det tar för givaren att nå 95 % av denna värdeförändring och hålla sig inom ett 5 %-band av det slutliga värdet. I det här fallet är svarstiden 85 s. Originaldata har hämtats från Ohmura *et al.* (*2019b*).

Kvantifiering: Internationella standarder (t.ex. ISO, 2003) ger detaljerad information om hur responstider kan kvantifieras, och även relaterade begrepp såsom fördröjning (*delay time*), sjunktid (*fall time*) och stigtid (*rise time*).

Figur 3.5 illustrerar responstidskonceptet baserat på experimentell utvärdering för en pH-givare. Det sanna värdet (baserad på kända experimentella förhållanden) visas i grått, och responsen som är som erhålls med givaren visas i rött. Responstiden kvantifieras som den tid det tar för givarsignalen att nå 95 % av värdeförändringen och är lika med 85 s.

Ytterligare noteringar: Förändringar i en givares svarstid kan ha flera orsaker, inklusive materialval (t.ex. permeabla membran), givaregeometri, givareplacering, blandningsförhållanden, kemisk deposition, nedsmutsning, uppkomna skador (membranrivning) och normalt slitage.

Stegförändring eller chock

Definition D3.26: **En stegvis förändring** eller chock är en drastisk och tillfällig förändring av givarens bias från en tidsstämpel till nästa.

Motivering: Identifiering av chocker i givarsignaler kan bidra till att upptäcka allvarliga skador på en givare. Dessutom bör chocker som orsakas av underhållsåtgärder (t.ex. rengöring, kalibrering) kvantifieras för att möjliggöra korrigering av givarsignaler (t.ex. genom att jämföra mätvärdena före och efter kalibrering).

En stegvis förändring eller chock är en förändring som sker på en gång vid ett specifikt tillfälle. Detta symptom är därför ofta lätt att upptäcka manuellt genom att granska en tidsserie. En chock kan bero på en fysisk skada eller en underhållsåtgärd (t.ex. rengöring, kalibrering). Avvikelsen mellan givarens värde och det verkliga värdet (bias) kan antingen förbättras eller försämras vid tidpunkten för chocken.

Kvantifiering: Den mest exakta utvärderingen av stegvisa förändringar i mätfel erhålls genom regelbunden utvärdering av givarens bias under hela dess livslängd. För att exakt kvantifiera

chockens storlek vid underhållsåtgärd (t.ex. rengöring, kalibrering) bör bias kvantifieras före och efter underhållsåtgärden. I bilaga 5.A i Kapitel 5 finns flera metoder för detta. I Kapitel 6 presenteras flera möjliga metoder för att upptäcka och kvantifiera oavsiktliga chocker (d.v.s. chocker som inte beror på underhållsåtgärder).

Drift

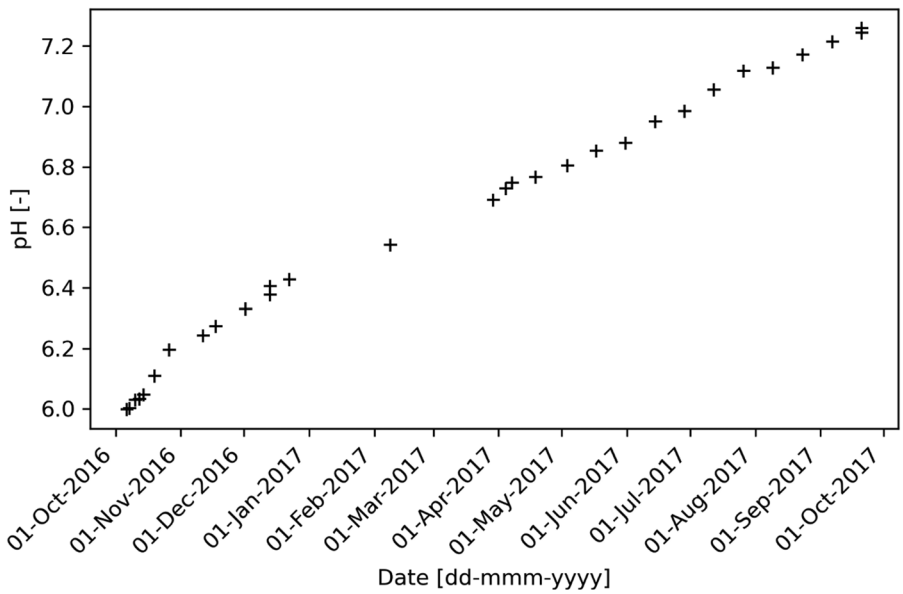
Definition D3.27: **Drift** är en långsam förändring av bias, varians eller responstid. Drift är ett symtom som utvecklas gradvis och kontinuerligt över tid.

Motivering: Insamling av information om givardrift rekommenderas för att kunna optimera underhållsintervall för givare (rengöring, kalibrering, utbyte). Förändringar i givarens bias är i de flesta fall ett prioriterat problem. Förändringar i responsstid kan dock kvantifieras utan att givaren behöver flyttas och är därför användbara för att effektivt spåra effekterna av påväxt och fel i skalning, se (Samuelsson *et al.*, 2018).

Kvantifiering: Givardrift kvantifieras genom regelbunden utvärdering av givarens bias, varians och svarstid under givarens hela livslängd.

Figur 3.6 visar ett experimentellt exempel på en förändrad bias som kvantifieras genom regelbunden validering. Figuren visar pH-värdet från en okalibrerad pH-givare i ett medium med pH 6. I avsaknad av korrigerande kalibrerings- och underhållsåtgärder ger denna givare en signal som kontinuerligt rör sig bort från referensvärdet (här: pH 6) allteftersom tiden går.

Ytterligare noteringar: Den riktning och hastighet med vilken bias, varians eller svarstid förändras är inte nödvändigtvis konstant. Den mest tillförlitliga metoden för att motverka effekterna av drift av bias är regelbunden validering och kalibrering. I Kapitel 5 listas metoder för att regelbundet samla in referensmätningar för att kvantifiera förändringar i bias och andra symtom. I Kapitel 6 presenteras flera metoder för att upptäcka och kvantifiera drift med hjälp av algoritmer.



Figur 3.6 Exempel E3.1 – utveckling av pH-givarens signal över tiden i ett medium med pH 6. De värden som visas är de förväntade pH-värdena om givaren aldrig kalibreras. Dessa pH-datapunkter beräknades utifrån de uppmätta data och kalibreringsvärdena för pH-givarens offset och känslighet som erhölls genom tvåpunktskalibreringar vid pH 4 och 7. Mer information finns i Ohmura *et al.* (*2019a*).

Extremvärden (“outliers”)

Till skillnad från stegvisa eller långsamma förändringar (drift) av bias, varians och responstid så är det vanligt med symtom under en kort tid. Det generella begreppet för datapunkter som uppvisar kortlivade symtom för outliers.

Definition D3.28: **En avvikelse** är en del av en datamängd som inte stämmer överens med de andra delarna av datamängden.

Motivering: De flesta dataanalysmetoder påverkas av avvikande värden. Det är därför viktigt att upptäcka avvikande värden tidigt. Outliers kan indikera en försämrad datakvalitet, men de kan också vara en korrekt mätning av ett avvikande processtillstånd. Därför är det viktigt att skilja mellan de potentiella grundorsakerna till en outlier innan enskilda datapunkter tas bort eller ersätts.

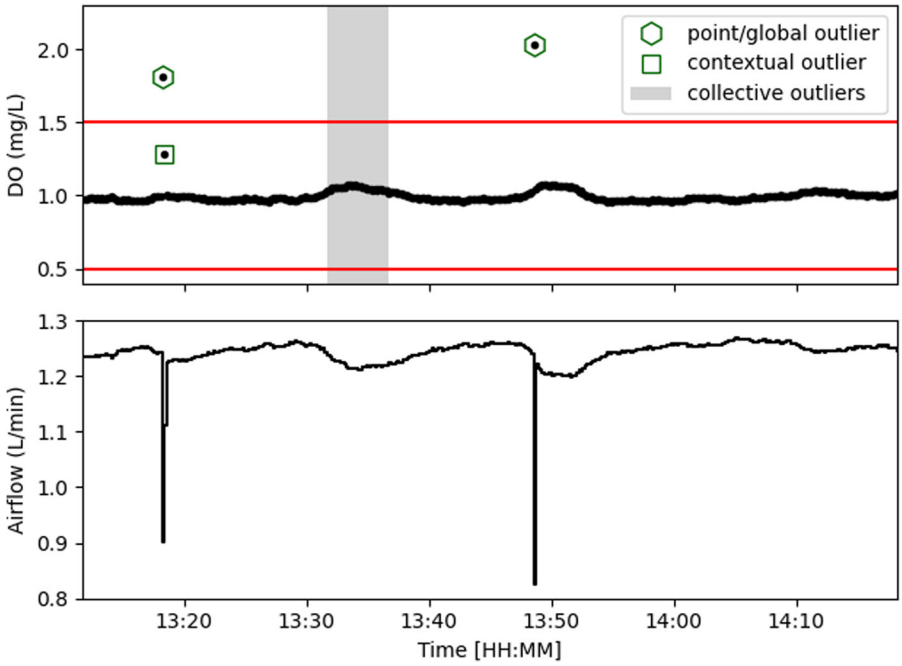
Hantering av outliers har blivit en central aktivitet då de kan ha stor påverkan på dataanalysresultat och modellkalibrering. Detektering och korrigering av extremvärden är grundläggande för att undvika fel i parameterskattning i dynamiska system (Albuquerque och Biegler, 1996).

Outliers kan orsakas av variabilitet som beror på själva i mätningen eller processen, vilket kräver olika åtgärder. I de fall outliers orsakas av avvikelser på grund av mätfel eller datainsamlingsfel bör avvikande värden förkastas när de bekräftas som felaktiga mätningar (Grubbs, 1969).

Outliers brukar delas in i en av tre kategorier:

- Enstaka extremvärde (“point outlier or global outlier”): En datapunkt vars värde ligger långt utanför intervallet för hela datauppsättningen. Vanligtvis kan sådana avvikande värden upptäckas automatiskt genom att jämföra dem med givarens metadata, t.ex. mätområdet eller variabelområdet (se Avsnitt 3.2.2.6–3.2.2.7). Detta illustreras med mätningar av löst syre i Figur 3.7 (överst). Variabelintervallet i detta exempel är 0,5–1,5 mg/L, och två datapunkter är avvikande extremvärden.
- Kontextuell outlier (“contextual outlier”): En datapunkt vars värde avviker avsevärt från resten av datapunkterna i ett visst sammanhang eller villkor, även om datapunkten skulle kunna vara möjlig i ett annat sammanhang. Figur 3.7 (överst) visar ett sådant exempel med en rektangel. Datapunktens värde ligger inom det förväntade spannet för variabeln, men avviker tydligt från mätningarna före och efter. I detta specifika sammanhang betraktas den därför som en outlier.
- Kollektiva outliers (“collective outliers”): Detta fall avser inte en enskild datapunkt utan snarare en delmängd av datasetet. Denna delmängd av data är en kollektiv outlier om de enskilda datapunkterna inte uppfyller de tidigare definitionerna av punkt- eller kontextuella outliers men ändå avviker avsevärt från det fullständiga datasetet. Ofta innebär detta att vissa mönster uppstår som kan betraktas som normala i ett visst sammanhang men inte i det aktuella fallet. Figur 3.7 (överst) visar ett exempel i grå skuggning. Datapunkterna under den markerade tiden är något högre än datapunkterna före och efter. Det finns också liknande data som samlats in cirka 20 minuter senare. Denna uppgång i koncentrationen kan dock betraktas som onormal eftersom en operatör bedömer att den inträffade vid en onormal tidpunkt. Det är naturligtvis ofta mer utmanande och subjektivt att identifiera kollektiva outliers än enstaka outliers eller kontextuella outliers.

Ytterligare noteringar: Konsekvenserna av extremvärden visas också i Figur 3.7. I det här exemplet styrs syrehalten med en PI-regulator genom att ställa ut ett luftflöde. Vid tiden 13:18 ökar den uppmätta syrehalten från 0,99 mg/L till 2,03 mg/L, troligen på grund av ett elektriskt fel. PI-regulatorn reagerar på den plötsliga syrehaltsförändringen och minskar luftningen. Efter 20 s återgår syrehalten till rätt börvärde (1,0 mg/L). Om den här typen av störningar inte beaktas kan de orsaka störningar i processen eller orsaka överdrivet slitage på manöverutrustningen (t.ex. ventiler, pumpar).



Figur 3.7 Exempel E3.4 – illustration av olika sorters avvikande värden – outliers. Övre figur: mätningar av syrehalt som en funktion av tiden med outliers: sexhörningar anger enstaka extremvärden som identifierats baserat på variabelintervallet (se Avsnitt 2.2.7); kvadraten indikerar en s.k. kontextuell outlier; det grå området anger datapunkter som är märkta som kollektivt avvikande värden. Nedre figur: luftflödesbörvärde som en funktion av tiden. Förekomsten av punkt- och kontextuella outliers orsakar stora störningar i reglering av luftflödet. Figuren baseras på originaldata som samlades av Zhang *et al.* (2021).

3.3.2 Beskriva mätfel genom deras grundorsaker

3.3.2.1 Översikt

Mätfel kan ha många olika orsaker. Genom att identifiera källan till dessa möjliga fel kan man föreslå strategier för att minska förekomsten av dem eller, om det inte är möjligt, filtrera bort dem från datasetet med minimal informationsförlust. Ovan har vi diskuterat hur givarens mätfel kan kvantifieras med

Tabell 3.2 Klassificering av mätfel utifrån rotorsakens natur.

Tidsspann	Typisk Effekt	
	Grundorsaken förväntas ge upphov till konstanta eller långsamt varierande fel i mätvärden	*Grundorsaken förväntas ge upphov till fluktuationer i mätfelet*
Grundorsaken är ihållande, det vill säga den förekommer kontinuerligt eller under långa perioder	Systematiska fel	Slumpmässigt fel
Grundorsaken är tillfällig, det vill säga vid en specifik tidpunkt		Tillfälligt fel

mått som noggrannhet, bias, varians och responstid. I de följande styckena diskuteras mätfel utifrån typiska orsaker eller källor.

Avsnittet följer i stort sett de definitioner som ges av McGhee (2005). Här delas fel in i tre kategorier utifrån rotorsakens natur: systematiska fel, tillfälliga fel och slumpmässiga fel. Dessa tre kategorier beskrivs nedan. En översikt ges i Tabell 3.2. Observera dock att

Att skilja mellan dessa kategorier är inte trivialt. Eftersom slumpmässiga fel främst beskrivs genom att kvantifiera varians och icke-slumpmässiga fel främst beskrivs genom att kvantifiera bias, är det vanligt att betrakta systematiska fel och tillfälliga fel som en enda kategori (se t.ex. Narasimhan och Jordache, 1999).

3.3.2.2 Systematiska fel

Definition D3.29: Ett **systematiskt fel** är beständigt i tiden och kommer sannolikt att ge upphov till ett mätfel som är konstant eller förändras långsamt över tid.

Systematiska fel orsakas ofta av instrumentfel, felaktig kalibrering och strukturella defekter i datainsamling eller bearbetning. På grund av att felen kvarstår är det möjligt att kvantifiera symtomen genom upprepade mätningar. Flera datapunkter förväntas påverkas på samma sätt, det vill säga med samma storlek och riktning på felet.

Vanliga orsaker som leder till systematiska fel kan vara:

- **Kalibreringsfel**. Kalibreringsfel kan uppstå på grund av många faktorer, t.ex. felaktigt val av eller beredning av standardlösningar, bristfällig rengöring av givaren före kalibrering, temperaturskillnader mellan den standardlösning som används för kalibrering och temperaturen på mätplatsen, blockeringsproblem i slangar eller ventiler i automatiska kalibreringssystem och brist på korrekt respons från givaren inom kalibreringsområdet. Kalibreringsfel kan ge en felaktig karakterisering av givarens kalibreringskurva, inklusive offset och skalning (se Kapitel 5). Symptomen på kalibreringsfel kan i sin tur vara ökad bias och varians. Kalibreringsfel kan minimeras genom att använda standardrutiner för givarkalibrering, vilket kan inkludera exakt kontroll av mediet under kalibreringen (se Kapitel 5).
- **Mätstörande faktorer**. Alla variabler eller faktorer som påverkar mätningen av ett instrument kallas mätstörande faktorer (*confounding factors*). Påverkan från sådana faktorer kan antingen ge upphov till ett (felaktigt) högre eller lägre mätvärde (*positive or negative confounding*). De mätstörande faktorerna hanteras vanligen genom att de mäts och därigenom kan deras på mätsignalen kvantifieras. Ett exempel är s.k. jonselektiva givare (*Ion selective electrodes–ISE*) för ammonium som störs av kaliumjoner. En ökning av kaliumhalten kan leda till en ökad elektrodrespons som felaktigt registreras som en ökning av ammoniumhalten. För att undvika detta, innehåller jonselektiva givare för ammonium även en extra elektrod för mätning av kalium. Det vanligaste symtomet på en mätstörande faktor är bias i mätsignalen. Observera att denna bias kan förändras över tiden. I exemplet med ammoniumgivaren kan kaliumkoncentrationen påverkas av avisningssalter under vintern, vilket leder till en tidsvarierande bias. Detta kan i sin tur beskrivas som drift. Att kontrollera eller mäta mätstörande faktorer är särskilt viktigt vid givarkalibrering.
- **Nedsmutsning**. Igensättning och nedsmutsning av givare är ett vanligt problem på reningsverk. I många fall kan det påväxten tas bort ganska enkelt genom varsam mekanisk rengöring (vatten och/eller trasa). I sådana består påväxten typiskt av biofilm. Påväxt leder normalt till minskad massöverföring mellan vätskan och givaren. Det vanligaste symtomet på detta är en ökning av responstiden. Detta kräver då längre väntetider vid kalibrering. En bieffekt av den begränsade massöverföringen är att slumpmässiga variationer i bulkvätskan får mindre effekt på mätningarna, vilket kan leda till en minskad varians i signalen. Igensättning kan även öka variansen i signalen, vilket visar på nyttan med att använda varians som ett sätt att övervaka nedsmutsning. När det gäller biofilmspåväxt kan ansamling av mikroorganismer förändra koncentrationen av den uppmätta variabeln (t.ex. syre, nitrat), vilket leder till en förändring av offset och/eller känslighet (se Kapitel 5) och försvårar kalibreringen. Det mest uppenbara

symtomet på detta fenomen är bias. Fel som orsakas av nedsmutsning kan normalt minskas genom regelbunden inspektion och rengöring.

- **Olinjäritet**. Givarkalibrering förutsätter ofta en linjär respons på förändringar i det uppmätta värdet. Det är dock vanligt med olinjära mätkurvor, t.ex. ger jonselektiva ammoniumgivare en olinjär respons på mätvärden under 1 mg/L (när den verkliga halten minskar linjärt). Givare kan också uppvisa hysteres, vilket innebär att en uppåtgående trend i den uppmätta variabeln leder till en negativ bias (mätningen är för låg) och vice versa. Symptomen på dessa typer av olinjära beteenden inkluderar bias och förändringar i varians.
- **Ofullständigt omrörd bassäng**. De flesta reglerstrategier utgår från en perfekt omblandning i tanken, så att mätningarna vid en viss punkt representerar hela reaktorn. I praktiken är det inte säkert att vattnet blandas lika mycket i hela tanken. Detta kan leda till dödzoner (zoner med ingen eller begränsad omblandning) och kortslutningsströmmar. Detta kan också orsaka variationer i mätvärdena. Symtomen på ofullständig blandning kan vara bias, varians och förändringar i responstiden. Korrekt placering och givarinstallation minimerar problemen med dessa fenomen. Beräkningsmodellering av strömningsdynamik (CFD, *computational fluid dynamics*) kan användas för att analysera dessa utmaningar och identifiera bättre positioner för givarplacering.
- **Instrumentfel**. Symtom på instrumentfel kan vara bias, varians, drift och förändringar i responstid. Det finns många orsaker till instrumentfel såsom felaktig konstruktion, begränsningar av mätningar på grund av hysteres, mätning utanför instrumentets rekommenderade mätintervall, felaktig användning och brist på underhåll.
- **Utfällningar**. Även kemiska ämnen kan leda till påväxt och utfällningar på givaren. I motsats till nedsmutsning kan det vara svårt att få bort utfällningar med enbart mekanisk rengöring. Symtomen på utfällningar är en ökning av responsstid, precis som vid nedsmutsning. Utfällning leder inte nödvändigtvis till bias om inte utfällningen helt begränsar massöverföringen till givaren. Fel som orsakas av utfällningar kan normalt åtgärdas genom regelbunden inspektion, kemisk rengöring och, i slutändan, byte av delar.
- **Slitage**. Många givare är inte tillräckligt tåliga för att klara av långvarig exponering mot reningsverkets tuffa miljö. Detta leder till att sådana givare kommer ge en sämre datakvalitet allteftersom tiden går. Detta kan leda till symtom såsom bias, varians, drift och förändringar i responstid, eller ett fullständigt haveri och förlorade mätvärden. Slitage kan minimeras med en genomtänkt placering och installation av givaren, samt motverkas genom underhåll och byte av delar.

3.3.2.3 Tillfälliga fel

Definition D3.30: Ett **tillfälligt mätfel** (*gross error*) orsakas av ett fenomen som inträffar sällan och under korta perioder och som ofta bara påverkar en enda datapunkt.

På grund av den låga frekvensen är det i allmänhet svårt att beskriva och kvantifiera symptomen enbart genom upprepade mätningar. På samma sätt kan de exakta förhållanden under vilka felet uppstår vara svåra att reproducera. Av detta skäl anses denna typ av fel inte kunna upprepas.

Tillfälliga fel induceras ofta av den mänskliga faktorn. De kan uppstå vid insamling eller beredning av vattenprover (t.ex. felaktig utspädningsfaktor) eller vid analys (t.ex. felaktig instrumentavläsning). De kan också orsakas av tillfälliga givarfel. Vi skiljer mellan följande generiska orsaker:

- **Felaktig avläsning eller registrering**. Mätningen utförs korrekt, men det mätvärdet avläses eller registreras felaktigt. Detta är vanligast vid laboratoriemätningar.
- **Felaktig beredning av vattenprovet**. Vattenprovet är felaktigt förberett innan det analyseras. T.ex. kan detta inkludera ändringar i rutinerna såsom, titrering och kemikaliehantering, utspädning, filtrering och blandning. Dessa fel kan även uppstå hos laboratoriemätningar och online-analysatorer (t.ex. ammoniumanalysator).

- **Beräkningsfel**. En felaktig mätning uppstår på grund av ett beräkningsfel. Sådana fel kan uppstå i laboratoriemiljö (t.ex. utspädningsfaktor vid applicering) och i online-givare (t.ex. på grund av elektriska fel eller programvarufel).

3.3.2.4 Slumpmässiga fel

Definition D3.31: Ett **slumpmässigt fel** (*random error*) orsakas av oförutsägbara men ihållande variationer som påverkar alla mätningar av samma givare på samma eller liknande sätt.

Slumpmässiga fel orsakas av oförutsägbara, alltid förekommande variationer i beredningen av vattenprovet, givaren och andra experimentella förhållanden. Slumpmässiga fel skiljer sig från systematiska och tillfälliga fel eftersom de inte kan reproduceras exakt. Det främsta symtomet på ett slumpmässigt fel är en konstant variation, som ofta beskrivs som brus. Slumpmässiga fel har flera orsaker som ofta är svåra att urskilja vilka inkluderar:

- Okontrollerbara variationer i beredningen av vattenprover.
- Oförutsägbara variationer i mätmiljön, inklusive temperatur och tryck.
- Elektriskt och magnetiskt fältbrus som orsakas av variationer i elektrisk spänning, ström och/eller magnetfält. Denna typ av fel är särskilt relevant när analoga mätningar kommuniceras genom oskärmade kablar.
- Variationer i elkvaliteten. Strömförsörjningen, oavsett om det är elnätet, ett batteri eller en alternativ energikälla (t.ex. solel), kan orsaka oförutsägbara variationer i strömkvaliteten. Detta kan i sin tur påverka datapunkter i dåligt avskärmade givare.

3.3.3 Diagnostik: identifiera de orsaker som förklarar symtomen

De flesta dataanalytiska verktyg för givardata syftar till att upptäcka symtom och ta bort data med försämrad kvalitet. Denna strategi är dock ofta otillräcklig för att säkerställa god eller förbättrad datakvalitet. Det är istället mer effektivt att åtgärda grundorsaken till den låga datakvaliteten.

Det finns många studier i litteraturen som beskriver olika metoder för diagnostik, det vill säga metoder som kan identifiera den grundorsak som bäst förklarar de observerade symtomen på försämrad datakvalitet (se t.ex. Venkatasubramanian, *2003a*, *b*, *c*)). Diagnostik är utmanande eftersom det inte finns någon generellt samband grundorsaker och symtom. Tabell 3.3 ger några rimliga matchningar mellan de symptom och grundorsaker som beskrivits i ovan avsnitt.

Tyvärr är många symtom inte unika för en enda grundorsak, så flera potentiella grundorsaker måste undersökas. Drift kan t.ex. orsakas av normalt slitage och mätstörande faktorer. En jonselektiv ammoniumgivare kan driva på grund av en koncentrationsförändring i referenselektroden, vilket kan åtgärdas genom kalibrering. Drift kan emellertid också uppstå på grund av en förändring i en mätstörande faktor såsom kaliumhalten vattenfasen. Detta illustrerar utmaningen med att ett symtom av mätfelet inte alltid räcker för att identifiera lämpliga korrigerande åtgärder.

Utöver det faktum att det inte alltid finns en grundorsak som förklarar observerade symtom, kan diagnostik av mätfel ytterligare försvåras av den återkoppling som sker i reglerkretsar. När givare används för reglering (återkoppling) kan det vara svårt att upptäcka förändringar i datakvaliteten, eftersom regulatorn maskerar många symtom eller minskar deras omfattning. I sådana fall behöver även styrsignalen analyseras. I Kapitel 6 beskrivs multivariata tekniker som kan användas för detta ändamål.

När en försämrad datakvalitet upptäcks bör diagnosen till grundorsaken baseras på en kombination av operatörens erfarenhet om den övervakade processen, kunskap om givaren och en analys av mätvärdena. I Kapitel 5 beskrivs flera metoder som är användbara för att ställa en givardiagnos, särskilt genom att minimera inverkan av mätstörande faktorer och andra omständigheter. Kapitel 6 innehåller dataanalytiska metoder som kan användas för att upptäcka och ibland hjälpa till att identifiera grundorsaken till en försämrad datakvalitet från givare. Naturligtvis behöver kostnaderna för att använda dessa verktyg beaktas och vägas mot nyttan med en förbättrad datakvalitet.

Tabell 3.3 Möjliga grundorsaker till typiska symtom på försämrad datakvalitet från givare.

	Symptom				
Grundläggande orsaker	**Bias**	**Avvikelse**	**Otillräcklig responstid**	**Drift**	**Outlier**
Systematiska fel					
Fel i kalibreringen	X	X			
Mätstörande faktorer	X			X	
Biofilmspåväxt	X	X	X	X	
Olinjäritet	X	X			
Ofullständig blandning	X	X	X		
Givarfel	X	X	X		X
Utfällning	X	X	X	X	
Slitage	X	X	X	X	
Tillfälliga fel					
Felaktig avläsning					X
Felaktig beredning av vattenprov					X
Beräkningsfel					X
Slumpmässiga fel					
Okontrollerbara variationer i beredningen av vattenprover		X			X
Oförutsägbara variationer i den uppmätta miljön		X			X
Brus från elektriska och magnetiska fält		X			X
Variationer i elkvalitet		X			X

Möjliga samband mellan grundorsaker och symtom i datakvaliteten för givare markeras med *X*. Tyvärr är vanligen åtgärder för att kvantifiera symtomen otillräckliga för att isolera en unik grundorsak. Både experimentella metoder (Kapitel 5) och dataanalytiska verktyg (Kapitel 6) kan användas för att för att identifiera grundorsaker och förfina diagnosen.

3.4 METADATA SOM GER KONTEXTUELL INFORMATION: ANNOTERING

Det är vanligt att data kompletteras med anteckningar för att ge dataanvändaren mer information om den kontext data samlats in i, så kallad kontextuell information. Med visualiseringsverktyg (pilar, färger, skuggning, rutor) samt strukturerade eller ostrukturerade kommentarer (se definitioner nedan) kan man påverkan hur data tolkas. Denna tolkning kan i sin tur påverka dataanalysen, datadrivet modellbygge, och i förlängningen besluten som baseras på dessa data. Annoteringar är därför en form av metadata. Annoteringar kan beskriva en eller flera dataposter för en eller flera givarsignaler. Detta sker ofta genom en manuell process som utförs av en expert (peka, klicka, dra, skriva). På grund av att experternas uppmärksamhet varierar består tillgängliga kommentarer vanligtvis av en blandning av observationer av ett symtom (*t*.ex. *avvikande* värden) och förklaringar av orsakerna (t.ex. *elektrisk överspänning*). I vissa fall kan ett mer automatiserat tillvägagångssätt användas för att underlätta en mer konsekvent annoteringsprocess. Oavsett vilken metod som tillämpas så möjliggör annotering av data, när den är framgångsrik, effektiv kommunikation mellan experter med olika tillhörighet, kompetens och ansvarsområden.

Annotering av data har ännu inte studerats i detalj och det finns stora variationer i hur den genomförs. Den befintliga litteraturen om datahantering innehåller endast allmänna uttalanden och rekommendationer om annoteringsproceduren. Akkiraju *et al.* (2020) rekommenderar användning av etiketter (se definition D3.33) och föreslår tre förbättringsområden:

- Avlägsna tvetydigheter genom att skapa och kontinuerligt förbättra definitioner för termer och riktlinjer för annotering.

- Växla upp mängden annoteringar genom att kombinera ett internt team av annoterare och stöd från allmänheten (*crowd work*).
- Underlätta annoteringsprocessen med hjälp av verktyg för maskininlärning (se Kapitel 6).

Av dessa tre är troligen den första och den sista är de enklaste att tillämpa på reningsverk. S.k. *crowd work* är utmanande eftersom (a) data åtminstone tillfälligt måste delas med (en del av) allmänheten och (b) annotering kan kräva detaljkunskaper om tillämpningsområdet eller det specifika sammanhanget för data. Detta kapitel bör vara till hjälp för att implementera den första punkten. Den tredje punkten behandlas främst i Kapitel 6. I följande avsnitt så beskrivs den nuvarande variationen i annoteringsmetoder samt så ges förslag på hur nuvarande annoteringsmetoder kan förbättras.

3.4.1 Syfte

Huvudsyftet med dataannotering är att påverka hur data tolkas eller används i ett pågående eller framtida projekt. Det finns två grundläggande sätt att uppnå detta:

(1) Dataannotering används ofta för att ta bort datapunkter från användarens vy för vissa ändamål, t.ex. när avvikande punkter märks upp på ett sådant sätt att de ignoreras under modellkalibreringen. Ett typiskt exempel är att en datapunkt markeras som en outlier på grund av att den är av låg kvalitet.
(2) Annoteringarna kan rikta uppmärksamheten mot en annoterad datapunkt, t.ex. för att visa på en händelse av intresse.

Båda typerna av annoteringar kan lyfta fram viktig information för framtida analys och därmed påverka hur dessa data tolkas.

3.4.2 Grundläggande definitioner

Definition D3.32: **Annotering av givardata** innebär att addera ytterligare information till dataposter eller tidsstämplar för att underlätta framtida användning och tolkning av givardata.

Det är viktigt att annoteringsprocessen inte tar bort data. Observera också att en enda etikett eller kommentar kan tilldelas ett valfritt antal dataposter, tidsstämplar eller tidsfönster. Därför kan man använda annoteringar för att beskriva något av följande:

- Enskilda dataposter. Till exempel för att markera ett avvikande värde. Se Kapitel 2 för definitionen av en datapost.
- Enstaka tidsstämplar. Till exempel för att ange en händelse som kanske inte sammanfaller med någon datapost (t.ex. händelse vid givarinstallation).
- Flera dataposter, eventuellt producerade av flera givare. Till exempel för att markera ett kalibreringsproblem eller sammanfallande avvikande värden som uppträder samtidigt i flera givare.
- Tidsfönster, som markeras med en start- och sluttidsstämpel. Till exempel för att ange en period av en störning (t.ex. snösmältningsperiod) eller en förändring i driften (t.ex. ammoniumregulator avstängd).

En enda tidsstämpel, datapost eller tidsfönster kan vara försedd med flera etiketter eller kommentarer, antingen av olika experter eller av samma expert med ett annat syfte. Detta beror på flera faktorer, t.ex:

- Flera experter som har olika åsikter och ger olika kommentarer till samma dataposter.
- Olika anteckningar görs för olika syften eller vid olika tidpunkter.

Som ett resultat av ovanstående mångfald av möjliga annoteringar kräver lagring vanligtvis användning av många-till-många-relationer (*many-to-many relationships*) mellan annoteringarna och de tillgängliga dataposterna. Avsnitt 4.4.5.1 i Kapitel 4 visar hur man kan ta hänsyn till detta vid utformningen av en s.k. relationsdatabas.

Definition D3.33: En **etikett eller flagga** (*flag*) är en strukturerad och förutbestämd typ av annotering med följande egenskaper:

- Den används upprepade gånger för att kommentera liknande händelser eller företeelser
- Den innehåller minst två komponenter:
  - En unik identifierare (alfanumerisk sträng) som beskriver typen av händelse eller företeelse.
  - En beskrivning av en typ av åtgärd, kausal händelse eller symtom.

- Dessutom kan etiketten innehålla följande:
  - Beskrivning av det specifika instrumentet.
  - Tidpunkt eller tidsfönster som etiketten avser.
  - Namn på den person skapat annoteringen.
  - Tidpunkt då etiketten skapades.
  - Syftet med annoteringen.
  - Programvara och version som använts för anteckningar.

Den stora fördelen med etiketter är att de är strukturerade och gör det enkelt att annotera data. Nackdelen är att de förutsätter ett gemensamt språk och en gemensam förståelse bland de personer som utvecklar och använder etiketterna.

Definition D3.34: En **kommentar** (*comment*) är ett stycke fritext som används för annotering. En enda kommentar kan kopplas till ett valfritt antal dataposter, ett valfritt antal tidsstämplar eller ett valfritt antal tidsspann.

Till skillnad från etiketter är kommentarer helt flexibla och tillåter anteckningar på ett ostrukturerat och fritt sätt. Detta gör att nya fenomen enkelt kan beskrivas och annoteras. Nackdelen är att kommentarer, om de inte hanteras, kan leda till röriga annoteringar och därmed förhindra en systematisk tolkning av data i ett senare skede.

3.4.3 Strategier för att hantera utmaningar med annoteringar

Det kan vara klurigt att få till en effektivt arbetsflöde för annoteringar. I följande avsnitt ges några allmänna förslag på lösningar.

3.4.3.1 Koppla samman annoteringar med andra metadata och signaler

Akkiraju *et al.* (2020) rekommenderar etiketter (se definition D3.33) för annotering och en snabb återkoppling mellan den som genomför annoteringen och den som använder data. I dag genereras annoteringar ofta för en specifik uppgift eller ett specifikt projekt och finns i isolerade system, t.ex. på en dator som hanteras av en modelleringsexpert. När det är relevant så bör annoteringsarbetet inriktas på att skapa återanvändbara annoteringar, vilket förhindrar upprepning och främjar en gemensam kunskapsbas för dataanmärkare och användare. Utmaningen med att skapa ett effektivt arbetssätt för annoteringar är förstås inte begränsat till reningsverk, vilket för att inspiration kan tas från andra branscher där annoteringar är mer centralt för verksamheten.

3.4.3.2 Flexibla annoteringsmetoder leder till heterogena metadata

Annotering är vanligtvis en flexibel process där en (process)expert skapar metadata som är till hjälp för en annan användare. Annotering används ofta för att lyfta fram händelser som är sällsynta, kanske helt oväntade eller händelser som – aldrig tidigare skådats. Eftersom det finns ett oändligt antal möjliga händelser och förhållanden som skulle kunna vara relevanta att annotera så används fritextkommentarer En instrumenttekniker kan t.ex. ange en unik typ av skada på en givare, eller en processoperatör kan ange en anmärkningsvärd avvikelse från normala driftförhållanden. Dessutom kan de termer som används för att beskriva händelser variera över tid eller från en person till en annan, även om de tillämpas på samma system. En utmaning med fritextkommentarer är dock att deras flexibilitet kan göra det extremt svårt att tolka stora mängder kommentarer automatiskt. Detta begränsar i hög grad nyttjandet av sådana annoteringar.

Utöver alla dessa olika annoteringar kan annoteringar även användas för att markera händelser (t.ex. regnoväder), avsiktliga åtgärder (t.ex. ändring av, drift, kalibreringshändelse) eller en beskrivning av mänskliga observationer eller givarsignaler. En brist idag är att man sällan skiljer på annoteringar som beskriver orsaker (åtgärder, händelser) och effekter (observationer, symtom, givarsignaler).

Den heterogena karaktären hos flexibla annoteringar såsom fritextkommentarer är en del i variationen (*variation*) som är ett av de fyra V:na i big data. Denna typ av variation kräver en balans mellan en enkel annoteringsprocedur och en enkel åtkomst och senare tolkning. Denna balans verkar vara svår att uppnå, vilket också framgår av nedan avsnitt.

3.4.3.3 Kommunikation mellan olika expertdomäner, arbetsteam och över tid

Det är vanligen inte samma person som skriver ned annoteringar som den person som använder dem. En servicetekniker kan t.ex. genomföra ett underhållsarbete på en givare och därefter (an)notera ned detta i ett underhållssystem. Processingenjören är dock den som har nytta av dessa annoteringar för att tolka givarens prestanda datakvalitet. Dessutom kanske den person som annoterar data och den person som tar emot använder dem aldrig träffas, särskilt om historiska data används i ett projekt där annoteringar kan ha gjorts av en tidigare anställd. Detta innebär att det kan vara svårt att förstå det ursprungliga tänkesättet eller avsikten hos den person som noterade uppgifterna. Precis som med de utmaningar som nämns ovan kan effekterna av detta reduceras genom att använda standardiserade arbetsflöden och användargränssnitt för annotering och strukturerade anteckningsformat.

3.4.3.4 Standardisering av annoteringar

Vanligtvis så förespråkas ett strukturerat arbetssätt för annoteringsarbetet för att effektivt kunna nyttja dem i ett senare skede. Standardisering sker främst genom att begränsa antalet termer och kategorier som, var och en ha en unik och standardiserad beskrivning samtidigt som användningen av fritextkommentarer begränsas.

I vissa fall kan fritext helt överges till förmån för strukturerade etiketter, särskilt genom att välja ett begränsat antal beskrivningar som är tillgängliga för den som annoterar (t.ex. rullgardinsmenyer, alternativknappar eller kryssrutor). En risk med ett sådant tillvägagångssätt är att det kan leda till att värdefulla metadata går förlorade, t.ex. när den som annoterar data inte kan hitta ett lämpligt alternativ för att beskriva en ny typ av händelse. Man måste alltså vara uppmärksam på att inte begränsa den som noterar så mycket att hen blir avskräckt och väljer bort värdefull information.

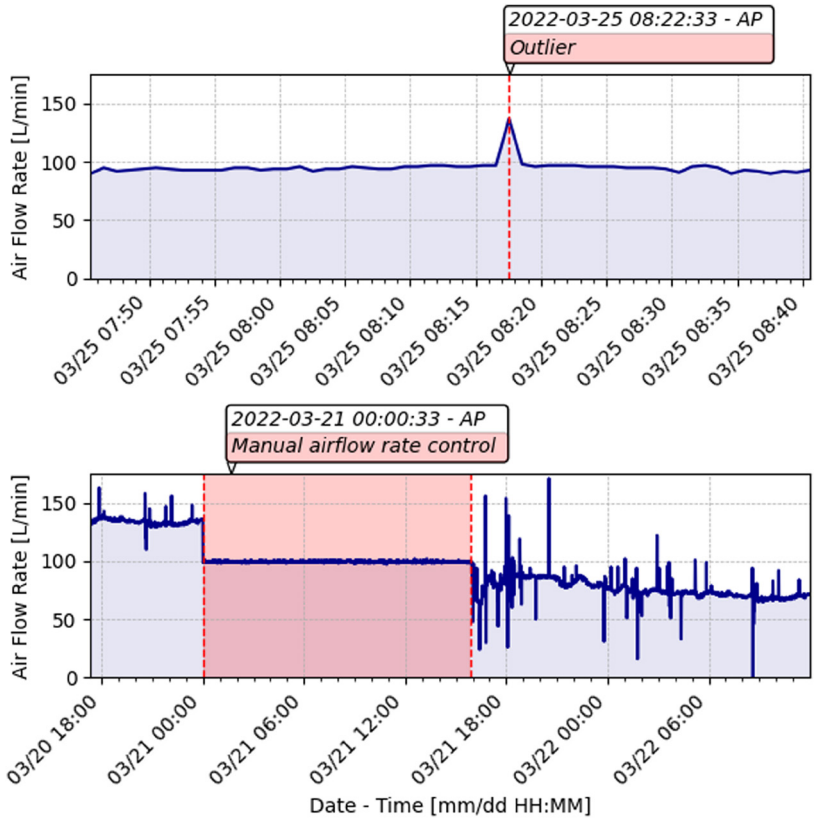
I praktiken behöver en avvägning göras mellan dataanvändarens bekvämlighet (mer struktur) och den annoterande expertens bekvämlighet (mer flexibilitet). Denna balans kräver en avvägning mellan de intressenter som är involverade i tillhandahållandet och användningen av annoterade data och är sannolikt unik för systemet eller teamet. Det är därför mycket viktigt att involvera alla personer som har ett intresse av att delta i dataannoteringsprocessen. I Avsnitt 3.4.4 beskriver vi flera olika sätt som annoteringar kan kategoriseras på.

3.4.3.5 Automatisera annoteringar

Förutom att noggrant välja en standardnivå för annoteringar kan det vara av intresse att använda automatiska verktyg. Det finns flera metoder för att automatiskt upptäcka symtom på försämrad datakvalitet hos en givare (se Avsnitt 3.3.1) eller förekomsten av process anomalier. Automatiska annoteringsverktyg möjliggör en effektivare och mer objektiv annotering och generering av metadata. I Kapitel 6 listas flera metoder som kan användas för detta ändamål.

3.4.4 Kategorisering av annoteringar

Annoteringar kan kategoriseras utifrån flera aspekter såsom tids, plats, förklaringsgrad, informationskälla och själva annoteringsprocessen. Dessa olika aspekter beskrivs i följande avsnitt. Genom att noga välja ut relevanta kategorierna kan man bidra till en systematisk annotering, vilket minskar röran i de metadata som produceras som i sin tur ökar värdet på de annoterade dataseten.



Figur 3.8 Exempel E3.5 – Annoteringar utifrån tid och plats för en tidsserie för luftflödeshastighet i en vanlig aktivslamprocess. En författarakronym, tidsstämpel och annoteringstexten visas. Överst – punktannotering; nederst – kollektiv annotering.

3.4.4.1 Kategorisering utifrån tid och plats

En enda etikett eller kommentar kan kommentera en eller flera dataposter vilket gör att man behöver kunna skilja på följande etiketter och annoteringar:

Definition D3.35: **En punktannotering** är en annotering som enbart nyttjas till en enstaka datapost.

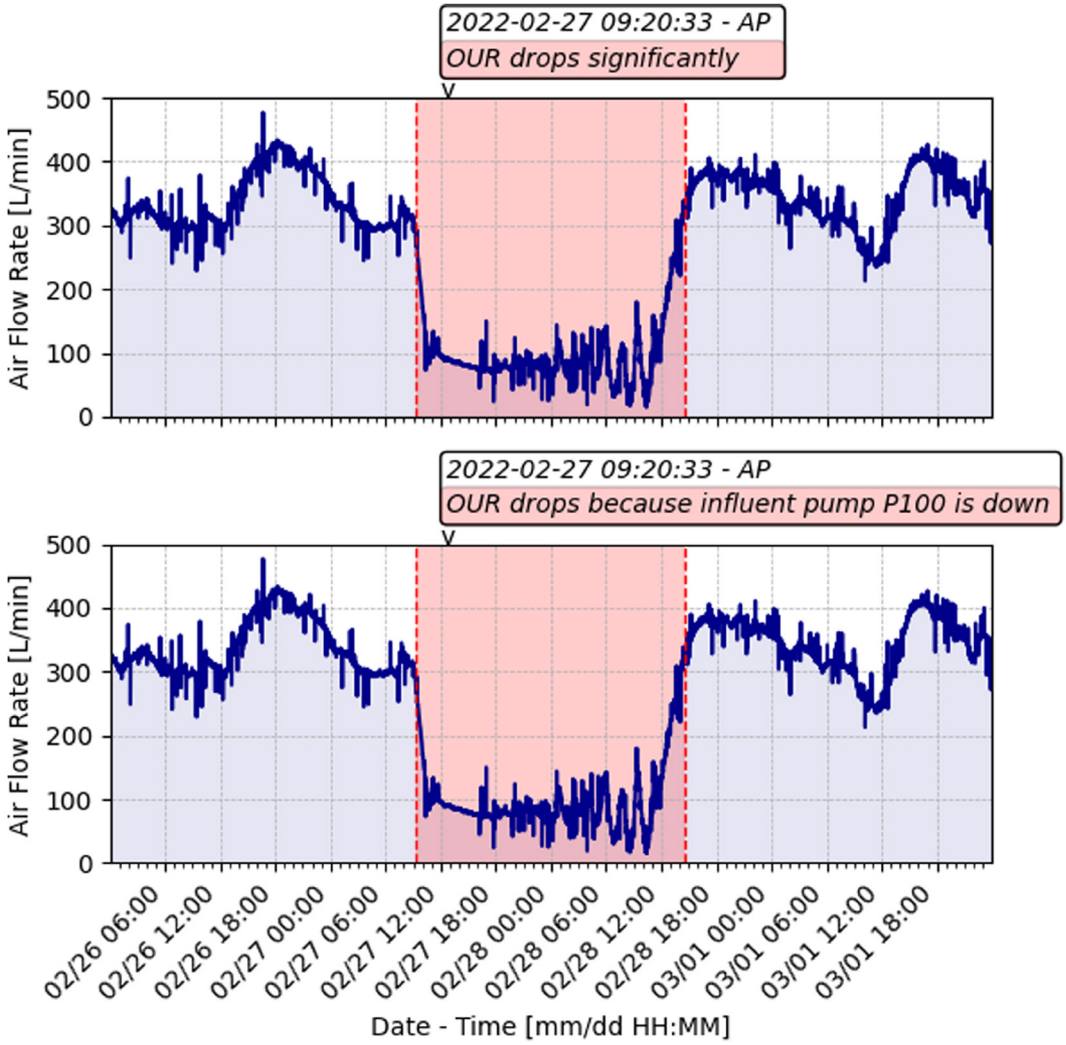
Definition D3.36: **En kollektiv annotering** som används för flera dataposter på en gång. De annoterade dataposterna kan ha samma eller olika tidsstämplar och kan tillhöra endast en signal eller flera signaler. Kollektiva annoteringar används vanligen för händelser som börjar och slutar vid olika tidpunkter.

Figur 3.8 (överst) visar en punktannotering medan Figur 3.8 (nederst) visar en kollektiv annotering.

3.4.4.2 Kategorisering utifrån förklaringsgrad

I många fall är syftet med annoteringen att ge en symtomatisk beskrivning av dataposter, t.ex. för att identifiera avvikande värden, signalförändringar, höga brusnivåer m.m. I vissa fall är det också möjligt att identifiera den orsak som ledde till de observerade symptomen. Alltså kan etiketter och annoteringar klassificeras enligt följande:

Definition D3.37: **En symtomannotering** är en annotering som beskriver observerade symtom men som inte ger någon kausal orsaksförklaring.



Figur 3.9 Exempel E3.6 – annoteringar för en tidsserie med luftflödeshastighet. Överst – symtomannotering; nederst – orsaksannotering. En författarakronym, tidsstämpel och annoteringstexten visas.

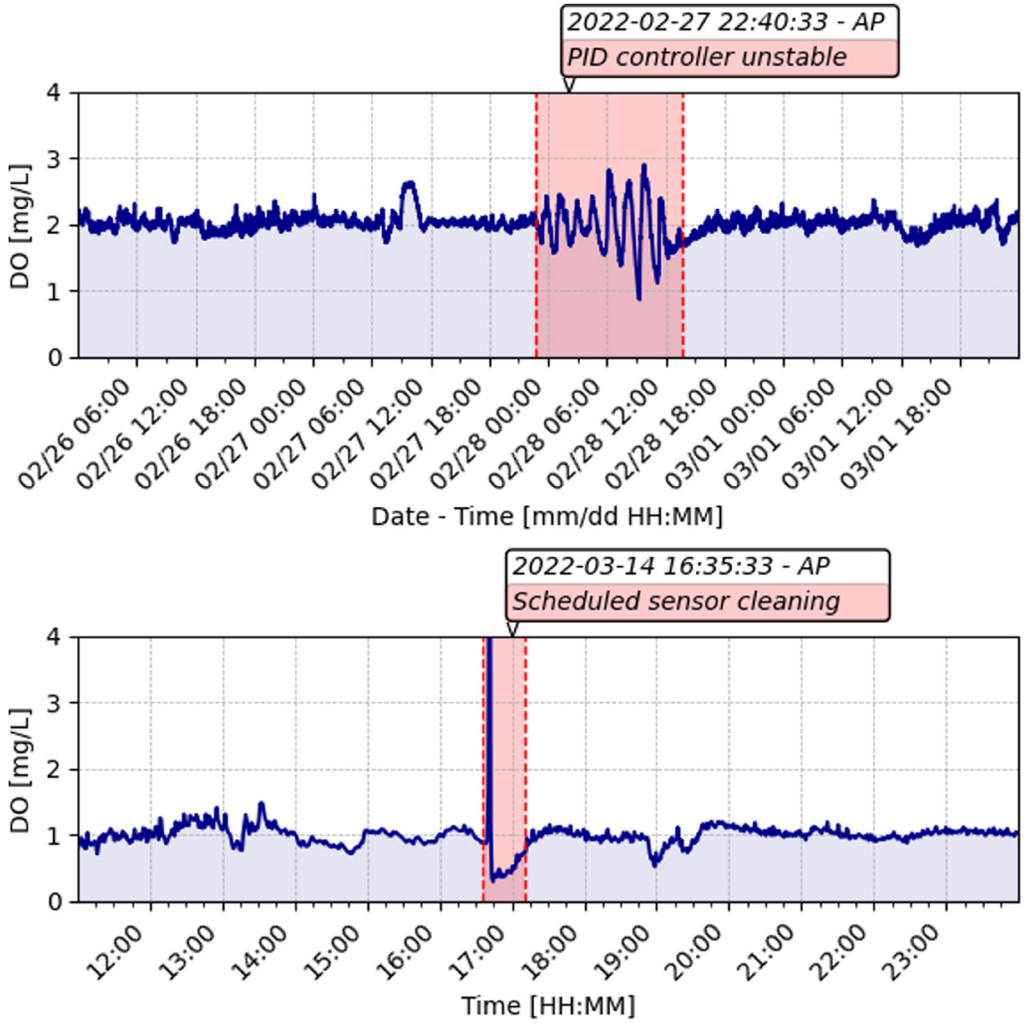
Figur 3.9 (överst) visar en symtomannotering.

Definition D3.38: **En orsaksannotering** är en annotering som innehåller information om grundorsaken till det observerade symptomet.

Figur 3.9 (överst) visar en symtomannotering medan Figur 3.9 (nederst) visar en orsaksannotering för samma fenomen.

3.4.4.3 Kategorisering utifrån informationskälla

Annoteringar kan också klassificeras dess informationskälla. En etikett kan t.ex. genereras utifrån hur väl en massbalans går ihop. I det fallet finns all information som leder till etiketten tillgänglig i samma datauppsättning. Men en etikett kan även baseras på information utanför den tillgängliga



Figur 3.10 Exempel E3.7 – exempel på annoteringar enligt informationens placering för en tidsserie för luftflödeshastighet. Överst – direkt anteckning; nederst – indirekt anteckning. En författarakronym, tidsstämpel och annoteringstexten visas.

datauppsättningen. Till exempel så kan en driftoperatör notera en väderrelaterad händelse såsom regnoväder eller förändringar i en infrastruktur uppströms (t.ex. avloppet) som inte fångas upp som mätningar i data. Etiketter kan därför klassificeras enligt:

Definition D3.39: En **direkt annotering** är en annotering som endast baseras på inspektion eller analys av de dataposter som annoteras.

Definition D3.40: En **indirekt annotering** är en annotering som baseras på information som finns utanför den datamängd som annoteras.

Figur 3.10 (överst) visar en direkt annotering. Det finns ingen annan information än den som direkt kan härledas från själva data. Figur 3.10 (nederst) visar en indirekt annotering. Annoteringen baseras på information som inte är uppenbar genom att endast inspektera de annoterade datapunkterna.

3.4.4.4 Kategorisering enligt annoteringsprocessen

Eftersom annotering kan vara besvärligt, subjektivt och helt enkelt tråkigt att utföra är det inte förvånande stor del av forskningen har fokuserat på automatiserad generering av etiketter eller kommentarer för att beskriva dataposter. Denna forskning fokuserar främst på detektion av fel och anomalier och diagnostik. Kapitel 6 ger en översikt över olika algoritmer som kan användas för detta ändamål. Beroende på användningen av sådana verktyg kan man skilja mellan följande kategorier:

Definition D3.41: En **expertannotering** är en anteckning som tilldelas efter att en mänsklig expert har granskat dataposter utan hjälp av dataanalys eller algoritmiska verktyg.

Definition D3.42: En **datorstödd annotering** är en annotering som tilldelas efter en inspektion av en mänsklig expert av dataposter som valts ut för inspektion genom dataanalys eller algoritmiska verktyg.

Definition D3.43: En **automatiserad annotering** är en annotering som har tilldelats genom dataanalys eller algoritmiska verktyg utan någon mänsklig inblandning.

3.5 SLUTSATSER

- Metadata förekommer i en mängd olika former och kan innehålla:
  - Information om den utrustning och de processer som används vid datainsamlingen (datagenererande process).
  - Information om kvaliteten på de producerade signalerna (datakvalitet).
  - Information om det sammanhang inom vilket data producerades (t.ex. driftförändringar, väder).
- En heltäckande metod för insamling av metadata börjar med en förteckning över de metadata som krävs för aktuella eller framtida projekt eller uppgifter. I det här kapitlet ges en översikt över de mest relevanta typerna av metadata.
- Noggranna annoteringsmetoder kommer att omfatta insamling av metadata för att:
  - Du vet med säkerhet var, när, hur och varför givarsignalerna producerades.
  - Bedöma datakvaliteten som produceras av en givare under hela dess livslängd.
  - Kunna sätta givarsignalerna i det sammanhang där de producerades.

TACKORD

Exempel E3.1 baseras på öppna data (Ohmura *et al.*, *2019b*). Exempel E3.4 tillhandahålls av Juan Baeza (opublicerade data, relaterade till Zhang *et al.*, 2021). Exemplen E3.5 till E3.7 tillhandahölls av modelEAU (Université Laval, Québec, QC, Kanada).

REFERENSER

Akkiraju R., Sinha V., Xu A., Mahmud J., Gundecha P., Liu Z., Liu X. and Schumacher J. (2020). *Characterizing machine learning processes: A maturity framework*. In: Fahland, D., Ghidini, C., Becker, J., Dumas, M. (eds) Business Process Management 2020. Lecture Notes in Computer Science, **12168**, 17–31. Springer, Sevilla, Spain. https://doi.org/10.1007/978-3-030-58666-9_2 (accessed 31 January 2024).

Albuquerque J. S. and Biegler L. T. (1996). Data reconciliation and gross error detection for dynamic system. *AIChE Journal*, **42**(10), 2841–2856. https://doi.org/10.1002/aic.690421014

Alavi N., Warland J. S. and Berg A. A. (2006). Filling gaps in evapotranspiration measurements for water budget studies: evaluation of a Kalman filtering approach. *Agricultural and Forest Meteorology*, **141**(1), 57–66. https://doi.org/10.1016/j.agrformet.2006.09.011

Ba-Alawi A. H., Loy-Benitez J., Kim S. and Yoo C. (2022). Missing data imputation and sensor self-validation towards a sustainable operation of wastewater treatment plants via deep variational residual autoencoders. *Chemosphere*, **288**(3), 132647. https://doi.org/10.1016/j.chemosphere.2021.132647

Chen J., Bandoni A. and Romagnoli J. A. (1997). Robust estimation of measurement error variance/covariance from process sampling data. *Computers & Chemical Engineering*, **21**(6), 593–600. https://doi.org/10.1016/S0098-1354(96)00295-5

Daszykowski M., Kaczmarek K., Vander Heyden Y. and Walczak B. (2007). Robust statistics in data analysis – a review: basic concepts. *Chemometrics and Intelligent Laboratory Systems*, **85**(2), 203–219. https://doi.org/10.1016/j.chemolab.2006.06.016

De Mulder C., Flameling T., Weijers S., Amerlinck Y. and Nopens I. (2018). An open software package for data reconciliation and gap filling in preparation of water and resource recovery facility modeling. *Environmental Modelling & Software*, **107,** 186–198. https://doi.org/10.1016/j.envsoft.2018.05.015

Efron B. and Tibshirani R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, **1**(1), 54–75. https://www.jstor.org/stable/2245500

Grubbs F. E. (1969). Procedures for Detecting Outlying Observations in Samples. *Technometrics*, **11**(1), 1–21. https://doi.org/10.1080/00401706.1969.10490657

HART Communication Foundation (2013). Hart Field Communications Protocol Application Guide Revision 7.1. 194 pp.

Hauduc H., Neumann M. B., Muschalla D., Gamerith V., Gillot S. and Vanrolleghem P. A. (2015). Efficiency criteria for environmental model quality assessment: a review and its application to wastewater treatment. *Environmental Modelling & Software*, **68,** 196–204. https://doi.org/10.1016/j.envsoft.2015.02.004

Isermann R. and Ballé P. (1997). Trends in the application of model-based fault detection and diagnosis of technical processes. *Control Engineering Practice*, **5**(5), 709–719. https://doi.org/10.1016/S0967-0661(97)00053-1

ISO (2003). ISO Water quality – on-line sensors/analysing equipment for water – specifications and performance tests. Standard ISO 15839:2003, International Organization for Standardization, Geneva, CH. https://www.iso.org/standard/28740.html (accessed 31 January 2024).

ISO (2006). Statistics–Vocabulary and symbols – Part 2: Applied Statistics. Standard ISO 3534-2: 2006, International Organization for Standardization, Geneva, CH. https://www.iso.org/standard/40147.html (accessed 31 January 2024).

ISO (2019a). Date and time – representations for information interchange – Part 1: Basic rules. Standard ISO 8601-1:2019, International Organization for Standardization, Geneva, CH. https://www.iso.org/standard/70907.html (accessed 31 January 2024).

ISO (2019b). Date and time – representations for information interchange – Part 2: Extensions. Standard ISO 8601-2:2019, International Organization for Standardization, Geneva, CH. https://www.iso.org/standard/70908.html (accessed 31 January 2024).

McGhee J. (2005). Calculation and Treatment of Errors. In: *Handbook of Measuring System Design*, https://doi.org/10.1002/0471497398

Narasimhan S. and Jordache C. (1999). Data Reconciliation and Gross Error Detection: An Intelligent Use of Process Data. Gulf Professional Publishing, Houston, Texas, 350pp. https://doi.org/10.1016/B978-0-88415-255-2.X5000-9

Ohmura K, Thürlimann C. M., Kipf M., Carbajal J. P. and Villez K. (2019a). Characterizing long-term wear and tear of ion-selective pH sensors. *Water Science and Technology*, **80**(3), 541–550. https://doi.org/10.2166/wst.2019.301 (accessed 31 January 2024).

Ohmura K., Thuerlimann C. M., Kipf M., Carbajal J. P. and Villez K. (2019b). pH Sensor Ageing Experiment (SoDAN-dataset1). https://doi.org/10.5281/zenodo.3229195 (accessed 31 January 2024).

Oppenheim A. V. and Schafer R. W. (1975). Digital signal processing. Englewood Cliffs, NJ: Prentice-Hall, 1975. New Jersey.

Osman M. S., Abu-Mahfouz A. M. and Page P. R. (2018). A survey on data imputation techniques: water distribution system as a use case. *IEEE Access*, **6,** 63279–63291. https://doi.org/10.1109/ACCESS.2018.2877269

Samuelsson O., Björk A., Zambrano J. and Carlsson B. (2018). Fault signatures and bias progression in dissolved oxygen sensors. *Water Science and Technology*, **78**(5), 1034–1044. https://doi.org/10.2166/wst.2018.350

Udugama I. A., Öner M., Lopez P. C., Beenfeldt C., Bayer C., Huusom J. K., Gernaey K. V. and Sin G. (2021). Towards Digitalization in Bio-Manufacturing Operations: A Survey on Application of Big Data and Digital Twin Concepts in Denmark. *Front. Chem. Eng.*, **3**, 727152. https://doi.org/10.3389/fceng.2021.727152

Venkatasubramanian V., Rengaswamy R., Yin K. and Kavuri S. N. (2003a). A review of process fault detection and diagnosis – part I: quantitative model-based methods. *Computers & Chemical Engineering*, **27**(3), 293–311. https://doi.org/10.1016/S0098-1354(02)00160-6

Venkatasubramanian V., Rengaswamy R. and Kavuri S. N. (2003b). A review of process fault detection and diagnosis – part II: qualitative models and search strategies. *Computers & Chemical Engineering*, **27**(3), 313–326. https://doi-org.ornl.idm.oclc.org/10.1016/S0098-1354(02)00161-8

Venkatasubramanian V., Rengaswamy R., Kavuri S. N. and Yin K. (2003c). A review of process fault detection and diagnosis – part III: process history based methods. *Computers & Chemical Engineering*, **27**(3), 327–346. https://doi-org.ornl.idm.oclc.org/10.1016/S0098-1354(02)00162-X

Wang L. Y., Yin G., Zhang J. F. and Zhao Y. (Eds.) (2010). System Identification with Quantized Observations, Birkhäser, Basel, Switzerland.

Widrow B. and Kollar I. (2008). Quantization Noise: Round off Error in Digital Computation, Signal Processing, Control, and Communications. Cambridge University Press, Cambridge.

Zhang C., Guisasola A. and Baeza J. A. (2021). Achieving simultaneous biological COD and phosphorus removal in a continuous anaerobic/aerobic A-stage system. *Water Research*, **190,** 116703. https://doi.org/10.1016/j.watres.2020.116703

Zoubir A. M., Koivunen V., Ollila E. and Muma M. (2018). Robust Statistics for Signal Processing. Cambridge University Press, United Kingdom, 312pp. https://doi.org/10.1017/9781139084291 (accessed 31 January 2024).

doi: 10.2166/9781789065107_0069

Kapitel 4

Strukturerad lagring av signaler samt deras metadata

N. Nicolai[1], K. Villez[2] och Q. Plana[1,3]

[1]*modelEAU, Université Laval, Québec, QC, Kanada*
[2]*Oak Ridge National Laboratory, Oak Ridge, TN, USA*
[3]*Eurecat - Kataloniens teknikcentrum, Barcelona, Spanien*

Nyckelord: datalivscykel, datalagring, databasteknik, fildatabas, relationsdatabas, tidsseriedatabas, historikdatabas, datalager, dataflöde, SQL

Detta kapitel syftar till att ge en översikt över hur signaldata och deras metadata kan struktureras, med betoning hantering och lagring. Specifikt fokuserar det på: (a) nyttan för organisation, (b) vilka data som ska lagras och vad som ska behållas, och (c) datahanteringsmetoder. I detta kapitel ges alltså svar på *var* och *hur* metadata ska lagras på ett effektivt sätt. I Kapitel 3 förklarades *vad som* anses vara metadata. I Kapitlen 5 och 6 beskrivs *hur man samlar in* vissa metadata genom särskilda valideringstester av sensorer (Kapitel 5) eller med dataanalytiska metoder (Kapitel 6).

4.1 LÄSANVISNINGAR

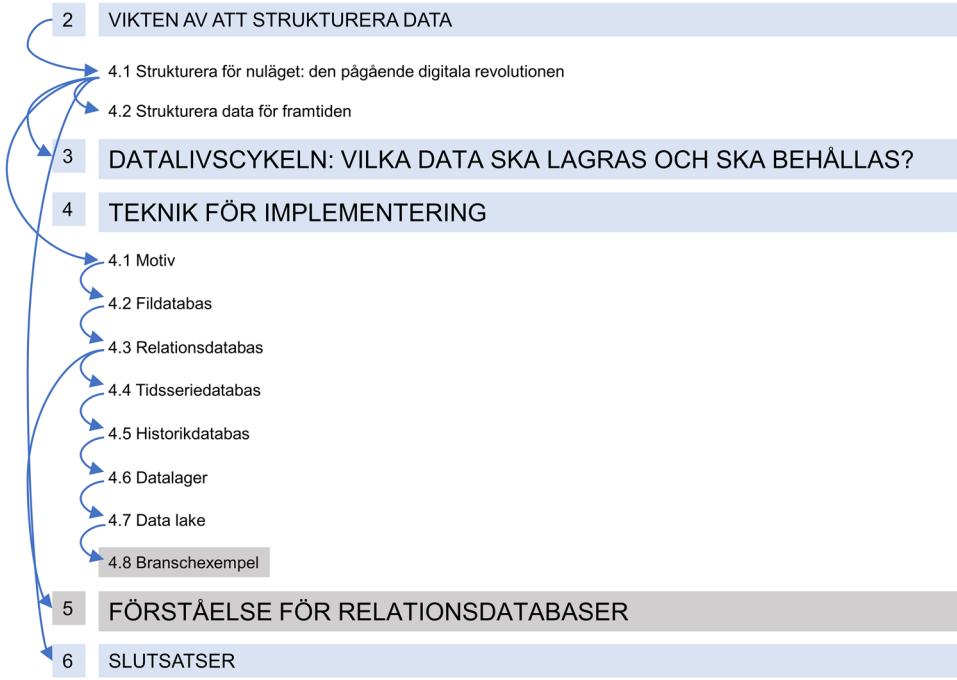
Huvudsyftet med detta kapitel är beskriva behovet av system för att strukturera, lagra och dokumentera insamlade data, även kallat datahantering. Dessutom beskrivs de viktigaste typerna av tillgängliga system, tillsammans med deras viktigaste egenskaper, tillämpbarhet och exempel. Figuren nedan visar kapitlets struktur och beroendena mellan avsnitten. Avsnitt som är markerade med blått eller vitt anses vara grundläggande och nödvändiga för att förstå datahantering och befintliga tekniker. De gråmarkerade avsnitten är kompletterande och användbara avsnitt för att lära sig hur dessa tekniker kan implementeras, vilka dokumenteras med branschexempel.

4.2 VIKTEN AV ATT STRUKTURERA DATA

I de tidigare kapitlen har vanliga typer av data och metadata beskrivits. Detta kapitel beskriver behovet av att prioritera, strukturera och organisera data och illustrerar hur detta kan uppnås i pratiken med hjälp av olika datahanteringstekniker, där datalagringstekniker definieras enligt följande:

Definition D4.1: **Datahanteringsteknik** (*data management technology*) avser en kombination av hårdvara och mjukvara som används för systematisk insamling, beskrivning, utforskning, modifiering, lagring och hämtning av givarsignaler och deras metadata.

© 2025 IWAP. Detta är ett Open Access-bokkapitel som distribueras enligt villkoren i Creative Commons Attribution License (CC BY-NC-ND 4.0) som tillåter kopiering och vidaredistribution för icke-kommersiella ändamål utan derivat, förutsatt att arbetet citeras korrekt (https://creativecommons.org/licenses/by-nc-nd/4.0/). Kapitlet är från boken *Bästa möjliga data från givare i avloppsreningsprocesser*, Kris Villez, Daniel Aguado, Janelcy Alferes, Queralt Plana, Maria Victoria Ruano, Oscar Samuelsson (redaktörer).



Det finns flera fördelar med data som är lätta att lagra, inspektera, och använda. I en ideal värld så är insamling, lagring och datahämtning automatiserad för alla användare. I praktiken kan det bli nödvändigt att balansera hur lätt det är att lagra respektive hämta data. De främsta fördelarna med en systematisk datahantering beskrivs i följande avsnitt där de omedelbara fördelar som följer av förbättrade datahanteringsrutiner diskuteras först, följt av fördelar som kan förväntas på lång sikt.

Definition D4.2: **Datahantering** (*data management*) består av metoder, arkitektur och verktyg för att få tillgång till och kunna hantera leverans av data på ett konsekvent sätt med olika datatyper i en organisation. Syftet med datahantering är att uppfylla organisationens alla datakrav för tillämpningar, applikationer och processer.

4.2.1 Strukturera för nuläget: den pågående digitala revolutionen

4.2.1.1 Data som en resurs för information

Givarsignaler utan sammanhang är i princip oanvändbara (se Kapitel 3). Utan kritisk information såsom enhet, plats där mätningen sker och information om givarunderhåll kan potentiellt informativa tidsserier snabbt bli inaktuella och värdelösa. Idag samlar de flesta reningsverk in givardata som främst används för övervakning och styrning. För att uppfylla dessa behov behövs vanligtvis enbart de senaste historiska dataposterna. Dessutom aggregeras sensordata online över längre perioder (t.ex. timmar till år) för rapportering till externa intressenter och myndigheter. Den låga upplösning som behövs för att tillgodose detta behov kräver ofta inte detaljerad kontextuell information. Nuvarande praxis förutsätter ofta att manuell inspektion är tillräcklig för att säkerställa tillräcklig noggrannhet.

Digitaliseringens intåg, inklusive det ökande antalet givare, användningen av modeller som digitala tvillingar (mekanistiska, empiriska och kombinationer av dessa) för anläggningen kräver ett nytt tillvägagångssätt. Den datahistorik som används för modellbygge och validering förlänger den historiska horisonten från dagar till år, vilket är längre att det går att förlita sig på det mänskliga minnet för kontextuell information. Dessutom kommer denna historiska horisont nu vanligtvis att omfatta

flera tillfällen för underhåll och feldetektioner av givare, som alla kan vara viktiga att ta hänsyn till vid identifiering och driftsättning av modeller. Traditionella och analoga loggböcker blir då ohållbart. Av detta skäl bör dataanvändare ha tillgång till en digitaliserad kontextuell information i form av strukturerade och tillgängliga metadata. Endast genom att ge tillgång till både data och de metadata som krävs för att tolka dem kan dataanvändaren ta bra beslut utifrån den tillgängliga informationen.

4.2.1.2 Bättre beslutsfattande och kostnadsbesparingar

En god praxis och hantering av metadata för datagenerering, kvalitet och sammanhang underlättar tolkning av data. Metadata kan också bidra till att minska osäkerheten vid datatolkningen och därmed minskar skepsisen mot beslut som baseras på data. Detta argument gäller särskilt för beslut som tas långt senare efter datainsamlingen, vilket vanligtvis är fallet för applikationer som inkluderar modeller.

Den största utmaningen för att uppnå detta är kunna integrera flera informationskällor och kunna ta hänsyn till alla tillgängliga data. Ofta innebär detta att olika datakällor (t.ex. historikdatabaser, system för hantering av laboratorieinformation, manuellt förda register, underhållssystem, IoT-plattformar, ekonomisystem m.m.) måste integreras på ett effektivt sätt för effektiv dataexport.

4.2.2 Strukturera data för framtiden

I ovan avsnitt beskrevs kortfattat de fördelar som en förbättrad datahantering direkt kan leda till. I följande avsnitt beskrivs istället de fördelar som en uthållig och förbättrad datahantering kan ge på lång sikt. Notera att vissa av de fördelar som diskuteras är *förväntade* effekter och nyttor, som ännu inte validerats i praktiken.

4.2.2.1 Förberedelser för det okända (s.k. unknown uknowns)

Reningsverkens datahanteringsrutiner bör inte enbart ses över med tanke på nuvarande behov och projekt utan även med tanke på framtida projekt som kan dra stor nytta av en övergripande strategi för data- och metadatahantering.

Mer specifikt återspeglar dagens datahanteringsmetoder ofta bara det omedelbara behovet av information, inklusive processövervakning, styrning och rapportering till intressenter och tillsynsmyndigheter. Det är vanligt att man inte tar hänsyn till hur data kan användas för andra uppgifter, t.ex. anläggningsutformning, eftermontering, utformning och justering av processtyrning. Konsekvensen blir att de stora datamängder som hanteras på reningsverken kan visa sig bli oanvändbara för sådana uppgifter. Detta leder i sin tur till att kompletterande och kostsamma mätkampanjer behöver genomföras för att komplettera data (Hauduc *et al.*, 2009; Rieger *et al.*, 2010). Sådana onödiga merkostnader och tidsfördröjningar kan undvikas genom att förbättra rutinerna för datainsamling och datahantering med dessa andra typer av uppgifter i åtanke.

4.2.2.2 Integrering av interna och externa data ("data brokerage")

Dagens vattenindustri är starkt fragmenterad. I stor utsträckning så produceras och konsumera data i inom vatten- och avloppsreningsverken. Endast en begränsad del av data kommuniceras till berörda parter utanför anläggningarna. På samma sätt är den del av data som samlas in utanför anläggningen begränsad till väderdata eller signaler som produceras av givare i ledningar eller anslutna system (t.ex. fastigheter, demografi, industrier, trafik etc.). Det finns dock en allt starkare strävan efter att se vattencykeln i städer på ett mer holistiskt sätt. För att detta mål ska kunna uppnås behöver VA-organisationerna dela data med varandra, och med andra intressenter. För att de delade uppgifterna ska vara till största möjliga nytta för alla intressenter krävs en integrerad datahantering. Där så är möjligt bör gemensamma metoder och standarder nyttjas.

Ett bra exempel på ett sådant holistiskt genom hela vattencykeln är det arbete som utfördes under Covid-19-pandemin. Pandemin illustrerade hur avloppsdata kan ge en mängd information om epidemiologi på stads- och delstatsnivå. VA-organisationerna bör betrakta sig själva som producenter

av värdefull, och eventuellt kommersiellt gångbar data av nytta för flera samhällstillämpningar. För att utnyttja denna potential kan det krävas samarbete med dataförmedlingssystem med ytterligare krav på datakvalitet och metadata (Torfs *et al.*, 2022).

4.3 DATALIVSCYKELN: VILKA DATA SKA LAGRAS OCH VILKA SKA BEHÅLLAS?

När data väl har samlats in kan de bli inaktuella och värdelösa med tiden. Ett koncept som beskriver den förändringen är tidsvärdet för data. Det enda sättet att motverka denna naturliga tendens är att samla in relevanta metadata (Michener, 2006). Med detta i åtanke är datamängder bäst att betrakta som resurser med ett "bäst-före-datum". För att nyttan av tillgängliga data ska maximeras bör data hanteras i flera faser under sin livslängd. De olika faserna listas nedan, specifikt för givardata.

(1) *Planering*. Detta är den fas då man planerar för nya givare, köper in, installerar, kalibrerar och ansluter dem till organisationens datainsamlingsstruktur. Under denna fas bör man försöka samla in och strukturera metadata om den datagenererande processen (se Kapitel 3, Avsnitt 3.2.2). Den avsedda lagringsplatsen för data och hur länge de kan och bör lagras bör också definieras vid denna tidpunkt. Nästa fas och tillhörande aktiviteter bör också definieras, i den mån det är möjligt.
(2) *Datainsamling (acquisition)*. Detta är den fas då data aktivt samlas in. Denna fas kan ha en definierad längd (t.ex. längden på en intensiv mätkampanj) eller vara odefinierad i längd (t.ex. en givare som installeras som del i en reglerkrets).
(3) *Underhåll och databehandling (pre-processing)*. Beroende på datainsamlingens syfte kan rådata behöva modifieras till primärdata inklusive (a) modifiera signalen för att förbättra datakvaliteten (t.ex. brusreducering, borttagning av felaktiga mätningar) och (b) aggregerade värden som härleds från originalmätningen för ett specifikt ändamål (t.ex. månatliga medelvärden för rapportering). Denna fas kan omfatta definition, implementering och uppdatering av de databehandlingssteg som anses nödvändiga för att uppnå och bibehålla önskad datakvalitet. Inspiration kan hämtas från verktyg för versionshantering av programvara (definition nedan) för att underlätta detta.
(4) *Nyttjandefas*. Någon gång efter att givaren har installerats används data för det avsedda ändamålet. Beroende på tillämpning så kan data användas långt efter datainsamlingen (t.ex. modellbaserad processdesign eller utbyggnation av reningsverket), direkt efter insamlingen (t.ex. realtidsreglering) eller någonstans däremellan (t.ex. dataanalys för felsökning av driftstörningar). Det är viktigt att syftet och den tänkte nyttjandefasen definieras tidigt, redan i planeringsfasen.
(5) *Åtkomst/delning/återanvändning*. Det är mycket troligt att de data som samlas in kommer att återanvändas efter att den ursprungliga dataanvändningsfasen har påbörjats eller till och med avslutats. Ny, tidigare oavsiktlig användning av data bör beskrivas väl och leder tillbaka till fas 4 (nyttjandefasen). I idealfallet bör nya tillämpningar av data undersökas med avseende på tillgänglighet, lämplighet och en kontroll av om datakvaliteten är tillräcklig för den nya föreslagna användningen.
(6) *Arkivering*. När insamlade data inte längre ändras eller används bör man överväga att arkivera dem för framtida bruk. De bör då dokumenteras för att säkerställa att en framtida användare kan använda och tolka uppgifterna på ett tillförlitligt sätt.
(7) *Rensning*. Data som har passerat sitt bäst-före-datum försämrar kvaliteten på andra data som finns i samma lager och bör därför tas bort i god tid (t.ex. så bör data rensas för processer som inte längre är i drift). I vissa fall kan det finnas en policy för datalagring som kan ge vägledning om hur rensning görs på rätt sätt.

Definition D4.3: **Versionshantering av mjukvara** (*software versioning*) avser processen att dokumentera stegvisa förändringar i programvara och tilldela unika versionsnamn eller unika

versionsnummer för att referera till unika, historiska tillstånd av programvaran. Målen med versionshantering inkluderar (a) effektivt samarbete mellan programmerare, (b) hantera kompatibilitet med annan programvara, och (c) garantera reproducerbarhet av resultat som erhållits med programvaran.

För att hantera detta stegvisa tillvägagångssätt på ett strukturerat och organiserat sätt rekommenderas starkt en strategi för datahantering (*data governance*). Strategin bör omfatta aspekter som metadatagenerering, metoder för datalagring, definition av syfte och förväntade analyser, och den bör helst definieras innan ett nytt projekt eller en ny datainsamling inleds. I förlängningen kan detta leda till att data hanteras väl i nuläget och förbereds för framtida återbruk. Det är viktigt att inkludera lämpliga back-up-system, policyer för användaråtkomst och beslut om vilka typer av data som kan/bör lagras i tredjepartsapplikationer (t.ex. molnlagring kontra lokal lagring). De potentiella fördelarna med en datahanteringsstrategi är många även om det idag saknas exempel från reningsverk som i praktiken har kvantifierat fördelarna. I framtiden kan denna kunskapslucka komma att fyllas om data blir en kommersiellt gångbara vara som är tillgänglig för branschens intressenter (Avsnitt 4.2.2.2).

Definition D4.4: En tillämpad **datahanteringsstrategi** ("data governance") innebär att organisationen har en verksamhetsstyrning för att hantera data (planering, övervakning och tillämpning).

Definition D4.5: En **policy för datalagring** beskriver vilka data som ska lagras eller arkiveras, var det ska ske och hur länge de ska lagras innan de ska rensas bort.

Faserna insamling, användning, arkivering och rensning kan också påverkas av en policy för datalagring som riktar sig till flera intressenter och omfattar olika aspekter. Det är därför viktigt att en sådan policy är väl definierad. Här ges några förslag. Först och främst är det viktigt att beakta olika lagkrav som anger hur länge vissa uppgifter ska finnas tillgängliga för framtida efterlevnad. Policyn kan tillåta att vissa data lagras i en reducerad eller komprimerad form (t.ex. lägre tidsupplösning), medan andra data ska lagras som rådata. Den kan också definiera lämpliga lagringsmetoder och därmed rekommendera eller kräva specifika lagringsmedier och utesluta andra enheter för datalagring (t.ex. backup-servrar, externa hårddiskar etc.). Notera att med molntjänstleverantörer finns billig långtidslagring tillgänglig med acceptabel hastighet och flexibilitet för datahämtning. Det bör också vara tydligt från början (dvs. innan arkiverade data återställs) om datasetet lagras som originaldata, vilka metadata som ingår, om det består av aggregerade data och om det innehåller dokumentation. Datalagring är också viktigt vid planering för säkerhetskopior (*back-up*). Säkerhetskopieringsplaner använder ofta en s.k. kaskadmetod när det gäller lagringstid för att förhindra att onödiga säkerhetskopior blir kvar för evigt. En policy för datastyrning kan t.ex. kräva att dagliga säkerhetskopior sparas i 7 dagar, veckovisa säkerhetskopior i 4 veckor, månatliga säkerhetskopior i 12 månader och årliga säkerhetskopior i 7 år.

4.4 TEKNIK FÖR IMPLEMENTERING

4.4.1 Motiv

Om data efterfrågas för ett beslut eller projekt är sannolikheten stor att en eller flera databaser behöver sökas igenom. Ibland underlättas detta med ett användarvänligt gränssnitt (*frontend-application*), andra gånger så hanteras export från databaser istället med ett programmeringsspråk för databaser (*querying language*). Oavsett exportteknik så kan det vara till stor hjälp att förstå hur data och metadata är strukturerade i en databas. Databastekniken är inte ny. Under en relativt lång tid i informationsteknikens historia var datalager med filer (*flat files*) och relationsdatabaser de bästa lösningarna för strukturerad lagring av en organisations data. I takt med att dataflödena växte (*big-data eran*) så utvecklades nya databastekniker som kunde hantera den ständigt ökande datavolymen, variationen och hastigheten på ett tillförlitligt och effektivt sätt.

Resten av detta kapitel introducerar viktiga tekniker som används för att lagra och strukturera data. Olika databastyper beskrivs, inklusive styrkor och svagheter med respektive teknik. Eftersom fokus i denna rapport ligger på givardata och metadata kommer endast relevanta databastekniker att diskuteras. I detta avsnitt visas hur metadata kan komma till användning och vilka databastekniker som är bäst lämpade för att strukturera och utnyttja kontextuell information.

Ett särskilt avsnitt beskriver s.k. historikdatabaser (*process historians*), även om dessa i huvudsak är mjukvaruapplikationer som baseras på grundläggande databastekniker såsom relationsdatabaser eller tidsseriedatabaser. Historikdatabaserna erbjuder dock särskild funktionalitet för lagring av givardata och metadata som är användbara på reningsverk.

Eftersom givardata och metadata i slutändan kan användas för beslut kommer även moderna databastekniker som används i allt större utsträckning inom dataanalys och datavetenskap att diskuteras (t.ex. datalager och datasjöar). Den stora variationen i olika databastekniker illustreras med relevanta exempel från industriella tillämpningar.

4.4.2 Fildatabas (”flat-file database”)

En fildatabas består vanligtvis av en textfil med liten eller inga beskrivningar om dess innehåll *(minimal mark-up*). Den är enkel i sin struktur, ofta utformad som en tabell, för att lagra data. Gemensamt för denna typ av datalagring är att dataposterna har ett enhetligt format och lagras sekventiellt i databasfilen. På grund av denna enkelhet finns det vanligtvis inga strukturerade relationer mellan dataposterna. Fildatabaser är oftast fristående, vilket innebär att en enda fil används för att lagra data och att den inte är beroende av andra filer eller program för att strukturera och tolka data.

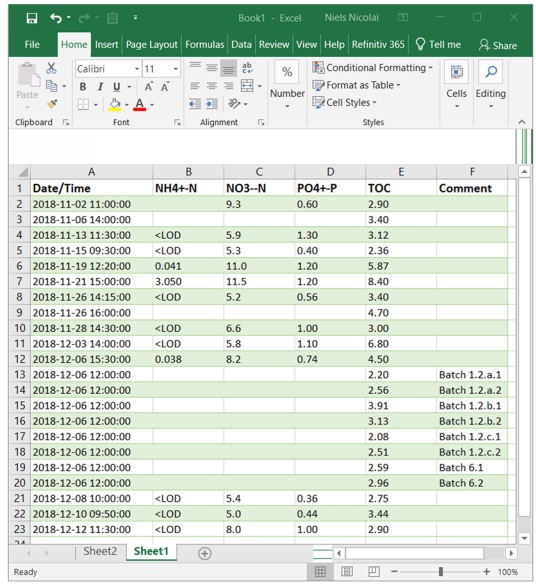
I Tabell 4.1 sammanfattas de största fördelarna och nackdelarna med en fildatabas. Fildatabaser används ofta för att lagra tabelldata, inklusive tidsserier. Dess enkla datastruktur kräver endast några få färdigheter för att spara och söka efter de data som lagras. Normalt räcker kunskaper i hur man använder kalkylbladsprogram (t.ex. MS Excel, Google Sheets, LibreOffice Calc etc.). När kalkylbladen blir stora minskar dock hastigheten för att hämta data och därmed också möjligheten att extrahera värdefull information. Databaser med filer förlitar sig på en linjär sökmetod för att söka efter data, där varje datapost i filen måste kontrolleras för en matchning. Den här metoden fungerar bra för en liten lista men är inte skalbar för större dataset där en mängd irrelevanta data måste analyseras innan rätt post kan lokaliseras. På grund av bristen på strukturerade relationer mellan data innehåller flatfilsdatabaser dessutom vanligtvis inte någon omfattande beskrivning i form av metadata som är starkt beroende av datarelationer.

Ett exempel på en fildatabas är en CSV-fil (*Comma Separated Values*) som är en radorienterad fildatabas. Varje textrad i sådana filer representerar en datapost som beskrivs av ett eller flera fält som är åtskilda med ett kommatecken (,) (se Figur 4.1). Om datavärdena innehåller kommatecken används alternativa avgränsare, t.ex. tabb (\t) (filformat för tabbseparerade värden (TSV)) eller fasta bredder (positionellt filformat där början av ett fält anges som en förskjutning från början av en rad). Modifiering av dataposterna i en filsdatabas kan göras med hjälp av grundläggande textredigerare, specialkodade applikationer eller med hjälp av kommersiellt tillgängliga applikationer. I det senare fallet lagras och komprimeras data som standard i det filformat som används i programvaran (t.ex. xlsx, GSHEET, ods etc.), men kan lätt konverteras till andra filformat som också stöder fildatabaser.

Tabell 4.1 Fördelar och nackdelar med fildatabaser.

Fördelar	Nackdelar
• Enkel datastruktur gör det lätt att förstå data. • Många yrkesverksamma känner till de vanliga programvaruapplikationerna för datahantering (textredigerare, kalkylblad etc.). • Kräver liten implementeringsinsats. • Snabba sökningar för små datamängder.	• Inte avsedd att ha strukturerade relationer eller hierarkier mellan dataposter, vilket komplicerar lagring av metadata. • Data är inte indexerade, vilket gör datahämtningen långsam för stora datamängder. • Fasta maximala filstorlekar. Större datamängder måste delas upp på flera filer. • Filer lagras ofta lokalt och är därför inte allmänt tillgängliga för att manipulera data av flera samtidiga användare.

```
Date/Time,NH4+-N,NO3--N,PO4+-P,TOC,Comment
2018-11-02 11:00:00,,9.3,0.60,2.90,
2018-11-06 14:00:00,,,,3.40,
2018-11-13 11:30:00,<LOD,5.9,1.30,3.12,
2018-11-15 09:30:00,<LOD,5.3,0.40,2.36,
2018-11-19 12:20:00,0.041,11.0,1.20,5.87,
2018-11-21 15:00:00,3.050,11.5,1.20,8.40,
2018-11-26 14:15:00,<LOD,5.2,0.56,3.40,
2018-11-26 16:00:00,,,,4.70,
2018-11-28 14:30:00,<LOD,6.6,1.00,3.00,
2018-12-03 14:00:00,<LOD,5.8,1.10,6.80,
2018-12-06 15:30:00,0.038,8.2,0.74,4.50,
2018-12-06 12:00:00,,,,2.20,Batch 1.2.a.1
2018-12-06 12:00:00,,,,2.56,Batch 1.2.a.2
2018-12-06 12:00:00,,,,3.91,Batch 1.2.b.1
2018-12-06 12:00:00,,,,3.13,Batch 1.2.b.2
2018-12-06 12:00:00,,,,2.08,Batch 1.2.c.1
2018-12-06 12:00:00,,,,2.51,Batch 1.2.c.2
2018-12-06 12:00:00,,,,2.59,Batch 6.1
2018-12-06 12:00:00,,,,2.96,Batch 6.2
2018-12-08 10:00:00,<LOD,5.4,0.36,2.75,
2018-12-10 09:50:00,<LOD,5.0,0.44,3.44,
2018-12-12 11:30:00,<LOD,8.0,1.00,2.90,
```



	A	B	C	D	E	F
1	Date/Time	NH4+-N	NO3--N	PO4+-P	TOC	Comment
2	2018-11-02 11:00:00		9.3	0.60	2.90	
3	2018-11-06 14:00:00				3.40	
4	2018-11-13 11:30:00	<LOD	5.9	1.30	3.12	
5	2018-11-15 09:30:00	<LOD	5.3	0.40	2.36	
6	2018-11-19 12:20:00	0.041	11.0	1.20	5.87	
7	2018-11-21 15:00:00	3.050	11.5	1.20	8.40	
8	2018-11-26 14:15:00	<LOD	5.2	0.56	3.40	
9	2018-11-26 16:00:00				4.70	
10	2018-11-28 14:30:00	<LOD	6.6	1.00	3.00	
11	2018-12-03 14:00:00	<LOD	5.8	1.10	6.80	
12	2018-12-06 15:30:00	0.038	8.2	0.74	4.50	
13	2018-12-06 12:00:00				2.20	Batch 1.2.a.1
14	2018-12-06 12:00:00				2.56	Batch 1.2.a.2
15	2018-12-06 12:00:00				3.91	Batch 1.2.b.1
16	2018-12-06 12:00:00				3.13	Batch 1.2.b.2
17	2018-12-06 12:00:00				2.08	Batch 1.2.c.1
18	2018-12-06 12:00:00				2.51	Batch 1.2.c.2
19	2018-12-06 12:00:00				2.59	Batch 6.1
20	2018-12-06 12:00:00				2.96	Batch 6.2
21	2018-12-08 10:00:00	<LOD	5.4	0.36	2.75	
22	2018-12-10 09:50:00	<LOD	5.0	0.44	3.44	
23	2018-12-12 11:30:00	<LOD	8.0	1.00	2.90	

Figur 4.1 Data som presenteras i Tabell 2.3 (Kapitel 2) uttryckt som en fildatabas. Vänster: CSV-filformat. Höger:. xlsx kalkylblad i Microsoft Excel-editorn.

Ytterligare noteringar: Även om de inte är utformade för ändamålet så kan fildatabaser användas för att representera relationer och samband mellan datapunkter. Vanligtvis görs detta med hjälp av ett extra textfält för anteckningar. Exempelvis kan följande kommentar *recordID Y is related to the event captured* by *recordID X* manuellt läggas till som en kommentar till *record* X för att ange dess relation till record Y (Tabell 4.1).

Vissa hävdar att fildatabaser inte bör betraktas som riktiga databastekniker (Dunning & Friedman, 2014). När det gäller datahantering är användningen av fildatabaser ofta begränsad till att dela rådata och aggregerade resultat mellan applikationer och slutanvändare. Utifrån det perspektivet så fungerar filen som en mellanliggande datalagring med det enda syftet att överföra data till en mer förfinad databasteknik. Lagring i fildatabaser kan verka föråldrad eftersom den inte har alla de fördelar som mer omfattande datalagringstekniker har. De flesta organisationer har dock fortfarande en hel del rådata lagrad som filer. Tänk på kalkylblad som innehåller mätdata från experimentella försök, CSV-filer som genereras av ett fristående mätinstrument och så vidare.

4.4.3 Relationsdatabaser

Relationsdatabaser är en beprövad teknik för att lagra stora datamängder tillsammans med strukturerade metadata. I detta avsnitt ges en övergripande översikt av denna databas. En mer detaljerad beskrivning, liksom några relevanta exempel, ges i Avsnitt 4.4. I Tabell 4.2 sammanfattas de största fördelarna och nackdelarna med relationsdatabaser.

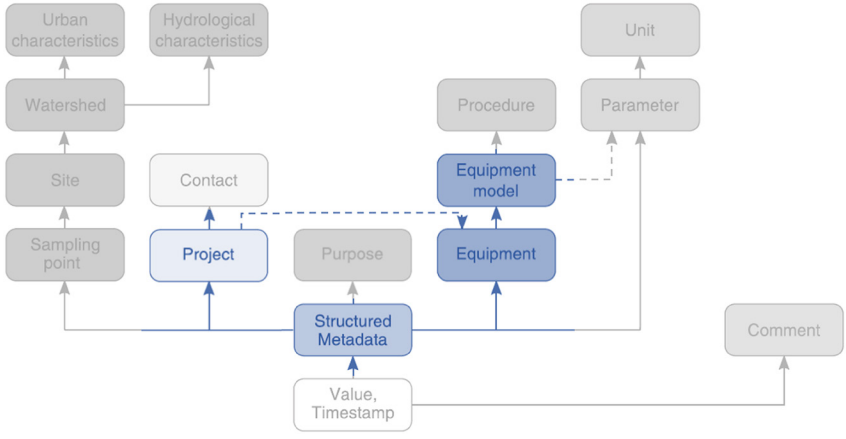
Relationsdatabaser lagrar data utifrån deras inbördes samband eller “relationer”, som visualiseras som tabeller för användaren. I motsats till datalagring i filer fångar en relationsdatabas upp de associationer som finns mellan olika tabeller. Varje rad i en tabell representerar en datapost med en unik identifierare (*primary key*). Varje kolumn i en tabell representerar ett attribut eller en egenskap som är relevant för tabellen. I stället för att lagra alla data i en enda stor tabell med många attribut, delas data upp i flera tabeller utifrån ett specifikt ämne, d.v.s. utifrån hur användarens perspektiv. På

Tabell 4.2 Fördelar och nackdelar med relationsdatabaser.

Fördelar	Nackdelar
• Relationer i data (och metadata) fångas intuitivt upp för att drastiskt minska dataredundansen. • En mycket strukturerad databasdesign med specifika regler finns på plats för att bevara dataintegriteten. • Det finns ett standardiserat språk (SQL) för att ställa frågor till relationsdatabaser oberoende av vilket system för hantering av relationsdatabaser (RDBMS) som används i bakgrunden. • De flesta RDBMS använder en klient-server-arkitektur som enkelt kan skalas när det gäller samtidiga användare och bearbetningsbelastning. • Olika behörighetsnivåer kan konfigureras i RDBMS för att skydda dataåtkomst.	• En väl utformad datamodell är en förutsättning för en väl fungerande databas. • För att göra en databasförfrågan krävs grundläggande kunskaper i SQL och den logiska datamodell som ligger till grund för databasen. • I idealfallet administreras och finjusteras relationsdatabaser aktivt (t.ex. säkerhetskopieringsplan, lagringspolicy, säkerhetsuppdateringar, resursallokering, designändringar etc.) • Utvecklad för transaktionsdata och därför inte ursprungligen avsedd för effektiv hantering av tidsseriedata (t.ex. långsammare lagrings- och hämtningshastigheter, komplexa SQL-frågor för tidsseriejustering och subsampling).

så sätt minskas dataredundansen för att säkerställa dataintegriteten och underlätta underhållet. Det sistnämnda kallas för att normalisera databasen (*database normalization*). För att detta ska fungera så behöver sambandet mellan olika värden i databasen beskrivas. Typiskt så behöver sambandet mellan en unik identifierare matchas med de värden för olika attribut som finns med i flera tabeller.

Hur data separeras i olika tabeller och hur relationerna mellan dessa tabeller definieras avgörs av den logiska datamodell som används för att strukturera databasen. Datamodeller är mycket tillämpningsspecifika. Det finns dock generiska principer för att utforma logiskt strukturerade, tillförlitliga och flexibla relationsdatabaser oberoende av tillämpning (Hernandez, 2021). Många relationsdatabaser stöder även halvstrukturerade data (t.ex. fildatabaser såsom XML, JSON) som komplement till alfanumeriska data (mätvärden, text, datatid, GPS-koordinater etc.). Figur 4.2 visar ett exempel på en sådan relationsstruktur där varje ruta representerar en tabell och pilarna de identifierade relationerna mellan dem.



Figur 4.2 Logisk datamodell som är relevant för lagring av vattenkvalitetsdata i en relationsdatabas. Blåmarkerade är de relationer som är relevanta för exemplet i Figur 4.3 och Tabell 4.3. Anpassad från Plana *et al.* (2019) och Therrien *et al.* (2020).

Exempel på detta: En stor skillnad mellan datalagring i en fildatabas och relationsdatabaser är hur data hämtas och ändras. För att söka i en fil krävs att man förstår filens fysiska uppbyggnad. En relationsdatabas kan däremot sökas med hjälp av ett *conditional statement* som går att läsa och förstå av användaren. Det kräver dock att användaren känner till relationerna mellan tabellerna i databasen. Det formella språk som används för att utföra dessa sökningar i relationsdatabaser kallas för *Structured Query Language* (SQL) och har sitt ursprung i relationsalgebran. Typiska SQL-satser kombinerar relationsoperationer som *select*, *project* och *join* för att manipulera en eller flera tabeller och skapa en annan tabell med resultatet i (se Figur 4.3 och Tabell 4.3). Notera att denna resulterande tabell inte lagras i databasen som sådan, utan att den istället är resultatet av en fråga. Resultattabellen kan fortfarande lagras som en virtuell tabell baserad på frågan, i stället för att definieras i själva datamodellen, en s.k. vy (*view*).

Ytterligare noteringar: Förutom den logik som används för att definiera en datamodell kan det vara viktigt att också ta hänsyn till hur informationen lagras för att databasen ska kunna komma åt den (*physical design*). Detta är särskilt viktigt när man arbetar med enorma databaser som består av miljontals dataposter och relationer. En term som ofta dyker upp i diskussioner om prestandahantering i databaser är *indexering*. På samma sätt som ett sakregister (*index*) i en teknisk bok kan användas för att hitta specifika termer, kan en databas använda index på specifika kolumner för att snabbt hitta rader i tabeller. Ett index kan ses som en speciell tabell som innehåller kolumner för att lokalisera rader i en annan tabell, tillsammans med information som beskriver var raderna är fysiskt lagrade. Den standardstruktur som används för databasindexering är vanligtvis baserad på B-träd (Beaulieu, 2020). Det faktum att relationsdatabaser kan använda sig av index ger dem en enorm fördel jämfört med fildatabaser som är begränsade till linjära sökningar.

Även om relationsmodellen och SQL är generell koncept, behövs ett särskilt program för att manipulera datafilerna och den fysiska hårdvara som de lagras på. Detta, och mycket mer, hanteras av ett RDBMS (*Relational Database Management System*). Det finns många RDBMS med både öppen och sluten källkod samt kommersiella och kostnadsfria RDBMS. Bland de mest populära

```
SELECT DISTINCT dbo.equipment.Equipment_ID AS 'Equipment ID',
    dbo.equipment.Equipment_identifier AS 'Name',
    dbo.equipment_model.Manufacturer AS 'Manufacturer',
    dbo.equipment_model.Equipment_model AS 'Model'
FROM dbo.metadata
    LEFT OUTER JOIN dbo.project ON dbo.metadata.Project_ID = dbo.project.Project_ID
    LEFT OUTER JOIN dbo.equipment ON dbo.metadata.Equipment_ID = dbo.equipment.Equipment_ID
    LEFT OUTER JOIN dbo.equipment_model ON dbo.equipment.Equipment_model_ID = dbo.equipment_model.Equipment_model_ID
WHERE dbo.project.Project_name='pilEAUte'
```

Figur 4.3 Exempel på en grundläggande SQL-fråga som kombinerar select-, project- och join-operationer för att hämta all aktiv utrustning som ingår i pilEAUte-projektet. Anpassad från Plana *et al.* (2019) och Therrien *et al.* (2020).

Tabell 4.3 Resultatmängd för frågan i figur 4.3.

Utrustningens ID	Namn	Tillverkare	Modell
6	pH_006	WTW	Sensolyt 700 IQ
15	pH_015	HACH	Hach pHD sc Online Process pH-sensor
26	Spektro_002	s::scan	spektro::lyser
38	Ammo_004	s::scan	ammunition::lyser
41	Varion_001	WTW	VARION PlusAN/A
…	…	…	…

Endast de fem första posterna presenteras här. Anpassad från Plana et al. (2019) och Therrien et al. (2020).

företagslösningarna finns MySQL, SQL Server, Oracle-databasen, DB2, SQLite och PostgreSQL, var och en med olika för- och nackdelar.

Ändå har alla samma grundläggande funktioner: (1) att definiera strukturella regler enligt en relationsdatamodell, (2) att möjliggöra samtidig inmatning och ändring av data med hjälp av SQL-standarden och (3) att skydda och säkra databasen utifrån användardefinierade roller med motsvarande behörighetsnivåer. Även om RDBMS-programvaran används för att hantera en relationsdatabas är interaktionen med data inte nödvändigtvis begränsad till RDBMS. Tack vare den utbredda tillgången på programvarugränssnitt är det möjligt att använda allmänna programmeringsspråk (t.ex. C++, Java, Python och R) för att direkt ställa frågor till de flesta relationsdatabaser.

4.4.4 Tidsseriedatabas

Tidsseriedata kännetecknas av mätningar av en och samma variabel över tid (ett givarvärde och en tidsstämpel). Tack vare de allt lägre kostnaderna för datalagring genereras tidsserier från alla möjliga anläggningar i stora volymer. Därför finns det ett behov av lösningar som effektivt lagrar och hämtar data från en eller flera tidsserier inom ett visst tidsintervall. Även om fildatabaser och relationsdatabaser kan användas, och används, för detta ändamål finns det särskilda tidsseriedatabaser som utnyttjar särdragen hos tidsseriedata för snabbare och mer intuitiv lagring och hämtning av data (Dunning & Friedman, 2014; Nielsen, 2019). För detta ändamål drar tidsseriedatabaser vanligtvis nytta av en eller flera av följande egenskaper:

(1) De dataposter som ingår i en tidsserie är kronologiskt ordnade. Denna ordning måste dessutom bevaras för att uppgifterna ska vara informativa.
(2) Tidsseriedata representerar ofta givardata. Interpolering mellan dataposter är därför i många fall acceptabelt.
(3) Många tidsserier består av regelbundet återkommande dataposter, det vill säga med ett konstant tidsintervall mellan efterföljande poster.
(4) För många tillämpningar är det mer sannolikt att de senaste dataposterna i en tidsserie är av större intresse jämfört med poster som lagrades för länge sedan.
(5) Tidsserier från stabila driftförhållanden är sällan de mest betydelsefulla. Mer intressanta är perioder med dynamiska förändringar som trender, transienter, extremvärden och så vidare.
(6) Tidsserier som har registrerats uppdateras sällan i efterhand.
(7) Enskilda dataposter är ofta mindre informativa jämfört med korta sekvenser eller aggregerad statistik i en tidsserie.
(8) Tidsseriedata kännetecknas vanligtvis av mycket få oberoende relationer i jämförelse med relationsdata.

Tabell 4.4 sammanfattar de största för- och nackdelarna med tidsseriedatabaser.

Med tanke på tidsseriernas speciella karaktär ger relationsdatabaser inte den höga prestanda och skalbarhet som ofta krävs för tidsserieanalys. Ett viktigt skäl är att relationsdatabaser inte naturligt tar hänsyn till kontinuiteten i tidsseriedata, vilket gör det komplicerat att ställa frågor som bygger på att flera tidsserier, som ofta registreras med olika frekvenser, ska sammanföras. Standard SQL har heller inte funktionalitet för modifierad tidsupplösning (*time-bucketing*) (TimescaleDB, 2017).

Tidsseriedatabaser kan vara en bra teknik att överväga om de flesta datasökningarna är data för olika tidsintervall. En begränsning är dock att tekniken för tidsseriedatabaser (ännu) inte är lika mogen som för relationsdatabaser. Tidsseriedatabaser kan vara fristående lösningar eller byggas ovanpå andra befintliga databasteknologier. Ett exempel på den förstnämnda tekniken är InfluxDB, medan TimescaleDB och OpenTSDB är byggda på en relationsdatabas (PostgreSQL) respektive en NoSQL-databas (Apache HBase eller Google Bigtable). Databaserna behöver implementeras på ett sätt som stödjer tillämpningen.

Tabell 4.4 Fördelar och nackdelar med tidsseriedatabaser.

Fördelar	Nackdelar
• Speciellt framtagen för högpresterande, skalbar lagring och hämtning av stora volymer tidsseriedata. • En mängd olika tidscentrerade funktioner tillhandahålls vanligtvis för tidsserieanalys. • Tidsseriedata kan enkelt kopplas ihop med metadata som lagras i en relationsdatabas. • Några av de mest populära tidsseriedatabaserna är licensfria och har öppen källkod.	• Inga standarder bland leverantörer av lösningar för tidsseriedatabaser, vilket gör att sökningar och funktioner är programspecifika. • Integration av tidsseriedatabaser med industriell utrustning (PLC, SCADA, DCS etc.) kräver speciallösningar. • De grafiska applikationer som användaren kan använda för att interagera med databasen har begränsad funktionalitet. Anpassade applikationer måste utvecklas om icke-specialister ska kunna analysera data.

Ytterligare noteringar: De tekniska detaljerna om hur prestandan kan förbättras för lagring av tidsserier beskrivs inte här. Det finns dock några viktiga punkter som är intressanta att förstå. Till exempel det faktum att tidsseriedata tenderar att visualiseras som långa listor (dvs. en rad för varje datapost i serien) medan ett brett format (dvs. flera dataposter i samma tidsserie fördelas över olika kolumner på ett smart sätt) är mer lämpligt, eftersom det drastiskt kan minska antalet rader som behöver skannas om man också tar hänsyn till seriens kronologiska karaktär. Den förlustfria komprimeringen (definition nedan) av efterföljande tidsstämplar och värden med minsta möjliga antal bitar kan också minska den lagrade datavolymen (Chiarot & Silvestri, 2021). Genom att ha en automatiserad lagringspolicy i en tidsseriedatabas kan äldre dataposter lagras med lägre tidupplösning eller aggregeras över tid. Genom att lagra nyare dataposter i cache-minnet i stället för att lagra dem direkt på disk kan fördröjningen slutligen minskas avsevärt.

Eftersom varje tidsseriedatabas använder olika sätt att modellera och lagra data finns det för närvarande inget standardspråk för att ställa frågor till tidsseriedatabaser. Vissa kommersiella lösningar använder ett språk som helt baseras på standard SQL och lägger till anpassade funktioner för att hantera tidsserier, medan andra har infört ett helt nytt och oberoende frågespråk (Figur 4.4).

```
-- Microsoft SQL Server
SELECT timestamp, value
FROM pilEAUte_values
WHERE timestamp >= 1662909036 AND timestamp <= 1663513836
AND signal_name = 'temperature'

-- TimescaleDB
SELECT time, temperature
FROM pilEAUte_values
WHERE time > now() - '7 days'

-- InfluxDB
from(bucket: "pilEAUte_values")
    |> range(start: -7d)
    |> filter(fn: (r) => r._measurement == "temperature")
```

Figur 4.4 Frågeställning för de senaste 7 dagarna för en temperatursignal som lagrats i en relationsdatabas (överst) (dvs. SQL-kompatibel) och två olika tidsseriedatabaser (mitten och nederst) (dvs. anpassat frågespråk). Observera att det är komplicerat att ställa frågor om en tidsserie för ett specifikt intervall i relationsdatabasen eftersom ett explicit tidsintervall förväntas. Dessutom används enheten Unix epoktid för att lagra, och därmed också hämta, tidsserierna, vilken inte är intuitiv.

Oavsett språk så erbjuder de flesta tidsseriedatabaser ändå samma funktionalitet. Till exempel finns det funktioner som utför filtrering, urval, aggregering, omvandling, urval, interpolering utifrån ett tidsintervall.

För att kunna ställa frågor om tidsserier krävs ofta att användaren i förväg definierar ett tidsintervall för vidare analys. Om det önskade intervallet för analysen bestäms på förhand kan det leda till flera iterationer innan den faktiska intressanta perioden har fastställts. Detta gäller särskilt när det handlar om tidsserier som registrerats under långa tidsperioder. Hur intressanta perioder i tidsseriedata kan identifieras på ett effektivt sätt är ett aktivt forskningsområde. Genom att använda detektionsalgoritmer (*motif discovery algorithms*) kan intressanta mönster identifieras i samma eller olika tidsserier. Baserat på det ursprungliga mönstret som ges av användaren (t.ex. en initialt vald intressant period eller baserat på en fritt formulerad skiss) kan en fråga konstrueras för att hitta liknande mönster och rangordna dem efter deras matchning (Correll & Gleicher, 2016; Torkamani & Lohweg, 2017). Därmed förbättras sättet att ställa frågor om tidsserier avsevärt.

Eftersom tidsseriedatabaser är mycket specifika för lagring av tidsserier kan de inte direkt användas för att lagra kontextuell information som rör den registrerade tidsserien (t.ex. typ och ID för den sensor som användes för att generera tidsserien). Eftersom sådana metadata mestadels är statisk information med flera relationer är det bättre att lagra dem i en mer strukturerad databas. Det är därför som tidsseriedatabaser gör det möjligt att koppla samman dynamiska tidsserier med statisk information som lagras i en fildatabas eller relationsdatabas. På så sätt kan kontextuella metadata användas i användargränssnittet för att ställa en fråga som avser rätt tidsserie.

4.4.5 Historikdatabaser

4.4.5.1 Allmän beskrivning

Historikdatabaser (*process historians*) är kommersiella mjukvaror som kan användas för datasamling, lagring, organisering och hämtning av givardata från industriella processer. Eftersom dessa mätningar registreras över tid är de i själva verket tidsserier. Observera dock att inom processautomation används ofta termen *tagg (tag)* för att hänvisa till en specifik tidsserie. Varje tagg i historikdatabasen har ett unikt och strukturerat namn som gör det möjligt att identifiera vara data producerats (exempelvis reningsverket, försedimentering, ammoniumgivare – db.arv.fs.nh001). Trots den lilla skillnaden i terminologi och användning är historikdatabaser i grund och botten faktiskt tidsseriedatabaser. På samma sätt kan de vara fristående lösningar eller byggas ovanpå en annan databasteknik (t.ex. en relationsdatabas). Ändå kan historikdatabaser göra mycket mer jämfört med generiska tidsseriedatabaser (Middleton, 2022).

Huvudsyftet med en historikdatabas är att samla in mätdata i realtid och lagra dem på ett effektivt sätt för senare tolkning och analys. Många historikdatabaser lagrar också ytterligare data för att ytterligare beskriva driftsstörningar. Till exempel kan historikdatabaser lagra larm- och händelsedata som är viktiga för felsökning. Aggregerade funktioner kan också beräknas automatiskt och lagras för att ge sammanfattande statistik för vissa variabler. En annan typ av data som lagras av de flesta historikdatabaserna är s.k. konfigurationsdata. Dessa data beskriver konfigurationen av programvaran och maskinvaran och används mest för att administrera historikdatabasens infrastruktur.

I Tabell 4.5 sammanfattas de största för- och nackdelarna med processhistorikdatabaser.

Ytterligare noteringar: Det finns några egenskaper som gör en historikdatabas särskilt lämplig till datainsamling på reningsverk. För det första är en historikdatabas kompatibel med annan programvara och maskinvara som används för industriell automation (t.ex. DCS, SCADA, MES, ERP, PLC etc.). Detta innebär att användarna inte behöver utveckla automatiska program för överföring av data på låg nivå mellan dessa datakällor och historikdatabasen (*data collectors*). Tidigare hade varje leverantör ett eget kommunikationsprotokoll för sina egna produkter. Detta innebar att industriella mjukvarulösningar, t.ex. historikdatabaser, ofta valdes utifrån tillgången till specifika gränssnitt. I och med införandet av standardiserade kommunikationsprotokoll (t.ex. ODBC, OPC UA, MQTT) har gränssnitten fått mindre betydelse för valet av historikdatabas. Även om interoperabiliteten har förbättrats genom åren är det viktigt att betona att precis som med de flesta andra industriella

Tabell 4.5 Fördelar och nackdelar med historikdatabaser.

Fördelar	Nackdelar
• Speciellt framtagen för högpresterande, skalbar lagring och hämtning av stora volymer tidsseriedata. • Programvarupaketet är i hög grad anpassat till industriella processer eller till och med specifika sektorer (t.ex. vatten- och avloppsindustrin). • Databasleverantören tillhandahåller ett helt system av ytterligare programvarukomponenter som kan organisera och analysera lagrad data. • Datainsamlare är lättillgängliga och följer branschens standardprotokoll. • Grafiska användargränssnitt gör det enkelt för användaren att söka efter rätt data utan att skriva någon SQL-kod. • Robust lösning för att garantera 100 % drifttid. • Teknikkonsulterna och supporten känner väl till processindustrins behov.	• Högt pris. Vanligtvis måste användaren betala per 100–1000 taggar på årsbasis. • Att sätta upp en historikdatabas är kostsamt, vilket kan öka projektkostnaden avsevärt. • System utan insyn (“black-box”) minskar flexibiliteten och konfigurationsmöjligheterna. Support från leverantören behövs för anpassade tillägg. • Beroendet av andra programvaruprodukter (Windows Server, Microsoft SQL Server etc.) kräver ytterligare licenser, vilket medför ytterligare kostnader. • Leverantörsinlåsning kans ske på grund av den (1) betydande investering i både kostnad och tid som krävs vid installation beroendet av leverantören för framtida uppdateringar och tillägg, och (3) beroendet av äldre system.

automationslösningar är historikdatabaserna fortfarande i hög grad optimerade för Microsofts programvaruteknik (Siemens SIMATIC Process Historian, GE Proficy Historian, AVEVA Historian, OSIsoft Pi, Rockwell Automation FactoryTalk Historian, Honeywell Process History Database, ABB S+ Historian, Emerson DeltaV Historian).

Historikdatabaser används för att lagra data för säkerhetskritiska system med stora ekonomiska omsättningar. Detta kräver en robust mjukvaru- och hårdvaruinfrastruktur med tillräcklig redundans, så att 100 % drifttid kan garanteras. En reservstrategi ingår ofta när fjärrdatakällor inte kan komma åt historikdatabasen på grund av kommunikationsproblem. I så fall lagras data tillfälligt i en lokal buffert tills historikdatabasen är online igen (“store and forward cache”). En annan funktion är “multi-tiering” där flera historikdatabaser konfigureras för att samla in data med olika frekvenser för olika ändamål. Till exempel kan det finnas en nivå-1 databas för rådata som samlar in mätdata med relativt höga frekvenser för driftens tillämpningar. Ovanpå den sitter en nivå 2-databas, som är placerad utanför anläggningens säkra interna nätverksinfrastruktur och fungerar som ett lager för aggregerade data eller mätdata med låg frekvens, som ett datalager som är specifikt för alla processoperativa data (Avsnitt 4.4.6). En nivå 2-databas kan användas för processoptimering på lång sikt, och därför kan den vara ansluten till internet för fjärråtkomst. Att integrera historikdatabasen i företagets nätverk passar in i den senaste tidens trend att integrera system för operativ teknik (OT) med system för informationsteknik (IT), det vill säga IT/OT-konvergens.

Ursprungligen var historikdatabaser endast avsedda för effektiv lagring och hämtning av mätdata av ett fåtal användare. Idag erbjuder moderna historikdatabaser många fler funktioner som är användbara för en bredd av personalen på reningsverket. I många fall utvecklades dessa ytterligare programvarukomponenter av leverantörerna och slutanvändarna själva för att lösa specifika problem, och med tiden integrerades och finjusterades de gradvis som en del av det paket som var tillgängligt för andra användare. De flesta historikdatabaser har t.ex. ett högspecialiserat grafiskt användargränssnitt eller en webbapplikation för att visualisera och analysera data av många samtidiga användare. Färdiga analyspaket kan användas för att processoptimering, massbalanser, kartlägga energianvändning, undersöka larmloggar, övervaka säkerhetssystem och så vidare. Dessutom kan automatiska utskick ske för rapportering eller avvikelsedetektioner, hantering av användare och deras behörigheter och mycket mer. Beroende på leverantör finns ibland även

programvara för underhållssystem och verktyg för annotering av data och för lagring av metadata (t.ex. OSIsoft AssetFramework).

4.4.5.2 Komprimeringstekniker som används i historikdatabaser

Stora processindustrier och reningsverk kan ha miljontals taggar som behöver sparas i historikdatabasen. För att hantera dessa stora volymer av realtidsdata har de flesta historikdatabaser en rad olika tekniker för datahantering. Precis som i generiska tidsseriedatabaser används förlustfria komprimeringstekniker för att lagra data med minsta möjliga antal bitar utan att förlora information. Dessutom finns det vanligtvis flera metoder för att få ännu högre komprimeringsgrad genom att utesluta redundanta mellanliggande dataposter från lagring (se även Avsnitt 3.2.4.2 i Kapitel 3). Det finns två motiv till att komprimera data. För det första så minskar det fysiska utrymme som behövs för att lagra data. För det andra så begränsas den bandbredd som krävs för dataöverföring till andra applikationer eller databaser. I denna rapport används följande definitioner:

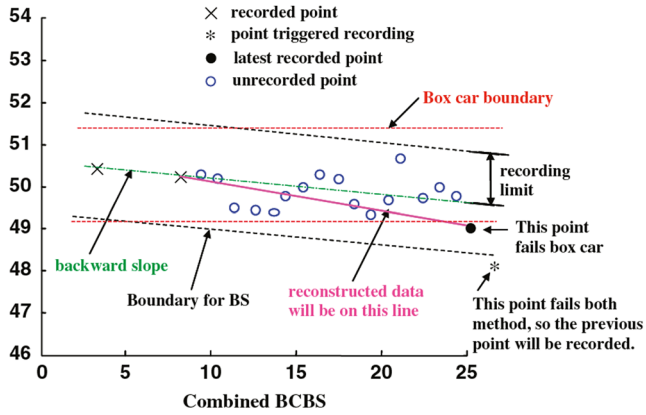
Definition D4.6: **Datakomprimering** är processen för att koda information med hjälp av färre bitar än den ursprungliga representationen. Detta minskar den volym som krävs på den fysiska lagringsenheten.

Definition D4.7: **Förlustfri komprimering (*lossless compression*)** är datakomprimering som säkerställer att originaldata kan rekonstrueras perfekt från de komprimerade data, det vill säga lagring utan informationsförlust. Komprimering uppnås genom att eliminera redundanta delar i data.

Definition D4.8: **Datakomprimering med förlust (*lossy compression*)** är datakomprimering som innebär att viss information går förlorad. Därför kan man bara konstruera en approximation av originaldata baserat från komprimerade data. Även om information går förlorad i processen kan höga komprimeringsnivåer ofta uppnås med små informationsförluster.

Den vanligaste metoden för att komprimera mätdata i realtid är att endast lagra dataposter med en tagg som skiljer sig avsevärt från den tidigare registrerade datapunkten. Detta kallas dödbandskomprimering (*delta storage or report on exception)*. Ett sätt att definiera storleken på dödbandet är att sätta fasta användardefinierade gränser för den tillåtna förändringen av ett värde innan det registreras (*deadband compression or Box-Car-compression)*. Alternativt kan storleken på förändringen mellan två på varandra följande registrerade datapunkter användas som grund för att avgöra vilka datapunkter som ska kasseras (*backward-slope compression or rate of change compression)*. Det är också möjligt att kombinera dessa båda metoder (se Figur 4.5). En annan populär komprimeringsmetod är svängdörrsalgoritmen (*swinging-door-algorithm*). Notera att när dessa komprimeringsalgoritmer används förloras de data som inte registreras för alltid. Tidsintervallet mellan de dataposter som bevaras varierar också från en post till en annan. En tidsserie med högre upplösning kan rekonstrueras från den komprimerade tidsserien med hjälp av enkel linjär interpolation (Choudhury *et al.*, 2008).

Det är intressant att de tidiga komprimeringsteknikerna med förlust är relativt enkla jämfört med de komplexa algoritmer som också används för bild-, ljud- och videokomprimering (t.ex. *wavelet*-komprimering som används i JPEG-standarden). En avgörande skillnad är hur komprimeringsalgoritmerna tillämpas. De enkla, och därför mycket intuitiva, metoder som används av en historikdatabas fungerar i realtid på punktbasis (dvs. de beslutar att behålla eller förkasta en ny mätning när den blir tillgänglig). Mer komplexa komprimeringsalgoritmer kräver vanligtvis en större sekvens, dvs. med en definierad start- och slutpunkt, innan komprimeringen kan utföras. Observera också att de diskuterade komprimeringsmetoderna inte garanterar en stor komprimering, eftersom de är starkt beroende av tidsseriens varians. Taggar som uppvisar ett lågt signal-brusförhållande kommer att ha låga komprimering. Dessutom kan komprimeringsmetoder med förlust försämra datakvaliteten genom att ändra sammanfattande statistik (exempelvis medelvärden), införa obefintliga tidsfördröjningar samt olinjäriteter i de dynamiska data (Thornhill *et al.*, 2004). Därför är det viktigt att förstå och dokumentera hur de algoritmer som används av historikdatabasen fungerar som kompletterande metadata. Det kan i själva verket vara värt att utvärdera effekterna av dessa komprimeringsalgoritmer för specifika signaler innan de införs.



Figur 4.5 Illustration av den kombinerade komprimeringsalgoritmen Box-Car-backward-slope (BCBS) som används av vissa historikdatabaser. När den aktuella datapunkten inte uppfyller kriterierna för både Box-Car- och backward-slope-metoden (*), lagras punkten före den aktuella mätningen (-) (Choudhury *et al.*, 2008).

Observera att man kan skilja mellan komprimering med förlust på kort sikt och aggregering på lång sikt. I det senare fallet kan medelvärden som tagits under långa tidsperioder beräknas, lagras och rapporteras. För drift och analys av anläggningar tenderar dock aggregering att komprimera data så mycket att de blir ointressanta för driftspersonalen.

4.4.6 Datalager

Ett datalager (*data warehouse*) är ett centralt datalager som lagrar mestadels aggregerade data och metadata från flera olika databaser. Syftet med ett datalager är primärt att möjliggöra övergripande analyser för hela verksamheten (*business intelligence*). Av den anledningen kategoriseras datalager som en analysdatabas i motsats till de operativa databastekniker som diskuterats tidigare. Tabell 4.6 sammanfattar de största för- och nackdelarna med datalager.

Nyttan med ett datalager illustreras bäst utifrån beslutshierarkin för databaserade beslut i större organisationer. På den lägsta nivån i en organisation finns användare och ledning för de operativa systemen. Deras roll är att driva den dagliga verksamheten, vilket inkluderar driftteknikerochdriftingenjörer som rapporterar processtörningar, dokumenterar underhållsaktiviteter, loggar händelser och avvikelser i loggböcker, håller reda på utrustningens status och så vidare. I sin vardag nyttjar de driftdata som lagras i olika driftsystem som ofta är applikationer som är anslutna till en relationsdatabas. Exempel på operativa system är historikdatabas, laboratorieinformationssystem (LIMS), underhållssystem, kalkylblad, loggböcker, IoT-plattformar och så vidare. Fokus för dessa system ligger i hög grad på utförande och på att lösa kortsiktiga problem relaterade till tillfälliga händelser. En nivå högre upp i hierarkin finns organisationens ledning som integrerar, analyserar och sammanfattar den information som lagras i de operativa systemens databaser. Baserat på de insikter som genereras på den här nivån fattas beslut på medellång till lång sikt. På ett reningsverk kan detta t.ex. innebära att olika processdata används för processutvärdering och för att identifiera framtida motåtgärder. I slutändan är det den högsta ledningen i en organisation som nyttjar data i kombination med extern information för att fatta strategiska beslut. Ett vanligt exempel är en genomförbarhetsanalys för utbyggnad av reningskapaciteten utifrån historisk belastning, tidigare reningskapacitet, befolkningsprognoser, klimatförändringar och så vidare. Den hierarki som presenteras ovan illustrerar tydligt att beslut på hög nivå *är beroende av korrekta data* från reningsverkets alla olika delar. Eftersom företagsledningen har andra behov, tidshorisonter och strukturer jämfört med den operativa driften så krävs ett annat system för att ledningen ska kunna

Tabell 4.6 För- och nackdelar med datalager.

Fördelar	Nackdelar
• Centraliserar olika data så att den kan användas av icke-specialister. • Intuitivt lagrar både verksamheten mätdata såväl som tillhörande metadata. • Det är enkelt att integrera olika datakällor (“data collectors”) Utifrån standardiserade databasprotokoll. • Tillgänglig som hanterad tjänst (PaaS) på publika moln med integrationer för datainmatning, bearbetning och analys. • Transparens och flexibilitet om hur data struktureras. • Strukturerat dataschema gör specifika sökningar möjligt som i sin tur bidrar till snabbt och datadrivna beslutsstöd.	• Centraliserad datalagring kan resultera i monoliter (fristående system) som är för komplexa för att underhålla och söka i. • Det är tidskrävande att definiera den datamodell som används i datalagret, vilket kräver samverkan mellan flera olika experter och domäner på reningsverket. • Den struktur som data lagras i (“data warehouse schema”) kommer troligtvis förändras för att inkludera nya datakällor allt eftersom att verksamheten utvecklas. • Dataanalys av stora datamängder försvåras på grund av långa söktider, begränsad beskrivningar och fördefinierade datastrukturer. • Data bör vara mycket strukturerade. Dock är detta sällan möjligt i praktiken eftersom ostrukturerade data kommer från många källor (IoT-sensorer, textdokument, video- och ljudfiler, och kan vara av olika format (JSON, PNG, PDF etc.).

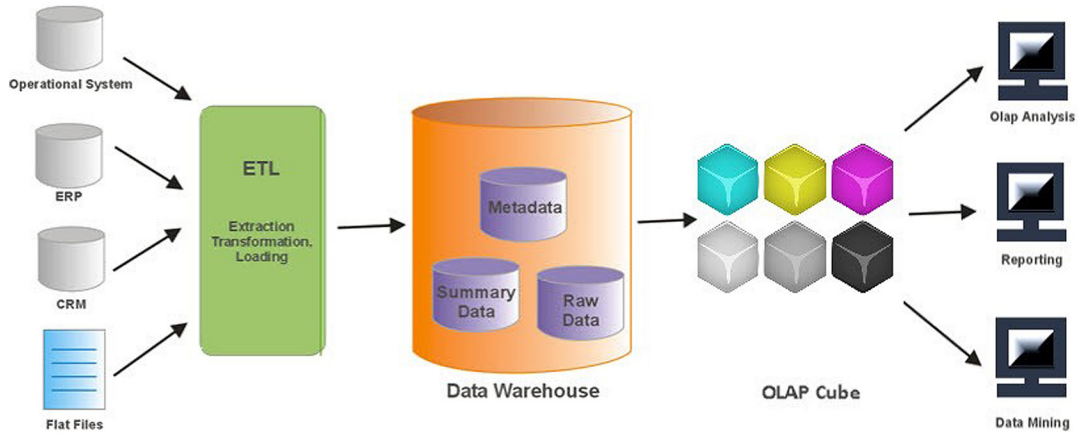
lagra och interagera med organisationens analytiska data. Att tillgängliggöra (*democratizing*) driftdata för icke-specialister är precis vad ett datalager utformades för. Notera att det är avgörande att förstå bakgrund och sammanhang för de data som används vid strategiska beslut, det vill säga metadata. Styrkan med ett datalager är att det lagrar både driftdata och de metadata som krävs för en grundlig analys. Hur ett datalager fungerar beskrivs bäst utifrån organisationens dataflöde.

Definition D4.9: Ett **dataflöde (*data pipeline*)** är en sekvens av processer och verktyg som används för att överföra och omvandla data mellan ett eller flera system och ett målarkiv. Processen bidrar till att strukturera data för att underlätta kommande analyser.

Varje enskilt dataflöde utgår från en datakälla, vanligtvis från de databaser som lagrar givardata som är den primära källan till information (se Figur 4.6). I nästa steg sorteras användbara data ut från dessa olika datakällor. Notera att detta ofta är komplicerat eftersom data produceras i olika mjukvaror från olika leverantörer och därmed levereras i olika format och på isolerade system.

Därefter behöver rådata förädlas till intressant information på ett standardiserat format som är sökbart och enkelt att analysera. Beslutsfattare kan behöva detaljerade data för att besvara en specifik frågeställning, men är nästan aldrig intresserade av enskilda dataposter i rådata. Många databaser med driftdata aggregeras därför och innehåller sammanfattande statistik, till exempel resultat på medellång till lång sikt. När rådata har aggregerats och transformerats så laddas den resulterande informationen in i datalagret. En sådan ETL-process (*extract-transform-load*) genomförs vanligen på ett batchliknande sätt, där en viss mängd rådata måste samlas in transformeringen kan genomföras. Så kallade strömmande ETL-processer kan dock konfigureras så att transformationen görs så snart nya data finns tillgängliga. Det senare är användbart när man är intresserad av att analysera stora dataflöden i realtid (Figur 4.6).

På samma sätt som i relationsdatabaser så måste strukturen i ett datalager utformas utifrån slutanvändarens perspektiv. Ett datalagers struktur syftar till att maximera insikter från data med minimal sökinsats, till skillnad från de mera driftnära systemen som baseras relationsdatabaser som istället prioriterar dataintegritet (korrekta data) och enkelt underhåll (dvs. normaliserade databaser utan redundans). Datalager är med andra ord utformade för att tillgängliggöra data för slutanvändaren. Detta resulterar ofta i datalager som avsiktligt innehåller överflödig information som förenklar sökningar, förbättrar prestanda och kan hantera förändringar i datastrukturen (*denormalized databases*).



Figur 4.6 Schematisk bild av det dataflöde som används för datalager. Hämtad från Panoply (2023).

Den metodik som används för att utforma den underliggande datalagerstrukturen kallas dimensionsmodellering eller datalagrets datamodell (*dimensional modelling*). Det går att skilja på två typer av modeller utifrån den fysiska implementeringen: stjärnmodellen (*star schema*) och OLAP-kuben (*Online Analytical Processing*). Stjärnmodellen implementeras i en allmän relationsdatabas, medan OLAP-kuben används när datalagret implementeras i en multidimensionell databas (MDB). Även om de implementeras på olika sätt är den underliggande logiken densamma (Kimball & Ross, 2013). Datamodeller för datalager baseras på en s.k. faktatabell (*fact table*). En faktatabell innehåller nyckeltal och kvantitativa data från organisationens normala processer. På ett reningsverk rör detta t.ex. driften, inköp och debitering till VA-abbonenterna. Varje rad med i en faktatabell innehåller information för en given tidsperiod (t.ex. elförbrukning per avskild mängd kväve under föregående dygn). Eftersom reningsverket kontinuerligt genererar sådana värden kan de underliggande ETL-processerna säkerställa att faktatabellen alltid fylls på med de senaste data. Det är därför inte ovanligt att en organisation har faktatabeller med miljontals rader. Data i en faktatabell kopplas ihop med olika s.k. dimensionstabeller (*dimension tables*). Begreppet dimensioner ska i detta sammanhang tolkas som olika perspektiv eller egenskaper för data, och inte som dimensioner i matematisk mening. De olika dimensionerna beskriver kontexten för insamlade data, d.v.s. metadata som förklarar vem, vad, var, när, hur och varför data samlats in i faktatabellerna. Varje dimension innehåller därför en mängd olika attribut som beskriver sammanhanget för data (se Kapitel 3). Attributen i en dimension kan sedan används vid dataanalys för att söka, filtrera eller klassificera faktatabellen och därigenom få en fördjupad och datadriven insikt i reningsverkets verksamhet. På samma sätt som faktatabellerna fylls på med data så uppdateras även dimensionstabellerna genom ETL-processen som styr innehållet i de underliggande databaserna. Men eftersom metadata är ganska statiska och begränsade i volym (dvs. en handfull tabellrader), så uppdateras dimensionstabellerna sällan. Notera att dimensionernas beskrivande karaktär gör att dess kolumner kan vara mycket breda med mycket textinnehåll. Sambandet mellan fakta och dimensioner hanteras slutligen med hjälp av relationer med den unika identifieraren (*primary key*) och tabellrubriken (*foreign key*) (dvs. på samma sätt som datalagring i relationsdatabaser). Metadata i ett datalager begränsas inte av den kontext och sammanhang som data samlats in (t.ex. typ av utrustning som användes för en mätning, operatör som ansvarade för datainsamlingen etc.) utan kan också omfatta tekniska detaljer om den teknik som används för att lagra och hantera data (t.ex. datalagrets struktur, dataflöde, information om algoritmer för sammanfattning). Inom datamodellering skiljer man därför på processmetadata (*business metadata*) och tekniska metadata (*technical metadata*). Faktum är att vissa stora datalager till och med lägger

till metametadata (!) för att förstå och hantera själva metadata (Hay, 2006). I den här rapporten är fokus på processmetadata som beskriver givardata. Två exempel på stjärnmodeller finns i Avsnitt 4.5.5. I exempel E4.1 fungerar tabellen med givardata (blå ruta – *signal*) som faktatabell, medan dimensionstabellerna i exempel 2 (övriga rutor) är länkade till metadatabellen som används direkt av de faktiska värdena (*values*).

Eftersom datalager baseras på tabeller är den typ av data som kan accepteras begränsad. Ursprungligen var det bara mycket strukturerade data (dvs. data som kan struktureras direkt i rader och kolumner) som kunde användas i datalager. Under de senaste decennierna har databastekniken förbättrats avsevärt och semistrukturerade dataformat såsom JSON och XML är numera tillgängliga. Detta har ökat flexibiliteten i moderna datalager eftersom de dimensionella attributen (dvs. kolumnrubrikerna i dimensionstabellerna) inte nödvändigtvis behöver ändras när ytterligare datafält läggs till (Epstein & Roberts, 2022). Trots den ökade flexibiliteten är det viktigt att förstå att de grundläggande dimensionerna i ett datalager skapas utifrån verksamhetens behov av dataanalys. Om vissa data anses irrelevanta under uppbyggnation är det kanske inte möjligt att helt enkelt lägga till dem senare utan att göra en betydande revidering av datalagret. En sådan revidering, när den tillämpas, kallas för *schema-on-write*.

Det sista steget i ett dataflöde är det roliga där användaren analyserar och visualiserar data. Som tidigare nämnts är de flesta datalager baserade på relationsdatabaser och kan därför genomsökas direkt med SQL eller med hjälp av ett allmänt programmeringsspråk (t.ex. C++, Java, Python och R) som kan konstruera SQL-frågor. En utmaning är dock att ledningen troligen inte prioriterar att lära sig att programmera SQL-frågor för att få svar på sina frågor. Detta har lett till att det idag finns ett stort antal intuitiva, flexibla och lättanvända BI-verktyg (*business intelligence*) som kan integreras med datalager utan att användaren behöver skriva en enda kodrad. Med tanke på att ett datalager har strukturerats enligt ett förutbestämt sätt så syftar dessa verktyg vanligtvis till att genomföra beskrivande analys (dvs. vad som hände) för att besvara återkommande verksamhetsfrågor i form av rapporter och instrumentpaneler (*dash-boards*) (omfattande analys av historiska data respektive tolkning av data i realtid). Även om det kan vara möjligt att genomföra specifika analyser i ett datalager är tekniken bäst lämpad för rutinmässiga frågeställningar.

Det finns två huvudskäl till denna begränsning: (1) Att söka efter stora delmängder av data för explorativ analys är mycket tidskrävande i ett datalager, (2) datalager lagrar vanligtvis inte rådata eller högupplösta data. På grund av detta används datalager främst av administrativ personal och inte av dataanalytiker eftersom de arbetar iterativt utan rutinmässiga frågeställningar (Gopalan, 2022).

4.4.7 Data lake

Konceptet med en datasjö (*data lake*) introducerades i början av 2010-talet som en lösning på det växande antalet utmaningar med att administrera och bearbeta data från stora datalager. På den tiden ökade kraven på snabb datainsamling och bearbetning på grund den snabba digitaliseringstakten. Samtidigt började halv- och ostrukturerade data att systematisk analyseras (binära filer eller i NoSQL-databaser såsom ljud, bilder, videor, dokumentsamlingar, inlägg på sociala medier, klickströmmar, loggfiler). Även om datalager var effektiva för att centralisera strukturerade driftdata för beslutsstöd och uppföljning, kunde de inte skalas upp för dessa nya big data tillämpningar. Dessutom saknade datalager stöd för lagring och analys av semistrukturerade/ ostrukturerade data.

Lösningen på denna tekniska utmaning blev att introducera ännu en teknik och teknikskifte som kallades *data lake*. I stället för att förse främst administrativ personal med förbehandlade data utifrån en noggrant utformad datalagringsmodell så lagrar en datasjö rådata (strukturerade och ostrukturerade) i sin ursprungliga form och format. Konsekvensen med det är att olika användare med olika behov kan hitta svar i data själva utan att först behöva förstå en abstrakt datalagringsmodell. Eller som den ursprungliga uppfinnaren av termen beskriver det: “Om du tänker på ett datalager som ett lager med buteljerat vatten – renat, förpackat och strukturerat för att enkelt kunna drickas – så är datasjön en stor vattenförekomst i ett mer naturligt tillstånd. Innehållet i datasjön strömmar in från

en källa för att fylla sjön, och olika användare av sjön kan komma för att undersöka, dyka i eller ta råvattenprover" (Dixon, 2014).

Eftersom data lagras som de är i en datasjö finns det inget behov av att motivera varför vissa datakällor bör lagras. All data kan göras tillgänglig, även när det inte finns något direkt förutbestämt syfte. Detta gör att en datasjön kan användas för mycket mer än rutinmässig rapportering och återkommande dataanalys. Ett bra exempel är explorativ dataanalys, som sker utan ett helt definierat mål. En dataanalytiker eller processingenjör kan få fördjupad kunskap genom att stegvis ställa frågor till datasjön (med hjälp av ett SQL-liknande språk) utan att på förhand veta vilka data som i slutändan kan vara användbara, vilket var svårt att genomföra i ett datalager. Nackdelen är att stora mängder data som automatiskt dumpas på en lagringsplats snabbt kan bli otympliga om de inte hanteras korrekt, vilket fått benämningen dataträsk (*data swamp*). Datasjöar består normalt sett av ett flertal komponenter som strukturera lagring och användning utifrån olika användargrupper.

Tjänster för att integrera olika datakällor är en centrala komponenter i datasjöar. De möjliggör strukturerad inmatning av rådata från olika isolerade källor. I motsats till ETL-processen i ett datalager så används ett något annorlunda tillvägagångssätt i datasjöar. Eftersom datasjön tar emot alla typer av strukturerade eller ostrukturerade data som rådata, finns det inget behov av att modifiera data innan den lagras. Det är istället användaren som modifierar data när det behövs under själva analysfasen. Den största fördelen med ett sådant ELT-förfarande (*extract-load-transform* jämfört med (*extract-transform-load*) är att det blir mer flexibelt, snabbt, och billigare eftersom det inte finns något behov av att omvandla data till en struktur innan den kan användas för analys. Detta kallas även *schema-on-read* i motsats till *schema-on-write*.

Medan datakonsumenten i en datasjö upplever att hen arbetar mot en enda central lagringsplats för data, är dessa data i praktiken utspridda på många underliggande datorer (vanligen i molnet) som arbetar effektivt tillsammans. Logiskt centraliserad, men fysiskt decentraliserad. Genom att använda en sådan arkitektur kan tidsbehovet minskas avsevärt för att bearbeta stora datamängder tack vara s.k. distribuerat och parallella beräkningar arbete (*parallell computation*). Dagens datasjöar baseras därför på billig lagring i molnet (t.ex. AWS Simple Storage Service (S3), Azure Blob Storage, Google Cloud Storage) i kombination med analysmotorer för storskalig databehandling och analys (t.ex. Apache Spark, Upsolver, Presto).

På grund av datasjöns till synes ostrukturerade natur kan den vara svårnavigerad för användarna. Av den anledningen finns en datakatalog (*data catalogue*) som liknar ett register på ett bibliotek. Genom att tillhandahålla intuitiva verktyg för att annotera dataset och attribut med ytterligare metadata kan en katalog över all data som lagras i datasjön skapas så att användare som inte är experter snabbt kan söka efter det de behöver med hjälp av vanliga söktermer. Observera att katalogisering av data kräver en gemensam terminologi och namngivning, och det är här som interna ordlistor, taxonomier, ontologier och namngivningskonventioner blir nödvändiga. När det finns ett gemensamt språk kan katalogiseringen utföras manuellt med hjälp av dataansvariga (*data stewards*) eller automatiskt med hjälp av artificiell intelligens (Gorelik, 2019).

Ett annat sätt att skapa ordning i datasjöns kaos är att bygga upp s.k. lager. Det är skikt (*zones*) där data uppfyller vissa krav. Till exempel: ett bronslager som innehåller all ursprungligen rådata och inte är tillgänglig för de flesta användare; ett silverlager som består av en delmängd av rådata som har verifierats men inte modifierats (primärdata) och är användbart för tekniska dataanalyser; ett guldlager där den mest förfinade data finns som är tillgänglig för ledningen. Genom att dela upp data i olika skikt uppnås inte bara förfinade data utan även en bättre struktur. Det är därför som mellanskikten i datasjöar ofta använder metoder som liknar relationsdatabaser, medan toppskikten innehåller färdigmodifierade dataset som är avnormaliserade (*denormalized*) och läsoptimerade datamodeller (Simon, 2021). Precis som datalager kan datasjöar kopplas till en uppsjö av verktyg för analys och visualisering. På grund av datasjöarnas frånvaro av struktur krävs skräddarsydda verktyg för själva dataanalysen vilket inkluderar verktyg för att utveckla, implementera och nyttja AI-modeller. Ett exempel är verktyg för maskininlärningsoperationer (MLOps) som versionskontrollerar den

Fördelar	Nackdelard
• Driftdata och metadata lagras centralt i råformat, vilket ger obegränsad flexibilitet för framtida användning. • Låg lagrings- och underhållskostnad. • Alla dataformat accepteras, det vill säga strukturerade/semistrukturerade/icke-strukturerade. • "Data collectors" för integration är lättillgängliga enligt standardprotokoll för databaser • Tillgänglig som hanterad tjänst (PaaS) på publika moln med optimerade ekosystemintegrationer för datainhämtning, bearbetning och analys. • Stöder både välplanerad BI och mera ad hoc-dataanalys.	• Centraliserad datalagring kan resultera i monoliter som är alltför komplexa och därför inte är lätta att underhålla eller att hämta data från. • Datakvaliteten upprätthålls inte eftersom det inte finns någon datamodell. • Stöder inte "data transactions" (strukturera om i databasen genom paste, insert, delete osv). • Begränsad hantering av dataexport med flera användare samtidigt ("concurrency control") • Det är svårt att göra precisa datassäkningar pga bristande struktur • Icke-standardiserade datalagringsformat kan behöva manuell handpåläggning.

kod och data som används för att träna datadrivna modeller. I och med utvecklingen av datasjöar förflyttades fokus från att enbart beskriva (dvs. vad hände) och diagnostisera (dvs. varför hände det) till att prediktera (dvs. vad kan hända) och i slutändan förebyggande åtgärder (dvs. påverka vad som händer) (*prescriptive analysis on big data*).

4.4.8 Branschexempel

Detta avsnitt avslutas med två exempel som illustrerar dataflöden och hur rådata och metadata som samlas in i driftdatabaser omvandlas till beslutsunderlag upp igenom beslutshierarkin.

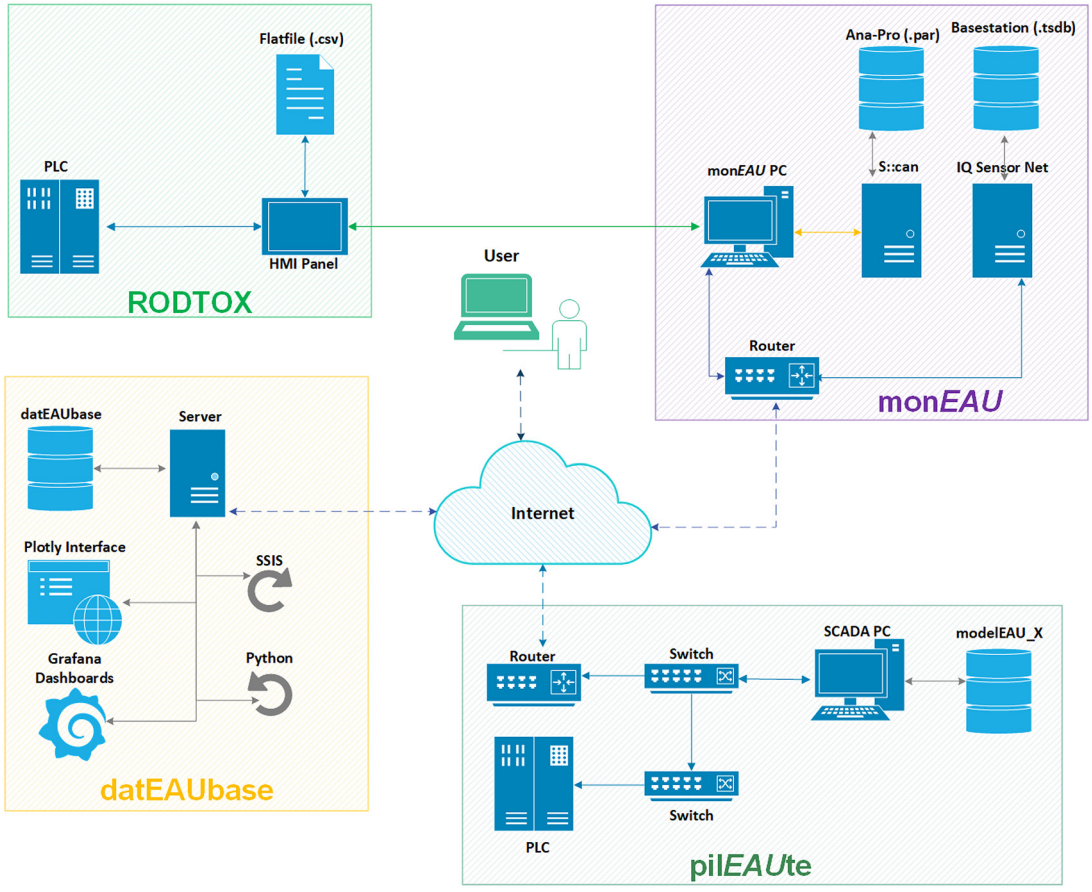
Som Avsnitt 4.3 och Avsnitt 4.4.6 beskrev livscykeln för data, respektive datalager, så flödar data genom olika faser i sin livscykel. Flödet från en fas till en annan kan liknas vid att följa dataflödet genom ett rör (*pipeline*). Therrien *et al.* (2020) presenterar de olika konceptuella faserna för ett sådant dataflöde (*data pipeline*) för reningsverk. Notera att även om visionen är en helt automatiserat dataflöde så krävs vanligtvis även viss manuell handpåläggning.

4.4.8.1 Exempel 1: pilEAUte pilotanläggning för avloppsvattenrening

pilEAUte är en välinstrumenterad pilotanläggning för avskiljning av BOD och kväve. Pilotanläggningen drivs av *modelEAU*, Université Laval, QC i Kanada och består av två parallella linjer för biologisk rening så att olika reglerstrategier ska kunna utvärderas. Pilotanläggningen utformades som en konventionell nitrifikations/fördenitrifikationsanläggning men är även förberedd för biologisk fosforrening och biofilmsprocesser med bärare.

Anläggningen har flera datakällor som stöd driftsstöd (Figur 4.7). Mätvärden från givare, ställdon, börvärden etc. genereras från automationssystemet (*PLC/SCADA*). Dessa signaler lagras som tidsserier i en relationsdatabas genom signalens TAG-namn (*modelEAU_X*), men utan någon kontextuell information. Det finns även ett fristående system som används för detaljerad karakterisering av inflödets och utflödets sammansättning (*monEAU* online-sensorer och *RODTOX* at-line batch respirometri). Dessa system är dock inte integrerade i det centrala automationssystemet. Data från dessa system lagras istället på lokala datorer i databaser i fildatabaser (filformaten.par och.csv) och en särskild tidsseriedatabas (filformatet.tsdb). Dessutom innehåller driftloggböcker noteringar om givarunderhåll (t.ex. kalibrering, validering, rengöring, anteckningar) och driftobservationer som fylls i av driftpersonalen. De senare lagras som fildatabaser i Excel-format.

Datakällorna i pilEAUte-anläggningen är isolerade från varandra, vilket försvårar en övergripande dataanalys. För att lösa detta integrerades de flesta datakällorna via säkra internettunnlar *(virtual private network – VPN*) till ett centralt datalager som fungerar som en historikdatabas (*datEAUbase*). Integrationen extraherar automatiskt nya data med jämna mellanrum, omvandlar dem till förväntat

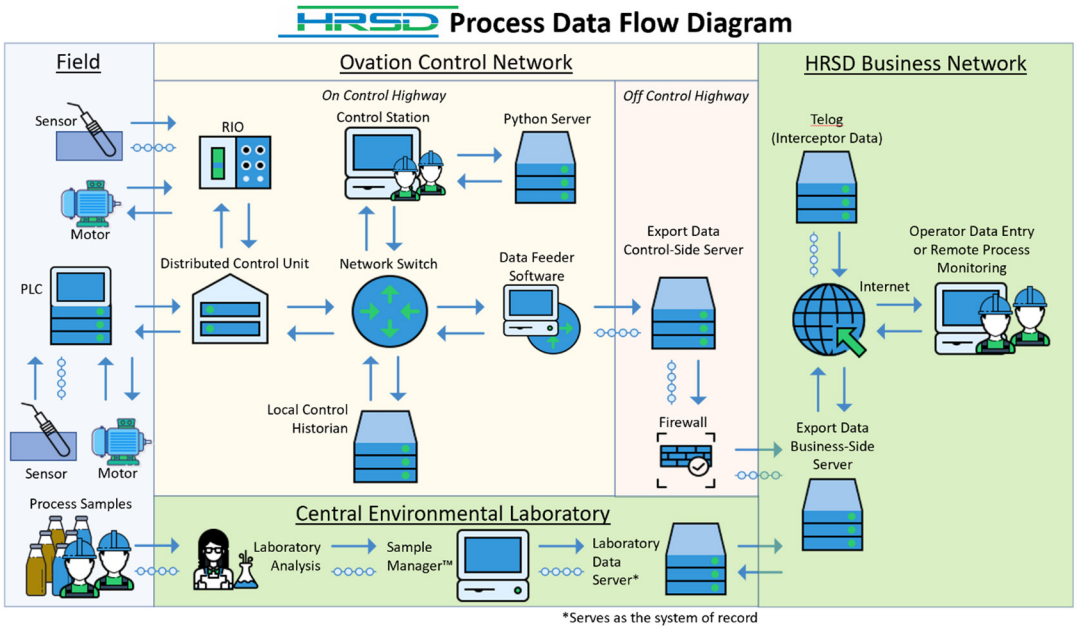


Figur 4.7 Dataflödesdiagram för pilEAUte-pilotprojektet för rening av avloppsvatten.

format (t.ex. SI-tekniska enheter) och lagrar dem i önskad databas (*ETL-process*) genom olika integrationstjänster (*MS SQL Server Integration Solutions - SSIS* och skräddarsydda python-skript). Användarvänliga gränssnitt har också utvecklats som stöd för operatören så att hen intuitivt kan visualisera givardata och metadata (Plotly-gränssnitt och dashboards från Grafana). Det centrala datalagret baseras på en relationsdatabas (*MS SQL Server*) med en skräddarsydd datamodell (Figur 4.2). Observera att datamodellen i stor utsträckning använder beskrivande metadata (relaterade till instrumentering, platser, kontaktpersoner etc.) för att strukturera all data. I det avseendet har det stora likheter med ett datalager för kommersiell affärsverksamhet. Slutresultatet är att operatörerna har tillgång till all information centralt för att kunna ta både snabba och långsiktiga beslut.

4.4.8.2 Exempel 2: förenklat flödesschema för processdata på Hampton Roads Sanitation District (HRSD) anläggningar

Hampton Roads Sanitation District (HRSD) är en VA-organisation som driver flera stora reningsverk i sydöstra Virginia, USA. Dataflödet, datastrukturen och datahanteringen är standardiserad i alla stora anläggningar enligt Figur 4.8. Data produceras och hanteras för att uppfylla lagkrav och för driftövervakning. De data som är relevanta för driften produceras huvudsakligen av fyra datakällor:



Figur 4.8 Dataflöde i HRSD:s reningsverk.

(1) Driftpersonal (t.ex. manuell mätning av slamtäcke).
(2) Givare och instrument (Electrical and Instrumentation equipment – E&I).
(3) Automations- och övervakningssystem (t.ex. styrsignaler och börvärden).
(4) Laboratorieanalyser (t.ex. stickprov av ammoniumhalten i utgående vatten).

Data lagras därefter på sex olika platser beroende på syfte. Notera att HRSD:s nätverk är uppdelat mellan en säker sida för styrsystemet och en del för allmän användning. Enligt IT-säkerhetspolicyn så tillämpas enkelriktad överföring, d.v.s. får enbart överföras från styrsystemssidan till det allmänna nätverket. Detta för att skydda verksamheten från eventuella cyberattacker och nätverksavbrott. Detta innebär att ändringar i driftskritiska komponenter endast kan göras inifrån styrsystemsnätverket. Denna policy var ovärderlig när HRSD utsattes för en ransomware-attack 2020 som stängde ned det allmänna nätverket, men påverkade inte reningsverkets drift. Datainfrastrukturen uppfyller därmed HRSD:s säkerhetsbehov, efterlevnad av tillstånd samt behov för driftövervakning och analys för både driftpersonal och företagsledning.

Styrsystemsnätverk

- Lokal historikdatabas - kortvarig lagring av processdata för kritiska driftparametrar (lagras upp till 30 dagar).
- *Export data control-side server* (EDS-CS) - klon av drifttaggar och trender. Inkluderar all givar- och driftdata.
- Python-server - på två av HRSD:s anläggningar används pythonskript för digitala tvillingar och maskininlärning i forskningssyfte. För dessa tillämpningar kommuniceras taggar via OPC-protokoll.

Allmänt nätverk

- *Export data business-side server* (EDS-BS) - klon av drifttaggar och trender. Inkluderar all givar- och driftdata, vissa laboratoriedata och data som producerats av driftpersonalen. Gränssnittet möjliggör processövervakning via tidsserietrender och dataexport.

- Server för laboratoriedata – alla data som genereras av laboratorieta och som är relevanta utsläppskontroll och som genereras av DCS eller anläggningsoperatörer. Servern fungerar som registreringssystem emissionsdata och för att generera av månatliga driftrapporter (kalkylblad för driftsuppföljning.
- Telog (interceptor) – data som genereras av pumpstationer, tidvatten eller nederbörd.

All data lagras som tidsserier. Data kommuniceras kontinuerligt från automationssystemet till den lokala historikdatabasen via ett nätverk med TCP/IP-protokoll eller till EDS-CS innan de skickas vidare till EDS-BS. Tidsupplösningen varierar beroende på signal från sekunder (givare och automationssystem) till dagar (manuella driftanteckningar). För att minimera onödig datalagring av högupplöst data så modifieras signalerna så att ett nytt värde endast lagras efter att en viss värdeförändring har observerats (dvs. Box-Car-komprimering, Avsnitt 4.4.5). Tröskelvärdet sätts som standard till 1 % av mätintervallet för analoga signaler. Till exempel så kommer en syrehaltsmätning i intervallet 0 till 10 mg/L endast att lagras om mätningen har ändrats med minst 0,1 mg/L jämfört med föregående registrering. För digitala binära signaler så registreras endast tillståndsförändringen. Både historikdatabasen och EDS-servrarna baseras på relationsdatabaser. Ytterligare metadata kan lagras per datapunkt (t.ex. signalkvalitet, punktintervall, enhet, beskrivning, larmprioritet och larmgränser samt datapunktkälla) i EDS-CS och EDS-BS.

Data som samlas in av driftpersonalen och korta textkommentarer (d.v.s. anteckningar) har en tidsstämpel med timupplösning, och dessa data skickas in via en webbapplikation som körs i webbläsaren och som har åtkomst till den underliggande EDS-BS-databasen. Tidigt på morgonen komprimeras de föregående dygnens data som krävs för tillståndefterlevnad och driftrapporter genereras och lagras i EDS-BS som en fildatabas (.CSV) och överförs till laboratoriets dataserver, som i sin tur nyttjar en relationsdatabas. De data som genereras av laboratorieanalyser eller beräknas som en kombination av laboratoriedata och data från EDS-BS (t.ex. retentionstid för aeroba fasta ämnen) kan också föras tillbaka till EDS-BS. Data för kvalitetskontroll (QA/QC) såsom standarder, kalibreringsdatum, och interferenskontroller lagras också i laboratoriets dataserver för att uppfylla kraven för laboratorieackreditering. Dataöverföringen mellan EDS-BS och laboratoriets dataserver sker med hjälp av en fildatabas via en säker webbplats.

HRSD:s allmänna nätverk kan nås externt via ett virtuellt privat nätverk (VPN) för fjärrövervakning eller för att exportera laboratoriegenererade processdata. En webbapplikation används för att registrera data som genereras av anläggningsoperatörerna och kan nås av alla med en internetanslutning och inloggningsuppgifter.

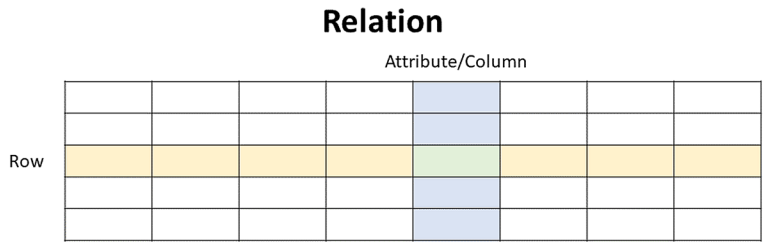
Data som rör det digitala underhållshanteringssystemet (CMMS), biologiska aktivitetstester och givarunderhåll lagras inte i EDS-BS.

4.5 FÖRSTÅELSE FÖR RELATIONSDATABASER

Relationsdatabaser är väldigt populära för att lagra de flesta typer av data som produceras på företag på ett strukturerat sätt i mjukvaruapplikationer. Som tidigare nämnts kan relationsdatabaser även användas för att lagra tidsseriedata och metadata från givare. Även historikdatabasers lagring och export av tidsserier bygger ofta på tekniker med stora likheter med relationsdatabaser. Eftersom relationsdatabaser är så vanliga diskuteras strukturerna och de grundläggande begreppen i detalj i kommande avsnitt, speciellt utifrån metadata och beskrivning av datakvalitet.

4.5.1 Definition

En relationsdatabas baseras på en relationsmodell för data som är en databas vars logiska struktur består av en samling relationer. En relation är i detta sammanhang en tabell med kolumner (s.k. attribut eller fält) och rader (*tuples* eller dataposter) (se Figur 4.9).



Figur 4.9 En relation (eller tabell) i en relationsdatabas.

En kolumn är den minsta organisationsstrukturen i en relationsdatabas och representerar de olika aspekter som definierar dataposterna i en tabell. Namnet på en kolumn i en tabell är unikt, men det kan vara detsamma i andra tabeller i samma relationsdatabas. Alla värden i en kolumnen behöver vara av samma datatyp, men kolumnerna kan visas i vilken ordning som helst utan att det påverkar tolkningen av data.

När det gäller raderna är varje datapost ett unikt värde (*instance*) av en viss typ av data (t.ex. signalvärde, sensortyp, analysmetod, metod för databearbetning etc.). Med andra ord finns det inga duplicerade rader i en relation. Varje rad identifieras med ett unikt värde, en s.k. primärnyckel, som innehåller en eller flera kolumner. Eftersom primärnyckeln är unik, blir även varje rad unik, ett koncept som kallas entitetsintegritet (*entity integrity*). I Tabell 4.7 så motsvarar varje rad ett värde som uppmätts vid en viss tidpunkt. Varje rad ges automatiskt en nyckel som identifiering, vilket gör den unik.

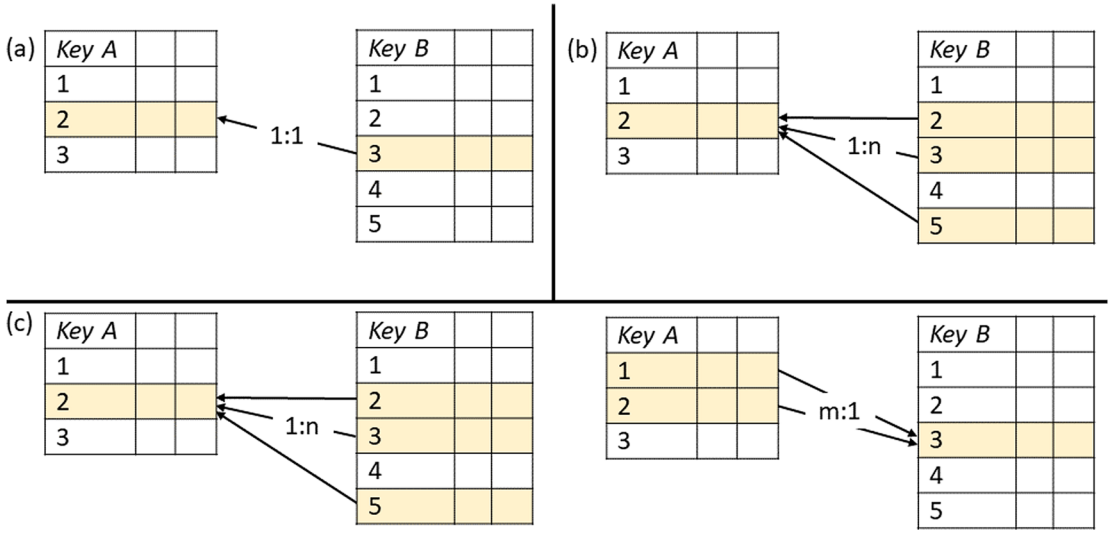
Den allmänna strukturen i en relationsdatabas innehåller två typer av tabeller: primära tabeller och uppslagstabeller (korsreferenstabell). Primära tabeller är de centrala tabellerna i en databas och de innehåller flest relationer till andra tabeller. En primär tabell innehåller vanligtvis en primärnyckel som är ett heltal som justeras automatiskt (för att säkerställa att den är unik) och minst ett textfält som innehåller eller grupperar den viktigaste informationen i databasen.

Uppslagstabeller fungerar som referenstabeller som huvudsakligen är uppbyggda med en primärnyckel och textfält som innehåller kategorier, ett kategorinamn och en beskrivning av kategorin. Den här typen av tabeller har i allmänhet en enda relation till den primära tabellen eller till en annan uppslagstabell. När relationen har upprättats blir primärnyckeln i en uppslagstabell en främmande nyckel i den primära tabellen eller i den andra uppslagstabellen.

Dessutom kan länken mellan primära eller främmande nycklar från olika tabeller vara en-till-en, en-till-många eller många-till-många (Figur 4.10). Ett exempel är en tabell som beskriver det specifika instrumentet (rader) med de viktigaste egenskaperna (kolumner), och en tabell som beskriver de analyserade vattenproverna (rader) och tillhörande resultat (kolumner). Ett vanligt ett-till-ett-förhållande uppstår när ett enstaka laboratorieanalysresultat erhålls från ett enda vattenprov. På samma sätt kan ett ett-till-många-förhållande vara att en koncentration i ett vattenprov mäts flera gånger med samma

Tabell 4.7 Exempel på relation.

Primärnyckel	Tidsstämpel	Värde	Parameter	Mätenhet	Sensor	Plats
1	03:45:10	7,123	pH-värde	–	PH18065	Utlopp försedimentering
2	03:45:15	7,324	pH-värde	–	PH18065	Utlopp försedimentering
3	03:45:20	7,163	pH-värde	–	PH18065	Utlopp försedimentering



Figur 4.10 Olika typer av relationer i en relationsdatabas: (a) en-till-en, (b) en-till-många och (c) många-till-många relationer.

instrument. När det gäller förhållandet många till många är ett exempel när ett instrument kan mäta flera parametrar, men att samma parameter också kan mätas med olika instrument.

4.5.2 De vanligaste datarelationerna

Som tidigare beskrivs i Kapitel 3 och Avsnitt 4.2 bör signaler som används för att processövervakning och reglering dokumenteras, det vill säga sättas in i sitt sammanhang, för att göra dem användbara och lättillgängliga för vidare analys och tolkning för andra ändamål. På ett reningsverk innebär denna dokumentation att man tillhandahåller information om när, var och hur en datapunkt samlades in, vem som samlade in den, av vilken anledning den samlades in och vilken kvalitet datapunkten har. Nedan beskrivs de huvudsakliga relationer som används för att knyta metadata till signaldata.

4.5.2.1 [datapunkt] producerad av [enhet]

Detta är en av de tydligaste datarelationerna på ett reningsverk. De insamlade datapunkterna kopplas samman med den enhet som producerade dem. Detta gör det möjligt att identifiera och tillhandahålla information om hur en datapunkt erhölls. I Figur 4.11 ges ett förenklat exempel på vilken typ av information som kan lagras för en enhet och för datapunkter i tabeller, och hur de båda tabellerna kan länkas.

4.5.2.2 [datapunkt] producerad vid [plats]

En annan viktig relation är att beskriva var datapunkterna producerades. Denna information ger mer bakgrundsinformation om den exakta fysiska platsen och därmed miljön där datapunkterna har samlats in. Figur 4.12 visar ett enkelt exempel på en typ av platstabell och möjlig information för att beskriva den specifika plats där en datapunkt samlas in.

4.5.2.3 [kvalitet] baserat på [datapunkter]

Det är även relevant att koppla information datakvaliteten till olika datapunkter. Detta kan vara all information som rör de tekniker eller metoder som används för att kvalitetskontrollera det uppmätta värdet. Detta sista förhållande beskrivs mer i detalj i Avsnitt 4.5.3.

Device table

Primary Key	*Device*	*Identifier*	*Manufacturer*	*Model*	*Year*
1					
2	pH sensor	PH18065	Company	pHmeter_v 3	2018
3					

1:n

Data point table

Primary Key	*Timestamp*	*Value*	*Foreign Key Device*
1			
2	03:45:10	7.123	2
3	03:45:15	7.324	2
4			
5	03:45:20	7.163	2

Figur 4.11 Schema över hur en datapunkt kan kopplas till information om den enhet som har producerat den.

4.5.3 Relationer som beskriver datakvalitet

Som tidigare nämnts (Kapitel 2 och 3 samt Avsnitt 4.5.2) så ökar tillförlitligheten, reproducerbarheten och robustheten i data om de kompletteras med beskrivningar om sammanhang som rör dataförändringar och kvalitetskontroller. Detta kommer att påverka dataanalytikerns förtroende för datasetet. Information om datakvalitet kan tillhandahållas direkt av de enheter som producerar datapunkterna eller av slutanvändaren själv. Kvalitetskontroller kan ingå i samma tabell som datapunktstabellen (se Figur 4.13) eller separeras i en oberoende tabell om mer omfattande beskrivningar behövs (dvs. flera fält och kombinationer) (se Figur 4.14). Båda exemplen, Figur 4.13 och 4.14, visar hur det går att annotera rådata och modifierade data efter att ett glidanande medelvärdesfilter har använts för brusreduktion. Tekniska detaljer som rör datamodifiering och/eller valideringsmetoder som tillämpas på en datapunkt (t.ex. datapunktens version eller helt enkelt en kvalitetsflagga) kan ingå i

Location table

Primary Key	*Site name*	*Sampling location*	*Latitude*	*Longitude*	*Address*
1					
2					
3	Primary settler 3	Outlet channel	42.460367	1.367154	2324 Water St, Water City

1:n

Data point table

Primary Key	*Timestamp*	*Value*	*Foreign Key Location*
1			
2	03:45:10	7.123	3
3	03:45:15	7.324	3
4			
5	03:45:20	7.163	3

Figur 4.12 Exempel på information som ingår i tabeller över plats och datapunkter.

Analytical methods table

Primary Key	*Method Name*	*Method type*	*Code version*	*Description*
1	Raw data	Measurement	N/A	N/A
2	Moving average	Univariate	3.2.1	Specific details of the code version
3	Principal component analysis	Multivariate	0.1.9	Specific details of the code version

1:n

Data point table

Primary Key	*Timestamp*	*Value*	*Foreign Key*
1	03:45:10	7.123	1
2	03:45:15	7.324	1
3	03:45:10	7.12	2
4	03:45:15	7.18	2
5	03:45:20	7.158	1
6	03:45:20	7.16	2

Figur 4.13 Exempel på tabell över analysmetoder och taggad till datapunktstabell.

Analytical methods table

Primary Key	*Method Name*	*Method type*	*Code version*	*Description*
1	Raw data	Measurement	N/A	N/A
2	Moving average	Univariate	3.2.1	Specific details of the code version
3	Principal component analysis	Multivariate	0.1.9	Specific details of the code version

1:n

1:n

Raw data point table
(link to the key 1 in analytical methods)

Primary Key	*Timestamp*	*Value*	*Foreign Key*
1	03:45:10	7.123	1
2	03:45:15	7.324	1
3	03:45:20	7.158	1

Data point table after moving average
(link to the key 2 in analytical methods)

Primary Key	*Timestamp*	*Value*	*Foreign Key*
1	03:45:10	7.12	2
2	03:45:15	7.18	2
3	03:45:20	7.16	2

Figur 4.14 Exempel på tabell över analysmetoder och separata tabeller beroende på dataversion och analysmetod.

relationsstrukturen. Genom att hålla reda på ursprunget för data är det lättare att förstå och hantera effekterna av vissa omvandlingar efter det att de har tillämpats.

Definition D4.10: **Dataproveniens** (*data lineage*) avser metadata som beskriver de processer som påverkat den aktuella datamängden och därigenom ger en historisk bild av data och deras ursprung. Detta inkluderar ursprung, modifiering och lagring. Detta leder till en större transparens som även underlättar felsökning vid dataanalysen.

Ett exempel kan vara att ett bias upptäcks och korrigeras automatiskt i en pH-signal. Databasen lagrar dock bara de korrigerade värdena, vilket försvårar efterföljande analyser. Om det istället finns information om korrigeringsmetoden är det möjligt att rekonstruera den ursprungliga och felaktiga signalen för att studera effekterna av de modifieringar som gjorts efter det att rådata samlades in.

Den information som ska ingå i en relationsdatabas kan förenklas genom att tillämpa en hierarkisk struktur. All information som rör datakvalitet kan exempelvis lagras som en uppsättning tabeller i en trädstruktur.

4.5.4 Definiera en relationsstruktur

För att bygga upp en relationsdatabas behöver man ta hänsyn till att rådata kan ha ett mycket varierande format; databasen måste vara flexibel och anpassningsbar för att kunna utvidgas kontinuerligt; och personalen som samlar in och hanterar data varierar också över tiden (Camhy *et al.*, 2012; Horsburgh *et al.*, 2008). En relationsdatabas bör därför:

- Bevara korrekta data (*accurate*) genom att gruppera flera samtidiga databasoperationer när databasen uppdateras.
- Säkerställa felfri lagring (*consistency*) och att data hanteras på förväntat sätt utifrån databasens regelverk.
- Separata datatransaktioner (*isolation*) som genomförs oberoende av andra transaktioner.
- Vara beständig (*durability*) så att alla slutförda databasoperationer är permanenta, trots avbrott och systemkrascher.

Enligt Harrington (2016) finns det flera steg för att definiera en databas struktur, vilken listas nedan utifrån en tänkt reningsverkstillämpning:

(1) Definiera syftet med en relationsdatabas för vattenkvalitetsdata.
(2) Lista de data som ska lagras (dvs. information som presenteras i Kapitel 3). För ett instrument kan detta t.ex. innefatta givarens namn, tillverkare, tillverkningsår, mätprincip och så vidare.
(3) Kartlägg länkarna mellan de data som ska lagras. En givare kan t.ex. vara kopplad till en plats och en parameter, eller ett värde kan vara kopplat till en givare, en parameter, en plats och en analysmetod.
(4) Gör en genomförbarhetsstudie för att bedöma om en ny databas behöver implementeras, alternativt modifiera en befintlig databas.
(5) Utveckla databasen utifrån sambanden som kartlades i steg 3.
(6) Utveckla och testa databasen.
(7) Implementera databasen och koppla upp SCADA-systemet till den.
(8) Utvärdera databasen.

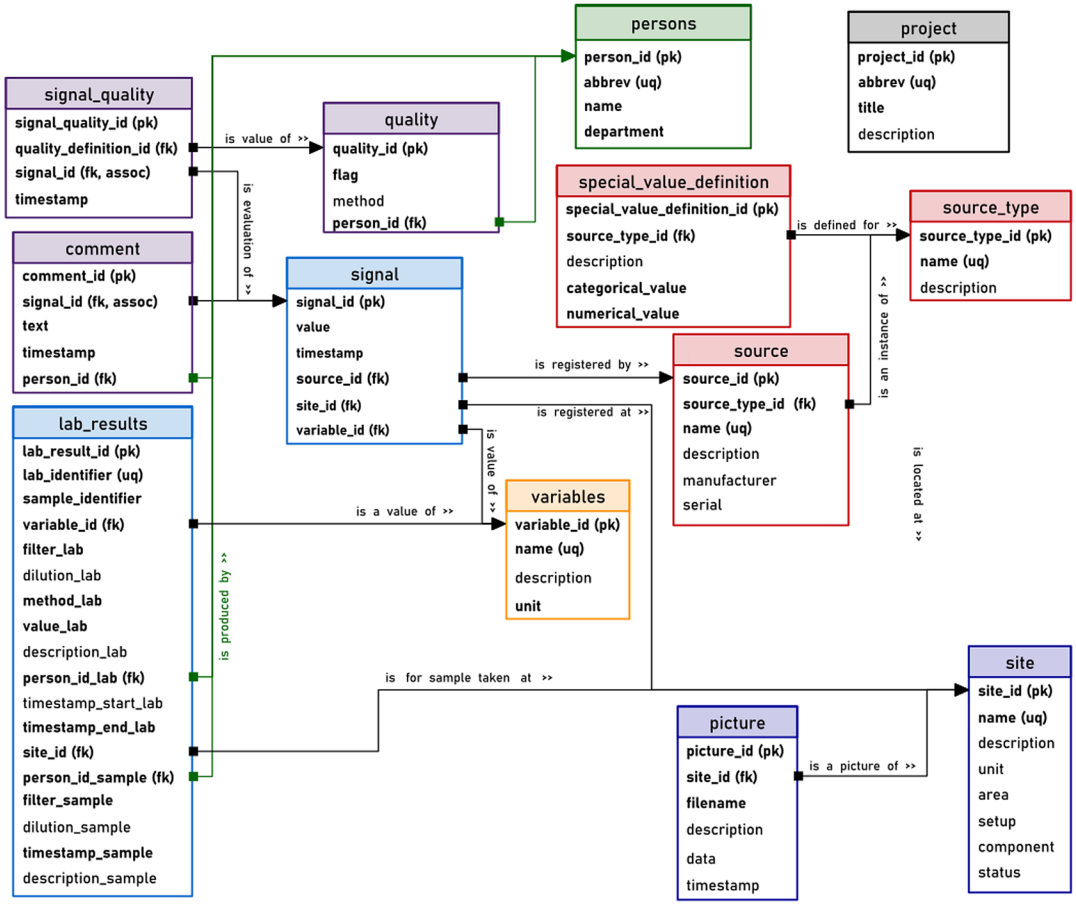
Olika databaser för tillämpningar inom hydrologi och vattenkvalitet har publicerats. Till exempel har Consortium of Universities for the Advancement of Hydrologic Science, Inc (CUAHSI) presenterat observationsdatamodellen (ODM) för hydrologisk datalagring (Horsburgh *et al.*, 2008) och Water Quality Portal av US Environmental Protection Agency (EPA) för vattenkvalitetsdata (EPA, 2023). Informationslagringen har dock varit begränsad till information om platsen eller enheten för mätningen snarare än om datakvaliteten. Nyligen har nya relationsdatabaser för vattenkvalitet presenterats med tonvikt på metadata som främjar information om datakvalitet, till exempel WaterHubDataPool

(WHDP) från Eawag (ETH Zurich, 2019) och datEAUbase från forskargruppen *ModelEAU* från Université Laval (Plana *et al.*, 2019).

4.5.5 Två branschexempel

4.5.5.1 Exempel E4.1: WaterHubDataPool

WaterHubDataPool (WHDP) är en databas som är utformad för att samla in och lagra data som produceras inom WaterHub på Eawag. WaterHub är en del av försöksanläggning NEST (ett s.k. living lab) där insamling, rening och fördelning av vattenresurser studeras. WaterHub hanterar olika avloppsvattenströmmar inklusive urin, ljust gråvatten (t.ex. regn), mörkt gråvatten (t.ex. badvatten) och svartvatten (toalettspolvatten). WaterHub har en mängd olika reningstekniker för att behandla dessa vattenströmmar och återvinna användbart vatten eller andra resurser, som dessutom förändras över tid. WHDP utformades för att samla in och lagra data från dessa reningsprocesser på heltäckande, framtidssäkert och lättanvändbart sätt. I följande avsnitt beskrivs den datamodell som används i relationsdatabasen. Datamodellen illustreras i Figur 4.15 och ytterligare öppen information finns tillgänglig på ETH Zurich (2019).



Figur 4.15 Exempel E4.1: Layout av en relationsdatabas för WaterHubDataPool (WHDP) vid Eawag.

Databasens layout innehåller komponenter för högfrekventa signaler (t.ex. ställdon och givare) samt lågfrekventa signaler (t.ex. laboratorieanalyser). I detta exempel diskuteras de viktigaste tabellerna som är relevanta för lagring av högfrekventa signaler och motsvarande metadata. Varje tabell innehåller en primärnyckel (*pk*) som tilldelas automatiskt och som vanligtvis inte är synlig för dataanvändaren. Dessa diskuteras inte närmare.

Mätvärdena lagras i signaltabellen, som är centralt placerad i databasens layout. Denna tabell innehåller alla dataposter för ställdon och givare. För att en datapost ska bli komplett måste man känna till datapunkten (lagrad som *värde*), *tidsstämpeln*, den dataproducerande enheten (lagrad som *source_id*), platsen för den uppmätta variabeln (lagrad som *site_id*) och den *variabel* som mäts (lagrad som *variable_id*). Posterna för *source_id, site_id och variable_id* är främmande nycklar (*foreign keys*, *fk*) och används för att peka på poster i andra tabeller, nämligen tabellerna *source*, *site* och *variables*. Dessa tabeller används främst för att beskriva datagenereringsprocessen (se Kapitel 3, Avsnitt 3.2.2):

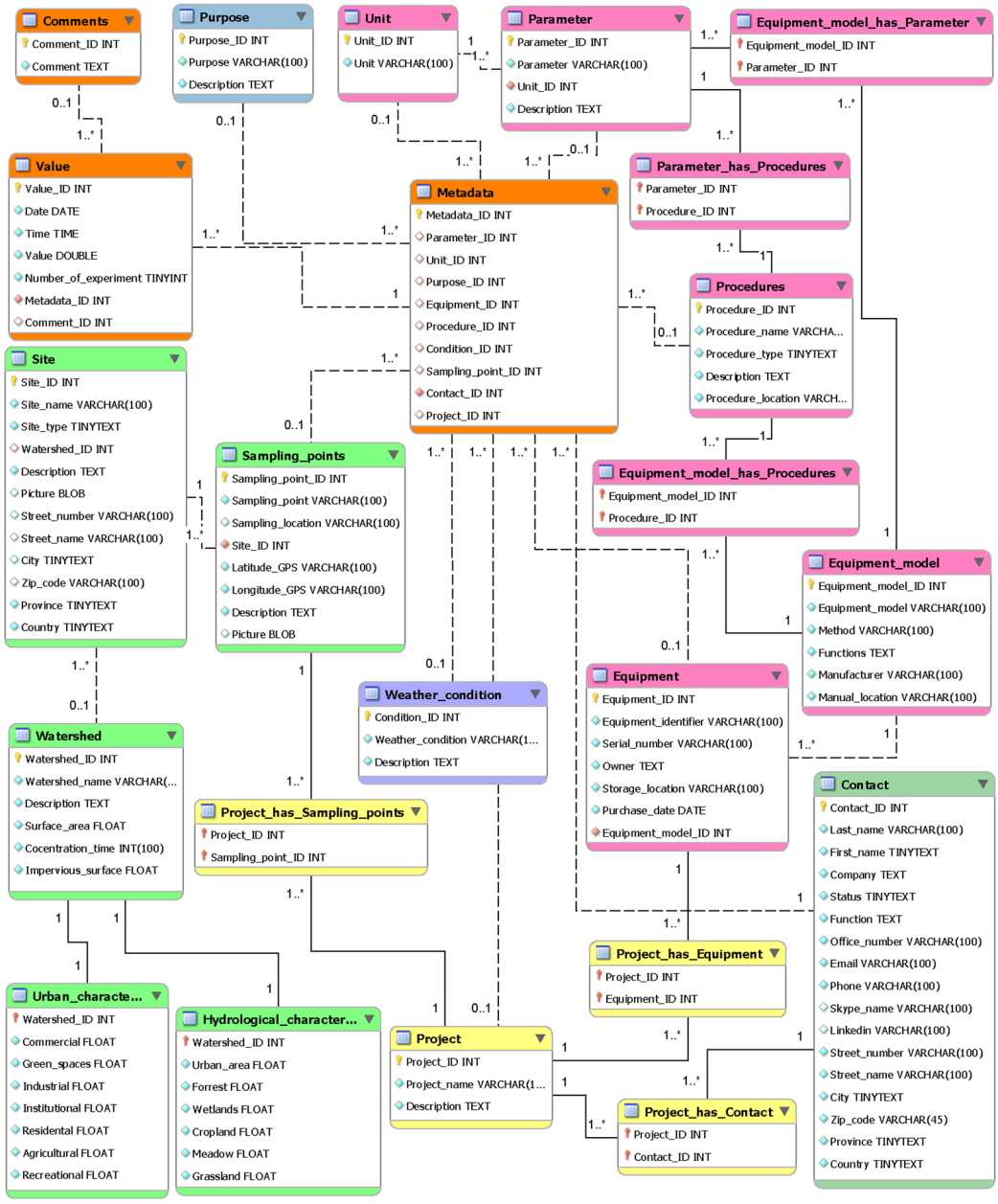
- Tabellens *källa("source")*: Varje post i denna tabell beskriver ett ställdon eller en givare. Ett ställdon eller en givare beskrivs med enhetstypen (lagrad som *source_type_id*), ett unikt *namn*, en *fritextbeskrivning*, *tillverkaren* och ett serienummer (*serial*). *Source_type_id* pekar på tabellen *source_type*, som listar och beskriver alla kända enhetstyper med ett *namn* och en *beskrivning*. Med hjälp av denna struktur behöver information som delas mellan enheter av samma typ bara anges en gång. Det kan t.ex. finnas flera pH-elektroder i bruk som delar samma information om fabrikat och konstruktion.
- Tabell *plats (site):* varje post i denna tabell beskriver en mätplats med ett unikt *namn*, en *beskrivning* av platsen, ett rum (lagrat som *enhet*), en yta i rummet (lagrat som *yta*), en experimentell uppställning (lagrat som uppställning), en *komponent* i uppställningen och en indikator på platsens status (t.ex. planerad, installerad, aktiv, inaktiv).
- Tabell över *variabler*: Varje post i denna tabell beskriver en variabel med ett unikt *namn*, en unik *beskrivning* och en unik *enhet*.

Tabellen *picture* används för att länka filnamn på bilder som lagras någon annanstans till specifika poster i tabellen *site*. Tabellen *special_value_definition* används för att dokumentera värden för en datapunkt som har en unik innebörd för en viss enhetstyp. Denna typ av information används för att automatiskt härleda och dokumentera saknade data och data utanför intervallet (se Kapitel 2, Avsnitt 2.4.1.3 och Kapitel 3 Avsnitt 3.2.2.6–3.2.2.7).

WHDP-layouten skapar ett flexibelt sätt att lagra metadata som beskriver signalkvalitet och kontextuell information i kommentartabellen. Detta gör det möjligt att lagra fritext för att kommentera data. En enskild kommentar beskrivs som den unika dataposten (lagrad som *signal_id*), en fritextkommentar (lagrad som *text*), tidpunkten då kommentaren skrevs (lagrad som *tidsstämpel*) och den *person som lämnade kommentaren (lagrad som person_id*). Värdena för *signal_id* pekar på posten i signaltabellen. På samma sätt pekar *person_id* på tabellen *persons*, som innehåller information om de mänskliga experter som finns tillgängliga för att kommentera data.

WHDP-layouten ger också ett strukturerat sätt att lagra kontexten som metadata. Det är främst avsett för lagring av metadata som beskriver signalkvalitet (se Kapitel 3 Avsnitt 3.3–3.4 för relevanta metadata). I tabellen "*Quality*" listas därför alla annoteringar/etiketter som finns tillgängliga för att beskriva kända avvikelse- och feltyper. Varje etikett beskrivs med ett namn (*t*.ex. *outlier*, lagrad som *flagga*), ett namn på den metod som används för att skapa etiketten (t.ex. *EWMA*, *visuell inspektion*, lagrad som *metod*) och den *person* som har tillhandahållit etiketten (lagrad som *person_id*). En särskild tabell för *signalkvalitet* används för att länka samman *signal-* och kvalitetstabellerna. Detta gör det möjligt att tilldela en enda *kvalitetsetikett* till flera dataposter i signaltabellen samtidigt som flera *kvalitetsetiketter* kan kopplas till samma datapost i signaltabellen.

De återstående tabellerna (*lab_results*, *project*) är inte relevanta för att beskriva signaldata och beskrivs därför inte här.



Figur 4.16 Exempel E4.2: Schema för datEAU-basen av modelEAU (Plana *et al.*, 2019).

4.5.5.2 Exempel E4.2: dateEAUbase

DatEAU-databasen utvecklades för att lagra alla data som samlats in som en del av forskningsprojekt av forskningsgruppen *ModelEAU* vid Université Laval, Québec, Kanada. Huvudsyftet med databasen är att centralisera och lagra data från onlinemätningar och laboratorieanalysresultat från

mätkampanjer som genomförts i olika vattenförekomster (t.ex. floder, avloppssystem, reningsverk) (Plana *et al.*, 2019). Databasen utformades på ett sådant sätt att den intuitivt lagrar relevanta metadata som behövs för att dokumentera och spåra de lagrade uppgifterna under hela deras livscykel (se även Figur 4.2 och Avsnitt 4.4.8.1).

Databasstrukturen som visas i Figur 4.16 kan delas in i två huvuddelar:

(1) Huvudtabellerna består av tabellerna *Values*, *Metadata* och *Comment*. Tabellen *Values* kan betraktas som en faktatabell i ett datalager, eftersom den innehåller onlinemätningar, resultat av laboratorieanalyser, tidsstämplar och motsvarande främmande nyckel till den beskrivande tabellen *Metadata*. *Metadatatabellen* grupperar de främmande nyckelnumren för resterande datadimensioner och fungerar därmed som ett mellanled som lagrar alla befintliga dimensionskonfigurationer. Anteckningar som är specifika för ett eller flera värden kan också läggas till (t.ex. sensorfel, rengöringsperiod, avvikelse etc.) med hjälp av tabellen *Comment*.

(2) Ytterligare datadimensioner som innehåller beskrivande metadata för mätningen. Detta inkluderar information om den exakta platsen där mätningarna utfördes, den enhet eller metod som användes för mätningen, syftet med mätningen, den person som ansvarade för mätningen, väderförhållandena vid mätningstillfället och det projekt som är kopplat till mätningen.

De mest relevanta metadatatabellerna beskrivs nedan:

- Platsinformation: Huvudtabellen för platsinformation är tabellen *Sampling_points* som gör det möjligt att ange den exakta platsen för mätningen inklusive GPS-koordinaterna (t.ex. inloppet till försedimentering nr 2; 46°41004″N 72°42059″W). Tabellen *Sampling_points* är kopplad till tabellen *Site* som innehåller mer omfattande information om platsen, t.ex. adress, bild och beskrivning. I anslutning till denna tabell finns också ytterligare information om vattenförekomsten, urbana och hydrologiska egenskaper som är länkade till tabellen *Site*.
- Information om enhet eller metod: denna del av databasen består av flera tabeller, t.ex. tabellerna *Equipment*, *Equipment_model*, *Procedure*, *Parameter* och *Unit*. Dessa tabeller innehåller information om de enheter som används för onlinemätningar eller utrustningen för laboratorieanalys, den procedur som används för att erhålla ett värde samt parametern och måttenheterna.

4.6 SLUTSATSER

- Det finns ett stort antal olika databaser och datahanteringstekniker som gör det möjligt att lagra signaldata tillsammans med beskrivande metadata.
- Relationsdatabaser är den vanligaste databastekniken som kan användas för att enkelt lagra signaldata med tillhörande metadata på ett strukturerat sätt.
- När en databas eller annan teknik för datahantering installeras behöver en avvägning göras mellan det arbete som krävs för insamling och lagring av data (*data ingestion*) och det arbete som krävs för att använda data (*data retrieval*).

TACKORD

Att skriva detta kapitel om strukturerad lagring av data, specifikt för VA-branschen, kräver en hel del domänexpertis. För att göra denna sammanställning så heltäckande som möjligt har författarna förlitat sig på synpunkter från flera experter inom området. Vi vill därför tacka Dr Stijn Van Hoey

(Fluves, Gent, Belgien); Frederick Van Leeckwyck (Factry BV, Gent, Belgien); Chris Hidalgo (Signature Automation LLC, TX, USA); Annika Schönfeld (Ruhrverband, Essen, Tyskland); och Manoj Yegnaraman (International Society of Automation - ISA Water/Wastewater Division, NC, USA) för deras bidrag och givande diskussioner som bidragit till detta arbete.

REFERENSER

Beaulieu A. (2020). Learning SQL: Generate, Manipulate, and Retrieve Data. 3rd edn. O'Reilly Media Inc., Sebastopol, CA, USA.

Camhy D., Gamerith V., Steffelbauer D., Muschalla D. and Gruber G. (2012). Scientific data management with open source tools – an urban drainage example. Proceedings IWA/IAHR 9th International Conference on Urban Drainage Modelling, September 4–6, 2012, Belgrade, Serbia. URL: http://hikom.grf.bg.ac.rs/starisajt/9UDM/ShortAbstracts/158-ShortAbstract.pdf

Chiarot G. and Silvestri C. (2021). Time series compression survey. *ACM Computing Surveys*, **55**(10), 1–32, accepted. https://doi.org/10.1145/3560814

Choudhury A. A. S., Thornhill N. F. and Shah S. L. (2008). Diagnosis of Process Nonlinearities and Valve Stiction: Data Driven Approaches. Springer-Verlag Berlin, Heidelberg, Germany.

Correll M. and Gleicher M. (2016). The semantics of sketch: Flexibility in visual query systems for time series data. IEEE Conference on Visual Analytics Science and Technology (VAST), Baltimore, MD, USA, pp. 131–140, https://doi.org/10.1109/VAST.2016.7883519

Dixon J. (2014). Pentaho, Hadoop, and data lakes. URL: https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/ (accessed 27 January 2023).

Dunning T. and Friedman E. (2014). Time Series Databases: New Ways to Store and Access Data. O'Reilly Media Inc., Sebastopol, CA, USA.

EPA. (2023). What is the Water Quality Portal (WQP). URL: https://www.waterqualitydata.us/wqp_description/ (accessed 27 January 2023).

Epstein B. and Roberts P. (2022). Accelerate Machine Learning with a Unified Analytics Architecture: Deploy Machine Learning Models in Minutes, not Months. O'Reilly Media, Inc., Sebastopol, CA, USA.

ETH Zurich. (2019). WHDP documentation. URL: https://whdp.readthedocs.io/en/latest/ (accessed 27 January 2023).

Gopalan R. (2022). The Cloud Data Lake. O'Reilly Media, Inc., Sebastopol, CA, USA.

Gorelik A. (2019). The Enterprise Big Data Lake. O'Reilly Media Inc., Sebastopol, CA, USA.

Harrington J. L. (2016). Relational Database Design and Implementation: Clearly Explained. 4th edn. Elsevier, Cambridge, MA, USA.

Hauduc H., Gillot S., Rieger L., Ohtsuki T., Shaw A., Takács I. and Winkler S. (2009). Activated sludge modelling in practice: an international survey. *Water Science and Technology*, **60**(8), 1943–1951, https://doi.org/10.2166/wst.2009.223

Hay D. C. (2006). Data Model Patterns: A Metadata Map. Elsevier, San Francisco, CA, USA, https://doi.org/10.1016/B978-0-12-088798-9.X5000-1

Hernandez M. J. (2021). Database Design for Mere Mortals: 25th Anniversary Edition. 4th edn. O'Reilly Media, Inc., New York, NY, USA.

Horsburgh J. S., Tarboton D. G., Maidment D. R. and Zaslavsky I. (2008). A relational model for environmental and water resources data. *Water Resources Research*, **44**, W05406, https://doi.org/10.1029/2007WR006392

Kimball R. and Ross M. (2013). The Data Warehouse Toolkit: the Definitive Guide to Dimensional Modeling. 3rd edn. John Wiley & Sons, Inc., Indianapolis, Indiana.

Michener W. K. (2006). Meta-information concepts for ecological data management. *Ecological Informatics*, **1**(1), 3–7, https://doi.org/10.1016/j.ecoinf.2005.08.004

Middleton E. (2022). Myths about historians. AVEVA. URL: https://www.aveva.com/content/dam/aveva/documents/white-papers/WhitePaper_AVEVA_HistorianMyths_09-19.pdf

Nielsen A. (2019). Practical Time Series Analysis. O'Reilly Media Inc., Sebastopol, CA, USA.

Panoply. (2023). Data Warehouse Testing 101. URL: https://panoply.io/data-warehouse-guide/data-warehouse-testing-101/

Plana Q., Alferes J., Fuks K., Kraft T., Maruéjouls T., Torfs E. and Vanrolleghem P. A. (2019). Towards a water quality database for raw and validated data with emphasis on structured meta-data. *Water Quality Research Journal*, **54**(1), 1–9, https://doi.org/10.2166/wqrj.2018.013

Rieger L., Takács I., Villez K., Siegrist H., Lessard P., Vanrolleghem P. A. and Comeau Y. (2010). Data reconciliation for wastewater treatment plant simulation studies – planning for high-quality data and typical sources of errors. *Water Environment Research*, **82**(5), 426–433, https://doi.org/10.2175/106143009X12529484815511

Simon A. R. (2021). Data Lakes for Dummies. O'Reilly Media, Inc., Sebastopol, CA, USA.

Therrien J. D., Nicolaï N. and Vanrolleghem P. A. (2020). A critical review of the data pipeline: how wastewater system operation flows from data to intelligence. *Water Science and Technology*, **82**(12), 2613–2634, https://doi.org/10.2166/wst.2020.393

Thornhill N. F., Choudhury A. A. S. and Shah S. L. (2004). The impact of compression on data-driven process analyses. *Journal of Process Control*, **14**(4), 389–398, https://doi.org/10.1016/j.jprocont.2003.06.003

TimescaleDB. (2017). TimescaleDB vs. PostgreSQL for time-series: 20x higher inserts, 2000x faster deletes, 1.2x-14 000x faster queries. URL: https://www.timescale.com/blog/timescaledb-vs-6a696248104e/

Torfs E., Nicolaï N., Daneshgar S., Copp J. B., Haimi H., Ikumi D., Johnson B., Plosz B. B., Snowling S., Townley L. R., Valverde-Pérez B., Vanrolleghem P. A., Vezzaro L. and Nopens I. (2022). The transition of WRRF models to digital twin applications. *Water Science and Technology*, **85**(10), 2840–2853, https://doi.org/10.2166/wst.2022.107

Torkamani S. and Lohweg V. (2017). Survey on time series motif discovery. *WIREs Data Mining Knowledge Discovery*, **7**, e1199, https://doi.org/10.1002/widm.1199

doi: 10.2166/9781789065107_0103

Kapitel 5

Praktiska metoder för validering av datakvalitet från givare

J. Alferes[1], O. Grievson[2], C. Heffernan[3], D. Lumley[4], T. Massiri[5], A. Menniti[6], O. Samuelsson[7], K. Villez[8] och E. Zegers[9]

[1]VITO – WaterClimateHub, Oostende, Belgien
[2]AtkinsRéalis/gästprofessor vid University of Exeter
[3]Metropolitan Water Reclamation District of Greater Chicago, Chicago, IL, USA
[4]tidigare vid DHI Group och Gryaab AB, Göteborg, Sverige
[5]Käppalaförbundet, Stockholm, Sverige
[6]Clean Water Services, Tigard, OR, USA
[7]IVL Svenska Miljöinstitutet, Stockholm, Sverige
[8]Oak Ridge National Laboratory, Oak Ridge, TN, USA
[9]Staden Toronto, Toronto, ON, Kanada

Nyckelord: bias, drift, datakvalitet, sensorunderhåll, sensorvalidering, kvalitetsbedömning, kvalitetskontroll.

Givare används idag för både övervakning och processreglering som oavsett syfte genererar högfrekventa onlinemätningar (sekunder, minuter). Parallellt med dessa mätningar så används även laboratoriemätningar som istället samlas in med låg frekvens (dygn, veckor). Kvalitetskontrollen av data från givare är sämre strukturerade än rutinerna för laboratorieanalyser på de flesta reningsverk. Detta leder till felaktiga givardata med okänd noggrannhet.

Detta kapitel beskriver hur en rutin (*standard operating procedure – SOP*) kan tas fram för att förenkla kvalitetsbedömning och kontroll av givardata och efterföljande underhållsåtgärder genom att producera relevanta metadata för givaren. Dessutom ges förslag på valideringsrutiner för de vanligaste givarna på ett reningsverk som inspirerats av praxis från reningsverk och universitet. Detta kapitel kräver förståelse för definitionerna i Kapitel 3 som här utvidgas med ytterligare definitioner relaterat till givarunderhåll. Metoderna i detta kapitel utgår från validering med referensmätningar, som kompletteras av Kapitel 6 som istället för referensmätningar utgår från dataanalytiska tekniker.

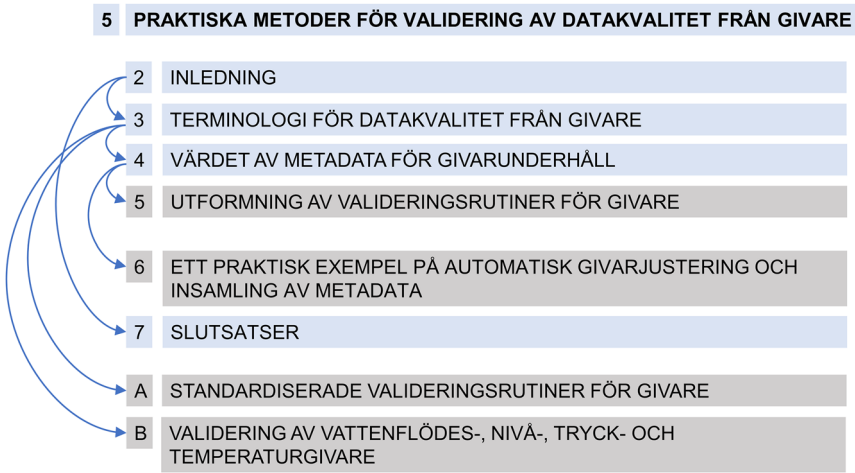
© 2025 IWAP. Detta är ett Open Access-bokkapitel som distribueras enligt villkoren i Creative Commons Attribution License (CC BY-NC-ND 4.0) som tillåter kopiering och vidaredistribution för icke-kommersiella ändamål utan derivat, förutsatt att arbetet citeras korrekt (https://creativecommons.org/licenses/by-nc-nd/4.0/). Kapitlet är från boken *Bästa möjliga data från givare i avloppsreningsprocesser*, Kris Villez, Daniel Aguado, Janelcy Alferes, Queralt Plana, Maria Victoria Ruano, Oscar Samuelsson (redaktörer).

5.1 LÄSANVISINGAR

Målen med detta kapitel är att:

(1) Illustrera hur standardiserade valideringsförfaranden av givare kan användas för att producera metadata som i sin tur är användbara för att bedöma sensorns långsiktiga prestanda och datakvalitet.
(2) Att tillhandahålla minst en standardiserad valideringsrutin (SOP) för vanliga givartyper såsom syrehalt, pH, ammoniumkväve m.m.
(3) Ge exempel på hur automatisk givarkalibrering och insamling av metadata kan genomföras i praktiken.

Figuren nedan visar kapitlets struktur. Avsnitt markerade med blått anses väsentliga för att förstå kapitlet. Grå avsnitt kan ses som fördjupning och detaljer. Vid behov bör definitionerna av datakvalitet i Kapitel 3 repeteras, alternativt genom att konsultera ordlistan i slutet av boken.



5.2 INLEDNING

Det blir allt viktigare att tillämpa ett systematiskt tillvägagångssätt för kvalitetsbedömning och kvalitetskontroll (QA/QC) av givardata i takt med att VA-organisationerna i allt högre grad förlitar sig på just givardata för ombyggnationer, driftövervakning och processoptimering och reglering. Detta blir extra viktigt när givardata dessutom delas inom en stor organisation eller med externa partners.

Givare på avloppsreningsverk behöver regelbundet underhåll såsom rengöring, komponentbyten, kalibrering och validering. Tyvärr registreras sådana underhållsåtgärder inte på ett systematisk sätt. Därför är det ibland svårt, eller till och med omöjligt, att utvärdera givarens skick eller kvaliteten på insamlade data efter längre tidsperioder, särskilt när nyckelpersoner slutat. Då återstår expertbedömningar av datakvaliteten, som kan vara subjektiv och eller svåra att genomföra.

Till skillnad från givare så kontrolleras kvaliteten på laboratoriemätningar på ett rutinmässigt sätt med standardmetoder. Många laboratorier nyttjar inte enbart standardmetoder, utan kontrollerar även kvaliteten på själva standardmetoderna regelbundet, t.ex. genom årlig granskning och revidering. Det här kapitlet är ett första steg för att korrigera denna obalans mellan valideringsrutiner av givare och laboratorieanalyser.

Genom att tillhandahålla mallar och exempel på valideringsrutiner för givare är förhoppningen att VA-organisationerna kommer tillämpa valideringsrutiner för givardata på samma nivå som

de standarder som nyttjas i laboratorieverksamhet, och integrera dessa i sina driftsrutiner. Dessa valideringsrutiner är framtagna för att producera pålitliga referensmätningar och givarsignaler under kontrollerade förhållanden så datakvaliteten kan undersökas med en enkel jämförelse. Dessa exempel på valideringsrutiner är också ett steg mot kunskapsutbyte och att dela valideringsrutiner för givare mellan olika VA-organisationer.

5.3 TERMINOLOGI FÖR DATAKVALITET FRÅN GIVARE

I det här avsnittet definieras centrala begrepp som rör datakvalitet från givare. Vissa praktiska aspekter betonas i separata stycken som markeras med rubriken *På reningsverket* (*utility*). Slutligen så illustreras de två viktigaste begreppen med två exempel (E5.1 och E5.2). Några av termerna har redan definierats i Kapitel 3 och finns även i ordlistan längst bak i den här boken: bias, riktighet, precision, varians och noggrannhet. Dessa begrepp bör läsaren vara väl bekant med eftersom de är återkommande i detta kapitel.

5.3.1 Definitioner

Definition D5.1: En **givare** är en enhet som kommunicerar värdet på en processvariabel (definition utifrån Asociación Española de Normalización, 2004).

Genom att använda denna definition exkluderas vissa delar av datainsamlingsinfrastrukturen som påverkar kvaliteten på givardata. Exempel på sådana faktorer är:

- Mätmiljöns tillstånd (t.ex. flödeskanalens tillstånd för nivåbaserad mätning av flödeshastighet)
- Tillståndet för andra enheter i dataflödet, t.ex. sändare, kabelintegritet, IT-infrastruktur.

Det främsta skälet till att utesluta dessa faktorer är att kvalitetsbedömning och kontroll av dessa faktorer inte kräver ett tillvägagångssätt som är unikt för tillämpningen reningsverk. Befintliga standarder från elektronik och/eller IT-området bör användas för validering av infrastruktur för datainsamling. Fokus i detta kapitel ligger istället på validering av själva givaren.

Definition D5.2: **Givarvalidering** (*Quality assurance – QA*) inkluderar alla åtgärder som utförs för att verifiera och dokumentera att en givare har full funktion och producerar data med önskad noggrannhet.

Givarvalidering kan omfatta, men är inte begränsat till, kalibrering, granskning av metadata, signalverifiering, validering och visuell inspektion.

På reningsverket: Givarvalidering styrs i praktiken av (a) specifika rutiner för olika givare, inklusive jämförande mätningar med laboratorietester, standardlösningar eller ytterligare givare och (b) analys av de resultat som produceras under sådana rutiner. Som exempel kan nämnas att trendanalys av givarens riktighet (*trueness*) är centralt för att validera en givare.

Definition D5.3: **Givarunderhåll** omfattar alla åtgärder som krävs för att bibehålla givaren i full funktion under hela dess livscykel.

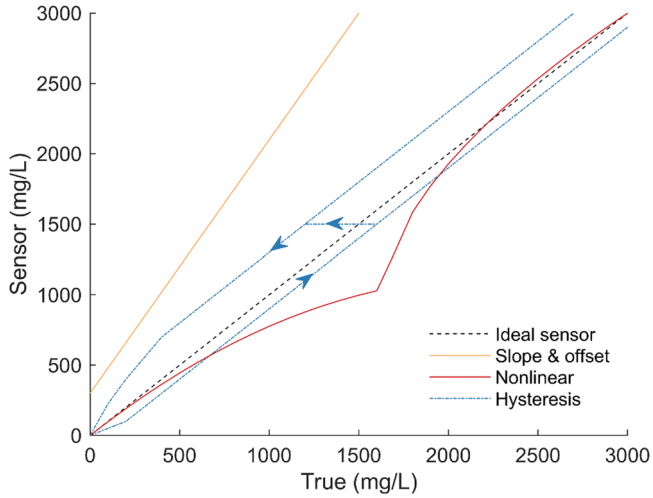
Vanliga underhållsåtgärder för givare är: kalibrering, rengöring, reparation och byte av reservdelar, uppdatering av programvara och dokumentering av utförda underhållsåtgärder.

På reningsverket: Dokumenterat givarunderhåll blir metadata för givarunderhåll. Det utförda och planerade givarunderhållet kan registreras i ett dedikerat underhållsprogram, i loggböcker eller i ett kalkylblad. Metadata för givarunderhåll är avgörande för givarvalidering. En viktig utmaning är att integrera metadata för givarunderhåll tillsammans med givarsignalerna.

Definition D5.4: En givares **offset** (*offset*) är en konstant avvikelse från det accepterade referensvärdet.

En bias som orsakas av en offset resulterar i samma bias oavsett mätvärde, i motsats till ett fel som orsakar en förändring av lutningen. Den gula linjens offset vid 350 mg/L i Figur 5.1 illustrerar detta.

Definition D5.5: En givares **lutning/känslighet** (*slope/sensitivity*) är den förändring av den uppmätta variabeln (utdata) som fås för en standardiserad förändring av de sanna variabeln/halten (indata).



Figur 5.1 Illustration av givarfelen: lutnings- och offsetfel, olinjäritet och hysteres.

Om in- och utdata representerar samma variabel mätt i samma skala (t.ex. ammoniummätning i laboratorium och en ammoniumgivare), är den ideala lutningen ett. Detta gäller för givare som utför s.k. direkt mätningar (se Definition D5.11). Den ökade känsligheten i den gula linjen, jämfört med den ideala känsligheten för den streckade linjen i Figur 5.1, illustrerar detta koncept.

I många givare är den uppmätta variabeln korrelerad med en variabel av intresse (se definition D5.11). I dessa fall så kommer den ideala lutningen inte att vara lika med ett. Som exempel kan nämnas att en pH-givare mäter en elektrisk potential (i mV) för att korrelera den med ett pH-värde (logaritmen av protonaktiviteten). I detta fall är den ideala och teoretiska känsligheten under standardförhållanden lika med 59,15 mV per pH-enhet. Den ideala lutningen specificeras vanligtvis av tillverkaren eller genom en teoretisk härledning.

På reningsverket: Givarens lutning kan vara en indikation på givarens tillstånd. Lutningen för en galvanisk Clark-cellelektrod (syrehaltsgivare) minskar t.ex. successivt på grund av normalt slitage tills membranet behöver bytas ut.

Definition D5.6: En givares **linjäritet** (*linearity*) beskriver i vilken utsträckning en förändring i den sanna halten/variabeln (indata) återspeglas som en linjär förändring i själva mätningen (utdata) över hela dess mätområde.

Mycket noggranna referensmätningar i hela mätområdet krävs för att fastställa en givares linjäritet. Av denna anledning bedöms linjäriteten sällan på plats, utan istället i givarleverantörernas laboratorier (fabrikskalibrering/validering). Betydande olinjäriteter kan dock upptäckas under valideringen om flera referensmätningar görs. Den röda linjen i Figur 5.1 exemplifierar en fiktiv olinjäritet i en givare.

På reningsverket: Olinjära effekter leder till fel i mätningen som inte kan korrigeras med kalibrering av offset eller lutning. Det är viktigt att förstå om det finns olinjära effekter för att förstå begränsningarna för givarens noggrannheten när en linjär kalibreringskurva används, eller för att mildra dessa begränsningar genom att ta fram en olinjär kalibreringskurva.

Definition D5.7: **Hysteres** är ett fenomen där givarens mätvärde beror på riktningen för förändringen av mätvärdet.

Den blå linjen i Figur 5.1 illustrerar detta koncept. För ett ökande sant värde följer mätningen den nedre blå streckade linjen. Men när den sanna koncentrationen minskar avger givaren ett konstant värde tills den övre blå streckade linjen uppnås och följer sedan denna linje för minskande värden.

På reningsverket: Mätningar av luftflödeshastighet kan uppvisa hysteres i förhållande till mätningen av ventilpositionen. Detta kan bero på glapp i ventilen, vilket innebär att ventilpositionen

beror av ventilens rörelseriktningen (öppnanden eller stängande riktning). Notera att det är själva ventilpositionen som ger upphov till en hysteres i detta fall. Om mätning av ventilpositionen saknas observeras istället hysteresen via luftflödet.

Definition D5.8: **Givarvalidering** eller givarverifiering avser ett systematiskt förfarande för att (a) samla in och registrera mätvärden under kontrollerade förhållanden för att kvantifiera datakvaliteten genom kvalitetsmått som noggrannhet, bias, precision, offset, lutning och/eller linjäritet och (b) för att utvärdera om kvalitetsmåtten uppfyller förutbestämda specifikationer för den specifika tillämpningen. Den enklaste formen av validering är att jämföra givarens mätvärde med en referensmätning.

Termen validering används ibland synonymt med verifiering och definitionerna varierar mellan olika områden och tillämpningar. I kvalitetsrutinerna enligt ISO 9000 beskrivs verifiering som ett förfarande som ger objektiva bevis för att specificerade noggrannhetskrav har uppfyllts. En givare anses verifierad när givaren är installerad, rengjord och kalibrerad enligt tillverkarens rekommendationer.

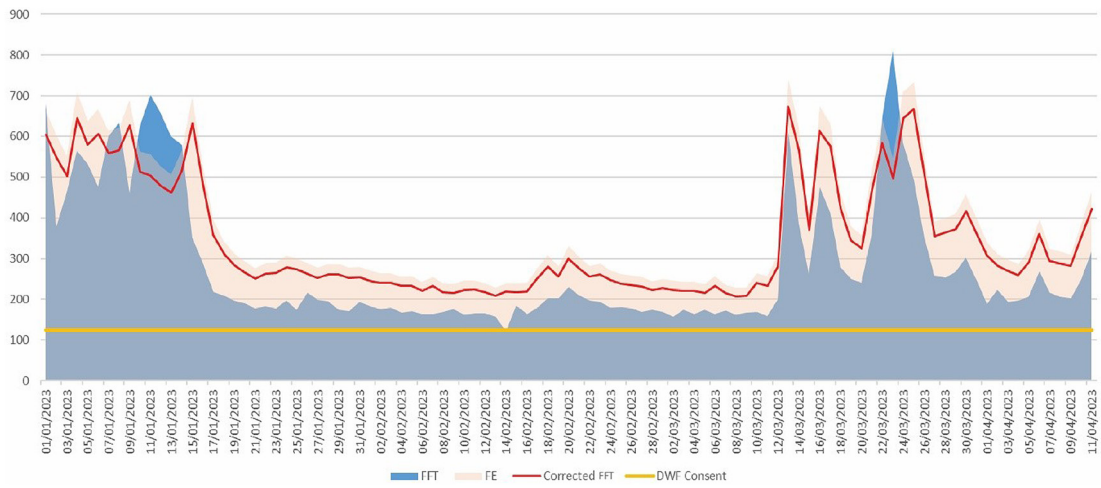
Validering anses kräva ytterligare anpassning till lokala kvalitetskrav och behöver uppfylla slutanvändarens krav på noggrannhet. I den här rapporten så används validering och verifiering synonymt.

Standardrutiner för att validera de vanligaste givarna på reningsverk finns i Bilaga 5.A. Det vanligaste sättet att validera en givare på är att jämföra och registrera avvikelsen mellan en givaren och en accepterad referensmätning.

På reningsverket: Den största fördelen med givarvalidering är att man givarens noggrannhet kvantifieras, vilket i sin tur leder till en bättre förståelse för den erhållna datakvaliteten från givaren utifrån befintligt givarunderhåll.

Definition D5.9: **Signalverifiering** (*end-to-end-testing*) avser ett systematiskt förfarande för att verifiera att givarsignalen förblir oförändrad mellan givaren och platsen för själva datalagring, t.ex. en historikdatabas.

Ett exempel på signalverifiering illustreras av Figur 5.2 som visar den totala dygnsvolymen (i m^3/ dygn) på ett reningsverk under en 3-månadersperiod. På detta reningsverk finns två flödesgivare, en i början av verket och den andra vid utloppet. I den bästa av världar så borde båda flödesgivarna ge samma resultat, bortsett från slumpmässiga mätfel. Men som figuren tydligt visar så är flödet in till



Figur 5.2 Flödesmätningar på ett avloppsreningsverk: FFT mäts vid inloppet och utflödet (FE) mäts vid utloppet. Ett fel i överföringen gör att FE systematiskt är större än FFT där de borde vara lika stora. Den här typen av fel kan upptäckas och elimineras genom att verifiera att signalöverföringen är korrekt.

reningsverket (FFT, blå skugga) betydligt lägre än utflödet (FE, rosa skugga). Detta skulle innebära att reningsverket producerar vatten, vilket förstås inte är fallet.

Om man jämför de båda givarsignalerna så kan man se att förhållande mellan signalerna är konstant, vilket tyder på ett systemfel i flödesmätningarna. Det beräknade förhållandet mellan flödesmätningarna kan användas för att korrigera utgående flödesmätning (linjen i rött) som följer trenden för utgående vatten bättre (även om den inte är exakt). Detta är ett typiskt mönster som ses vid datakvalitetsfel där en av flödesmätarna är felaktigt skalad – det vill säga instrumentets mätområde stämmer inte överens med det mätområde som används i mätsystemet. Verifiering av överföringen används för att upptäcka och eliminera den här typen av fel.

På reningsverket: Typiska fel som kan upptäckas under verifieringen av signalöverföringen är brus i den analoga överföringen och ändrad ström (mA-skalning), programmeringsfel, ändring av enheter och osynkroniserade tidsstämplar.

Definition D5.10: **Kalibrering** innebär att justera en givare för att uppnå önskad noggrannhet genom att matcha givarens utdata med ett eller flera accepterade referensvärden.

Kalibrering utförs vanligen efter en validering av givaren. Syftet med kalibreringen är att minska givarens bias. Kalibreringen kan kallas enpunkts-, tvåpunkts- eller flerpunktskalibrering, beroende på antalet referensmätningar som används. Givarens offset och/eller lutning justeras under kalibreringen beroende på grundorsaken till den identifierade biasen. Figur 5.3 illustrerar kalibrering av en givares lutning och offset.

De tekniska beräkningarna vid en kalibreringen utförs ofta automatiskt i givarens interna signalenhet, och endast referensmätvärdet matas in manuellt. Givarens interna kalibreringskurvor, inklusive offset- och lutningsvärden, kan dock vara tillgängliga som metadata, antingen genom manuell avläsning av signalenheten/transmittern eller automatiskt via digitala kommunikationsprotokoll såsom Modbus och Profibus. En alternativ kalibreringsmetod som inte kräver tillgång till givarens signalenhet beskrivs i Exempel E5.2.

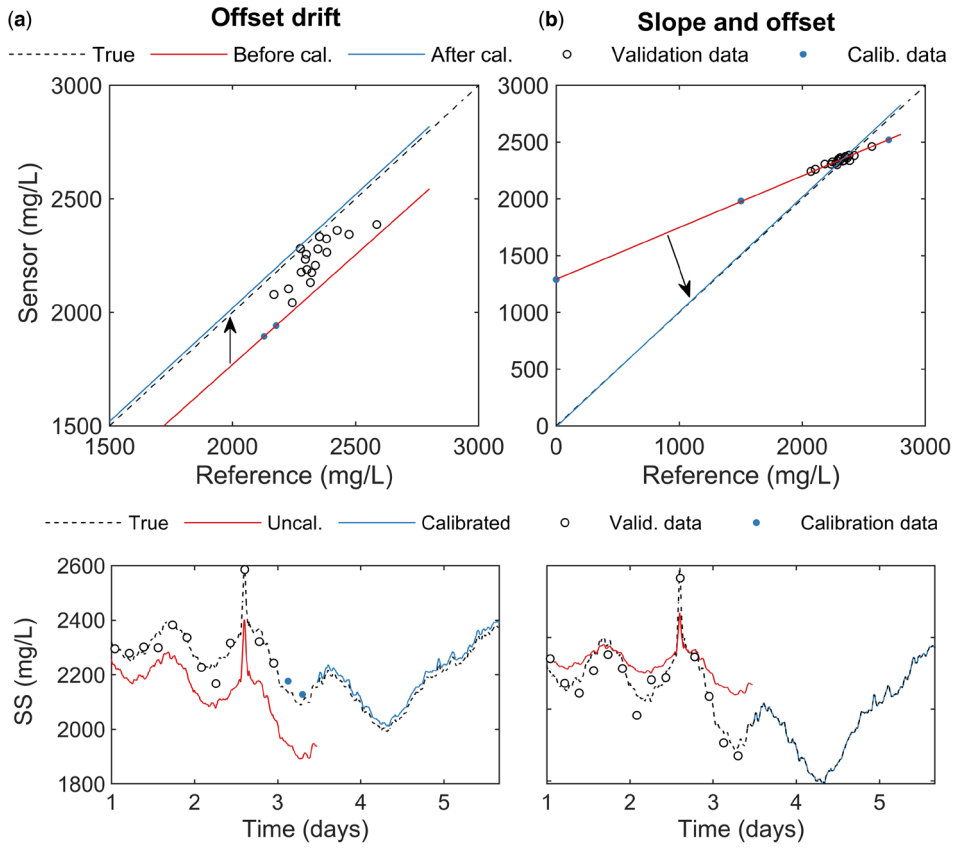
På reningsverket: Huvudsyftet med kalibrering är att säkerställa att den producerade givarsignalen matchar det verkliga värdet så nära som möjligt. Många givare kräver kalibrering med regelbundna tidsintervall för att säkerställa att effekten av åldrande och förändringar i mätstörande faktorer, vanligen observerade som drift, kan kompenseras. Observera att hela mätområdet bör täckas in enligt Figur 5.3b och man bör undvika att koncentrera mätningarna till ett smalt koncentrationsområde såsom i Figur 5.3a.

Definition D5.11: En **direkt mätning** är när givaren direkt mäter den komponent som är av intresse.

Ett exempel på en givare som använder sig av direkt mätning är en konduktivitetsgivare, som direkt mäter hur väl en ström överförs i mediet, det vill säga dess konduktivitet. De flesta givare använder sig dock av indirekta mätningar.

En indirekt mätning mäter en korrelerad komponent, t.ex. ljusabsorptionen vid en viss våglängd, som ersättning (*proxy*) för den önskade variabeln, t.ex. suspenderade ämnen. På samma sätt baseras jonselektiva elektroder (ISE) på mätning av en elektrisk potential som kan korreleras med koncentrationen av en eller flera joner. Denna teknik används vanligen för att mäta pH och koncentrationen av ammonium, nitrat och kalium. Den indirekta mätsignalen omvandlas sedan till den önskade komponenten i givaren. Detta innebär att en förändring i ett mätstörande ämne kan påverka den variabel som mäts, medan den intressanta variabeln inte påverkas av denna faktor, vilket leder till en felaktig mätning. Av denna anledning är det viktigt att förstå själva mättekniken och dess begränsningar.

På reningsverket: De flesta givare på reningsverk ger indirekta mätningar. Endast konduktivitet och oxidations- och reduktionspotential (redox) är direkta mätningar. Alla givare som kräver någon form av kalibrering före användning betraktas som indirekta mätningar. Till exempel så baseras pH-mätningar på mätning av en elektrisk potential och totalhalten suspenderade ämnen baseras på spridningen av ljus och dess intensiteten vatten. En kalibreringskurva används för att korrelera den uppmätta variabeln (t.ex. elektrisk potential) med den intressanta variabeln (pH). Kalibrering blir naturligtvis avgörande för att få korrekta mätningar.



Figur 5.3 (a) Korrigering av en givares offsetdrift med en offsetkalibrering baserad på medelvärdet av två datapunkter. (b) Lutnings- och offsetfel korrigerat med en lutnings- och offsetkalibrering baserad på tre datapunkter (jfr. Kapitel 2) utspridda över hela mätområdet. Observera att det verkliga värdet endast visas i illustrativt syfte i detta simulerade exempel och är okänt i praktiken.

5.3.2 Exempel E5.2: kalibrering av offset- och skalningsfel

Kalibrering utförs för att reducera systematiska avvikelser mellan det sanna värdet (som är okänt) och mätningen av denna, d.v.s. själva givarsignalen. Vid kalibrering så justeras givarens offset och/eller lutning. Minst två referensmätningar vid olika koncentrationer (kallas ibland tvåpunktskalibrering) behövs för kalibrering av både offset och lutning (Figur 5.3b), medan minst en referensmätning krävs för att enbart kalibrera offset eller lutning (Figur 5.3a; se även exemplet i Avsnitt 5.4.1.2). Kalibreringens effekt förbättras om fler mätningar än vad som är absolut nödvändigt används, och mätningar som helst täcker in hela intervallet som processen förväntas variera inom.

Vid en offsetkalibrering används en referensmätning för att justera givarsignalen. Notera att en referensmätning också kan baseras på medelvärdet av flera mätningar för att öka dess precision. En offsetkalibrering påverkar alla efterföljande mätningar genom att förskjuta de registrerade värdena efter kalibreringstillfället.

Termerna enpunkts- och tvåpunktskalibrering är vanligt förekommande men kan vara förvirrande eftersom de endast anger antalet referensmätningar som krävs för kalibrering och inte hur den faktiska kalibreringen utförs. Därför är det bra att läsa igenom givarinstruktionerna i detalj för att

fullt ut förstå effekterna av själva kalibreringen. Avancerade instrument kan kräva mer än två punkter för kalibrering. Exempel på detta är (a) att justera kovarians mellan olika variabler i multivariata spektrofotometrar eller (b) att detektera olinjära effekter i givare.

Kalibrering vid instabila förhållanden, t.ex. när givarens temperatur skiljer sig från omgivnings, kan resultera i ett kalibreringsfel, som i sin tur leder till systematiska mätfel. Potentiella störningskällor ska identifieras och, om möjligt, kvantifieras. Valideringsmetoderna i Bilaga 5.A antingen (a) eliminerar sådana störningar eller (b) gör det möjligt att korrigera sådan fel med olika beräkningar för olika typer av givare.

5.4 VÄRDET AV METADATA FÖR GIVARUNDERHÅLL

En ömsesidig förståelse för givarnas datakvalitet är avgörande när data delas och nyttjas av flera olika personer. Ett bra sätt för att uppnå detta är att dokumentera givarunderhållet på ett enhetligt och heltäckande sätt. Dokumentationen kan omfatta rengöring, kalibrering, validering och utbyte av delar av eller hela givare. I det här kapitlet rekommenderas att valideringsrutiner används specifikt för att validera givare, eftersom detta producerar metadata som är särskilt användbara för att övervaka givarnas tillstånd och deras datakvaliteten. Valideringsrutiner för givare kan integreras i befintliga drift- och kvalitetsrutiner.

På kort sikt kan rutinmässig utvärdering av datakvalitetsmått (noggrannhet, bias, offset, precision, lutning och/eller drift) användas för att fastställa en optimal frekvens för rengöring och kalibrering, vilket i sin tur gör åtgärderna mer effektiva. På lång sikt kan väldokumenterade valideringsrutiner av givare förenkla delning av data med personal inom organisationen samt med externa projektpartners. I följande avsnitt beskrivs några av de fördelar som insamlingen av metadata från givarunderhåll kan ge.

5.4.1 Optimering av kalibreringsfrekvensen

Det enklaste sättet att förbättra givarhanteringen på de flesta reningsverk är att systematiskt följa hur kalibreringskurvan för en givare förändras över tiden. Det primära syftet med kalibrering är att motverka drift. När en tvåpunktskalibrering utförs regelbundet kan man därför spara parametrarna för denna kurva (offset och lutning) varje gång en kalibrering utförs. Vissa moderna givare för även interna loggar över förändringar i kalibreringskurvan, som i många fall direkt kan överföras till en databas med hjälp traditionella kommunikationsprotokoll (t.ex. Ethernet och Modbus).

5.4.1.1 Optimering av kalibreringsfrekvens utifrån drift

Hur snabbt en givare driver beror ofta på vilken typ av avloppsvatten den mäter. Det innebär att noterade förändringarna i offset och skalning är specifika för givaren och dess plats på reningsverket. Drifthastigheten kan även förändras över tid, till exempel när avloppsvattnets sammansättning plötsligt förändras. Därför bör kalibreringsfrekvensen anpassas specifikt till givaren och det medium som den utsätts för. I de fall avdriftens hastighet förändras över tid kan en adaptiv strategi krävas där tiden mellan kalibreringarna justeras från en kalibrering till en annan. En optimal kalibreringsfrekvens är förstås enbart möjlig att uppnå om kalibreringskurvorna registreras och lagras på ett systematiskt sätt. Genom att justera givarnas underhållsfrekvens med väldokumenterade kalibreringskurvor kan man minska onödiga underhållskostnader, säkerställa att datakvaliteten är tillräcklig och därigenom öka förtroendet för de data som produceras.

Exempel E5.3 beskriver hur förändringar i både offset och lutning analyseras varje gång givaren kalibreras. Det finns även varianter på kalibreringsstrategier:

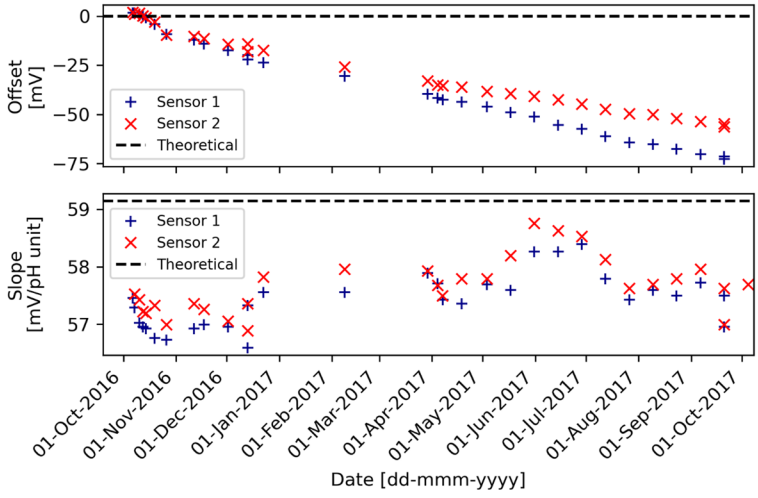
(1) Utvärdera om en kalibrering är nödvändig genom att först utvärdera offset och lutning och uppdatera endast kalibreringskurvan när en tillräckligt stor förändring noteras.
(2) Övervakning av offset och lutning ersätts med en procedur som endast följer en parameter i kalibreringskurvan (offset eller lutning). Detta kan vara användbart om (a) det är känt att

givaren inte har någon förväntad förändring i vare sig offset eller lutning under sin livstid, eller (b) det anses vara för besvärligt att producera konstanta mätningar för mer än ett referensvärde för de uppmätta variablerna.

5.4.1.2 Exempel E5.3: övervakning av offset och lutning för två pH-givare

Detta exempel illustrerar systematisk insamlingen av metadata från två pH-givare utifrån valideringsrutinen 'pH-rutin-2' (version 1.0, se Bilaga 5.A.2.2). För en typisk jonselektiv pH-givare som används på reningsverk beskrivs förhållandet mellan den uppmätta variabeln (elektrisk potential) och själva variabeln med ett linjärt samband som även kallas Nernst-ekvationen. Offset definieras som jämviktsvärdet för den elektriska potentialen som givaren ger när den exponeras för ett neutralt vatten (pH 7). Vid standard temperatur och tryck (SATP) är den teoretiska offset-värdet 0 mV. På samma sätt är lutningen förändringen i elektrisk potential för en pH-enhets förändring, givet mätvärden i jämvikt före och efter förändringen. Vid SATP är det teoretiska värdet för lutningen 59,16 mV för de flesta pH-givare.

Figur 5.4 visar uppmätta värden för offset och lutning för två vanliga typer av pH-givare i en nitrifikationsbassäng under 1 års tid. Valideringsrutinen 'pH-rutin-2' användes för att uppskatta givarnas offset och slope. Man kan se att offset ökar till över 60 mV efter 1 års drift (övre grafen), vilket skulle motsvara en förändring på 1 pH-enhet om givaren aldrig kalibrerades. Lutningen är däremot stabil och ligger inom 2 mV/pH-enhet under hela tidsperioden. Den uppmätta hastigheten med vilken offset ändras (ca 60 mV per år, eller 1 pH-enhet vid konstant lutning) kan användas för att välja en optimal kalibreringsfrekvens, där noggrannhet ställs mot tid för kalibrering. I det här exemplet innebär det att en kalibrering var 30:e dag är mer än tillräckligt för att bibehålla en offset med en storlek som är mindre än 6 mV (ca 0,1 pH-enhet).



Figur 5.4 Offset- och slope-värden erhållna med pH-rutin-2 (version 1.0, se bilaga 5.A). Offset- och slope-värdena registreras varje gång proceduren utförs. I detta fall tyder trenderna för offset och lutning på att det kan vara tillräckligt att ta hänsyn till förändring av offset under kalibreringen, med enbart en enpunktskalibrering. Avdriften i offset är ca 60 mV per år. Det innebär att en kalibrering var 30:e dag är mer än tillräckligt för att bibehålla kalibreringskurvans offsetparameter inom 6 mV från sitt sanna värde och därmed hålla det uppmätta pH-värdet inom 0,1 pH-enheter från det sanna värdet.

5.4.2 Optimal rengöringsfrekvens

Rengöring av givare är en grundläggande men tidskrävande underhållsåtgärd. Ofta utförs rengöring som en fast rutin, vilket gör det viktigt att optimera rengöringsintervallet.

Förändringar i givarens datakvalitet på grund av nedsmutsning (t.ex. offset och lutning) kan kvantifieras genom att notera givarens värde före och efter rengöring – förutsatt att givaren mäter vid konstanta förhållanden. Systematisk kvantifiering av rengöringseffekten är ett enkelt verktyg för att optimera rengöringsintervallet. Den systematiska insamlingen av metadata kräver bara att man lägger till tiden för rengöring om givarvärdena före och efter rengöring redan finns lagrade i en databas.

5.4.3 Delning av data inom reningsverket

Användarna av givardata skiljer sig oftast från den personal som ansvarar för själva givarunderhållet och bedömning av givarnas datakvalitet. När antalet givare ökar kan det vara svårt för en dataanvändare att fullt ut lita på tillgängliga data om det inte finns ytterligare metadata som beskriver rutinmässigt underhåll och kvalitetsbedömningar. Kvantifierade effekter av validering, kalibrering och underhåll av givare kan underlätta bedömning av givarnas datakvalitet och tillämpbarhet för olika ändamål. Det ökande beroendet av givare och data för att klara den vardagliga driften motiverar också ett större fokus på datakvaliteten. Ett systematiskt arbetssätt utifrån de beskrivna valideringsrutinerna i Bilaga 5.A rekommenderas för att fånga upp långsiktiga trender i givarnas tillstånd och kvaliteten på de data som de producerar.

5.4.4 Delning av data mellan organisationer och modellbygge

De givardata som samlas in på reningsverk används i allt större utsträckning av externa användare och data delas rutinmässigt med konsulter, tillsynsmyndigheter, andra industrier och forskare. Metadata som beskriver hur givarnas datakvaliteten bedöms kan vara avgörande för att: förstå data, bedöma deras tillämplighet för olika ändamål och för att kunna kommunicera viktiga händelser och trender i data med en gemensamt språk.

Ett exempel på när kvalitetsmått för givares datakvalitet är användbart är när data från två givare används för modellering. Om en datadriven modell (antingen en processmodell eller AI-modell) ska kalibreras utifrån data från två givare, A och B med mätningarna, y^A och y^B ($i = 1, \ldots, N$) så är det vanligt att givarnas data viktas utifrån deras respektive precision. Logiken med detta är att en låg precision innebär en hög varians som innebär att man bör lita mindre på en sådan mätning, jämfört med en mätning med hög precision. Vid modellkalibrering så normaliseras residualerna mellan modellens predikterade värde ($\hat{y}^A$) och det uppmätta värdet (y^A) med givarens standardavvikelse (σ^A). De normaliserade ('standardized') residualerna är då lika med:

$$r_i^A = \frac{\hat{y}_i^A - y_i^A}{\sigma_i^A}, \quad r_i^B = \frac{\hat{y}_i^B - y_i^B}{\sigma_i^B}$$

De normaliserade residualerna ges sedan lika stor vikt vid modellkalibrering. Summan av de residualerna i kvadrat (WSSR) är ett vanligt mått för att bedöma modellkvalitet och beräknas enligt följande:

$$\text{WSSR} = \sum_{i=1}^{N} \left(r_i^A\right)^2 + \sum_{i=1}^{N} \left(r_i^B\right)^2$$

Mätfelets standardavvikelse varierar vanligen från en givare till en annan och kan även variera i tiden. Tyvärr är det svårt att få tillförlitliga uppskattningar av mätfelens standardavvikelser direkt från mätningarna (y^A, y^B) under dynamiska, okontrollerade förhållanden.

Tillgång till rutinmässigt uppskattningar av mätfelets standardavvikelse underlättar definitionen av kvalitetsmått för alla modeller, oavsett om de baseras på maskininlärning eller fysikaliska principer. Ett standardiserat tillvägagångssätt för regelbunden och systematisk utvärdering av givarprecisionen kan därför vara värdefullt för att utvärdera modellers tillförlitlighet.

5.5 UTFORMNING AV VALIDERINGSRUTINER FÖR GIVARE

Det är enkelt att ta fram en rutin för givarvalidering. Däremot är det svårt att säkerställa att en givare valideras på ett enhetligt sätt (av alla inblandade instrumenttekniker) när valideringsrutiner saknas. Det här avsnittet ger några rekommendationer om vad som bör ingå i en valideringsrutin för givare som möjliggör ett konsekvent tillvägagångssätt för validering och underhåll av givare.

5.5.1 Ett strukturerat arbetssätt för validering av givare

Tre komponenter är centrala för att implementera en valideringsrutin för givare: syfte, repeterbarhet och lagring av metadata.

- **Syfte.** Syftet med valideringen behöver vara tydligt eftersom det avgör hur detaljerad och omfattande valideringen behöver vara. Följande frågor kan användas som stöd för att definiera syftet med givarvalideringen:
  - Vilka kvalitetsmått för givardata (t.ex. noggrannhet, bias eller precision) är relevanta för tillämpningen och inom vilka intervall/gränsvärden behöver de hålla sig?
  - Vilka praktiska motåtgärder kan sättas in (a) för att undvika att kvalitetsmått för givardata överskrider det acceptabla intervallet och (b) vilka motåtgärder kan sättas in när givarnas kvalitetsmått överskrider sina gränsvärden.
  - Vilken typ av information fås från valideringsrutinen och vad leder dessa insikter till i termer av åtgärder vid givarunderhåll?
- **Repeterbarhet.** En repeterbar metodik är nyckeln till konsekventa och tillförlitliga valideringsresultat. Precis som ett recept så ska valideringsrutinen vara tydlig med en steg-för-steg-guide för hela proceduren. Detta inkluderar förberedelser, förbrukningsmaterial och själva utförandet. Otillräckliga förberedelser eller brist på förbrukningsmaterial och/eller verktyg leder till stress, onödigt arbete med otillförlitliga valideringsdata som konsekvens. Med en tydlig valideringsrutin så förflyttas fokus till att noggrant genomföra valideringen i stället för att försöka komma ihåg all utrustning och hur valideringen ska utföras.
- **Lagring av metadata.** Metadata bör lagras automatiskt när så är möjligt. Till exempel kan vissa moderna givare automatiskt registrera tidpunkten för validering (den tid då givaren är placerad i en standardlösning) med en enkel knapptryckning. En sådan funktion kan tjäna två syften. För det första sker datainsamlingen på ett konsekvent sätt, vilket minimerar mänskliga fel och variabilitet. Detta förbättrar i sin tur repeterbarheten i valideringsresultaten. För det andra så underlättar automatisk datainsamling långtidslagring och återanvändning av data. Alternativet med manuell lagring av registrerade metadata i exempelvis en databas misslyckas ofta eftersom det är ett tråkigt arbetsmoment och lätt nedprioriteras. Automatisk datainsamling gör att metadata för givarvalidering blir en del av det organisatoriska minnet, snarare än otillgängliga och ostrukturerade i excelark. När metadata är lätta att hämta ut och analysera är det enkelt att följa givarens prestanda sin livslängd och även kvalitetsbedöma historiska data.

Slutligen så bör en valideringsrutin även innehålla viss generell information som listas nedan. Valideringsrutinen kan även kompletteras med en beskrivning som förklarar givarens syfte och funktion för den specifika placeringen. Bilaga 5.A. innehåller en mall för en valideringsrutin som uppfyller dessa krav vilka diskuteras i kommande avsnitt.

Viktiga komponenter i en valideringsrutin för givare

Följande delar är viktiga och bör inkluderas i alla valideringsrutiner för givare:

(1) Ange syftet med valideringsförfarandet.
(2) Bestäm hur många mätpunkter och replikat som behövs.
(3) Lista all hårdvara som krävs för genomförande.
(4) Lista alla förbrukningsvaror som krävs för genomförande.
(5) En detaljerad steg-för-steg-beskrivning av själva valideringsproceduren.

(6) En metod för att beräkna och registrera valideringsmätningar och kvalitetsmätningar av givardata som metadata.
(7) Definiera vilka andra metadata som ska registreras, t.ex. batchnumret för den standard som används eller den person som utför valideringen.

Valfria komponenter i en valideringsrutin för givare
En valideringsrutin kan även innehålla följande information, lite beroende på mätprincip, det uppmätta mediet och platsspecifika förhållanden:

- Beskrivning av mätprincip och störningsfaktorer.
- Versionshistorik för valideringsrutinen
- Syftet med givaren och dess mätning.
- Ytterligare kontroller som ska utföras före validering av givaren, t.ex:
  - Verifiering av korrekt infästning av givaren, strömförsörjningen och kablage.
  - Kontroll av givarens placering i bassängen.
  - Kontroll av och rengöring av smuts.

5.5.2 Allmänna rekommendationer för tillförlitliga referensmätningar

Generella framgångsfaktorer vid givarvalidering är representativa vattenprover och korrekta (*unbiased*) mätningar. Nedan listas fyra vanliga metoder som används för att göra korrekta jämförelser mellan en givare och dess referensmätning.

Den generella rekommendationen är att använda den s.k. hinkmetoden (Metod A). Denna metod är det enda tillvägagångssätt som säkerställer en korrekt (*unbiased*) jämförelse mellan givarsignalen och referensmätningarna, givet de begränsade studier som finns publicerade. Metadata för givares datakvalitet som tas fram med andra metoder (metoderna B, C och D) bör betraktas som mindre tillförlitliga. Av denna anledning så baseras alla valideringsrutiner i Bilaga 5.A på Metod A. Dessa valideringsrutiner kan förstås modifieras till att använda Metod B, C eller D utifrån lokala begränsningar.

(A) **Delprov med den s.k. hinkmetoden (*bucket approach*)**

En hink används för att ta ett prov från en representativ plats i processen. Koncentrationen kan ändras genom utspädning eller uppkoncentrering efter provtagningen för att justera sammansättningen på ett kontrollerat sätt. Eventuella variationer i vattenprovets sammansättning i förhållande till det ursprungliga vattenprovet betraktas som kända. Vattenprovet i hinken blandas sedan runt och givaren placeras i hinken och mäter under en längre tid tills mätningen är stabil. En automatisk omrörare eller luftning kan användas för att säkerställa en homogen blandning under mätningen och för att undvika sedimentering av suspenderade ämnen. Ett vattenprov tas sedan från hinken och analyseras i ett laboratorium som referensmätning till givarens mätning. Genom att hantera vattenprovet på detta sätt så minimeras okontrollerade variationer i vattensammansättningen, vilket gör att eventuella avvikelser mellan givaren och referensmätningen enbart anses bero på fel i givaren.

Denna metod ger en objektiv validering eftersom både givarens mätning och referensmätningen motsvarar samma välblandade medium. Dessutom är förhållandena i hinken konstanta, vilket gör det enkelt att få ett stabilt medelvärde oavsett hur stort brus givaren har.

Nackdelen med hinkmetoden är att det kan vara svårt och tidskrävande att få ett stort vattenprov från processen.

(B) **Jämförelse av flera givare på plats (in-situ)**

En referensgivare (t.ex. en handhållen enhet) placeras bredvid den stationära givaren under en längre tid. Tiden bör vara tillräcklig för att reducera eventuella effekter från signalfilter i den handhållna givaren och i SCADA-systemet (*Supervisory Control and Data Acquisition*). De två mätningarna kan jämföras med avseende på medelvärden och skillnader under processvariationer på plats.

En förutsättning är att referensgivaren är exakt, att tidsstämplarna i dataposterna för båda givarna är synkroniserade med samma klocka och att givarens signaldynamik på grund av givarens signalbehandling (t.ex. signalfilter) är försumbar. En jämförelse på plats kompliceras av det faktum att den uppmätta variabeln kan förändras under valideringsproceduren och att de två givarnas dynamiska respons kan vara mycket olika.

(C) **Flödesproportionellt 24-timmars prov**

Ett flödes- eller volymsproportionellt vattenprov tas i närheten av givaren. En laboratorieanalys av vattenprovet kan sedan jämföras med givarsignalen under samma tidsperiod som det sammansatta vattenprovet togs. Givarsignalen behöver dock justeras på samma sätt som vattenprovet. Det vill säga att givarens koncentrationsmätning justeras så att den blir flödesproportionell, innan dess medelvärde jämförs med laboratorieresultatet.

Det är sällan motiverat att installera provtagningsutrustning enbart för givarvalidering. Ibland finns dock sådan provtagningsutrustning redan på plats av andra skäl och kan då återanvändas för att erhålla en 24-h-referensmätning och givarvalidering.

(D) **De snabba stickproven**

Den vanligaste, men tyvärr också den minst exakta referensmätningen, är ett stickprov som tas så nära givaren som möjligt. Resultatet av vattenprovet kan enkelt jämföras med givarens mätvärde vid provtagningstillfället. Det snabba resultatet i kombination med att det är enkelt att lämna givaren på plats är troligen orsaken till att metoden blivit populär. Tyvärr påverkas skillnaderna mellan resultatet från vattenprovet och givaren av många okontrollerbara faktorer som gör att det inte går att använda denna metod för en opartisk och exakt utvärdering av datakvaliteten hos givare. Stora avvikelser mellan de två mätningarna behöver inte nödvändigtvis bero på felaktigheter i givaren, vilket kan leda till en pessimistisk utvärdering av givarens datakvaliteten, vilket i sin tur leder till onödiga underhållsåtgärder.

Det största problemet med stickprovsmetoden är att det är svårt att säkerställa att det vattenprov som tas är representativt för mediet. Givaren kan ligga för djupt för att man ska kunna ta ett vattenprov precis bredvid den. Det kan också vara svårt att ta vattenprovet på exakt samma plats hela tiden. Dessutom kan dynamiken i den uppmätta processen eller signalbehandlingen orsaka fördröjning och oönskade signalvariationer under valideringen, vilket gör att det registrerade värdet inte är representativt för givarens mätningar. På samma sätt kan registreringen av givarvärdet under provtagningen i SCADA (i motsats till givarens display) skilja sig åt i tid om tiden för vattenprovet registrerades från en felaktig klocka, t.ex. givarens display. Helst ska den registrerade tiden stämma överens med tiden för en klocka som är synkroniserad med SCADA-systemets klocka.

5.5.3 Mallar och exempel på standardiserade arbetsrutiner för validering av givare

Bilaga 5.A i detta kapitel innehåller 14 valideringsrutiner för fem parametrar som är vanliga på reningsverk. De parametrar som det finns valideringsrutiner för är:

(1) Löst syre (DO): tre valideringsrutiner
(2) pH: tre valideringsrutiner
(3) Ammoniumkväve (TAN): två valideringsrutiner
(4) Totalt suspenderat material (TSS): fyra valideringsrutiner
(5) Absorption av ultraviolett synligt ljus (UV/Vis): två valideringsrutiner.

Dessa parametrar valdes ut eftersom:

- Det finns mogna valideringsteknikerna för de givare och mätteknik som tillämpas idag.
- Det går att genomföra kalibrering och validering på plats för dessa parametrar.
- Givare för dessa parametrar kalibreras och valideras ofta av personal på reningsverket.

Av dessa skäl utelämnades vissa parametrar såsom:

- Löst lustgashalt: Tekniker för validering av givare är begränsade och kan komma att ändras inom kort.
- Flödeshastigheter för gas: Givarvalidering utförs vanligtvis av givartillverkaren. Ofta kan detta endast utföras vid särskilda anläggningar utanför anläggningen (t.ex. fabriksvalidering).
- Vattenflöden och nivåer: Det var inte möjligt att hitta någon stegvis procedur för valideringsrutin som tillämpas på reningsverk. För elektromagnetiska flödesmätningar är det vanligtvis givartillverkaren som ansvarar för detta. Ofta kan detta endast utföras på särskilda anläggningar utanför anläggningen (t.ex. fabriksvalidering). För nivåmätning är korrekt placering av nivågivaren och borttagning av eventuellt skräp viktigast. För nivåbaserad flödesmätning (flöde) är flödesmätarens fysiska funktion också avgörande. Själva nivågivaren anses ofta vara noggrann och driftfri. Detta kan förklara avsaknaden av befintliga metoder för validering av givare som skulle kunna samlas in för detta kapitel. Bilaga 5.B i detta kapitel innehåller anvisningar om validering av denna typ av givare. Flödeshastighetsgivare och nivågivare kan också valideras genom massbalanser som beskrivs i Kapitel 6.
- Tryck och temperatur: Dessa givare anses normalt vara noggranna och driftfria och systematiska mätfel beror oftast på fel i installationen. Detta förklarar varför upprepade valideringar sällan utförs. Bilaga 5.B i detta kapitel innehåller anvisningar för validering av denna typ av givare.

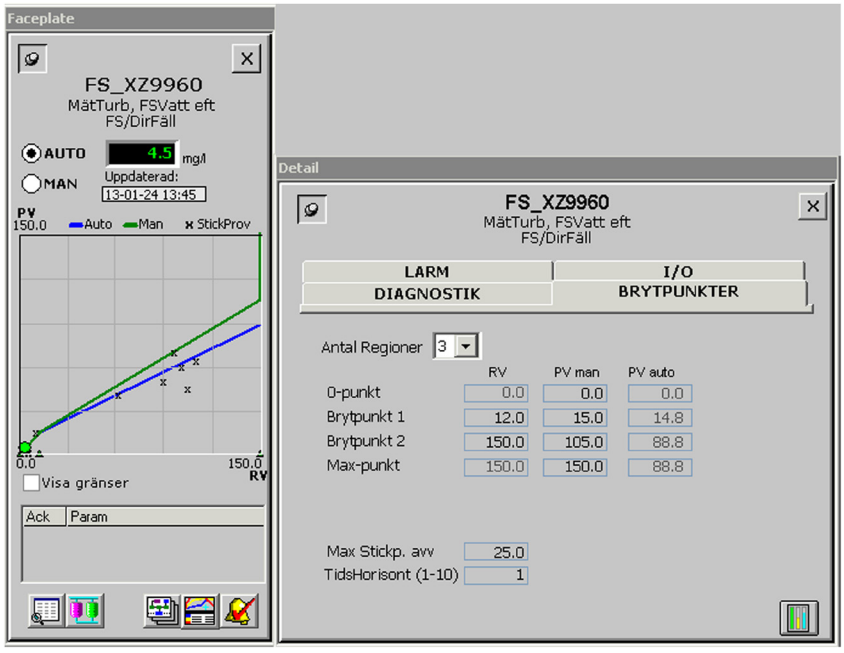
5.6 ETT PRAKTISK EXEMPEL PÅ AUTOMATISK GIVARJUSTERING OCH INSAMLING AV METADATA

En givares kalibreringskurvor lagras vanligtvis lokalt i givarens transmitter. Själva kalibreringen sker då vid givaren och endast den internt bearbetade signalen överförs till det centraliserade datainsamlingssystemet (t.ex. SCADA eller historik databas). Ett alternativ är att låta kalibreringskurvan i transmittern vara oförändrad och i stället utföra kalibreringsjusteringen i den centraliserade SCADA-miljön. Detta tillvägagångssätt har fördelen att alla metadata från kalibreringen bibehålls i den befintliga databasstrukturen och att ett mer generellt tillvägagångssätt används för samtliga givare, även om de produceras av olika tillverkare.

Det här exemplet visar hur givarsignaler justeras automatiskt efter att de har tagits in i det distribuerade styrsystemet (DCS) vid Ryaverket i Göteborg. Tillvägagångssättet kallas auto-justering (*auto-adjustment*) och exempel på vyer från Rya SCADA visas i Figur 5.5.

Auto-justering syftar till att minimera effekterna av systematiska och slumpmässiga fel i processen utan att behöva göra en omfattande instrumentkalibrering i fält. Den består av följande steg.

(a) Ta en referensmätning (vattenprov) bredvid den givare som ska justeras.
(b) Analysera vattenprovet på laboratoriet och lagra metadata, dvs. värdet på referensmätningen, i laboratoriets informationshanteringssystem (LIMS).
(c) Referensmätningen överförs sedan automatiskt från LIMS till styrsystemets DCS.
(d) Validera rimligheten i referensmätningen från (a) med en automatisk algoritm i DCS med fördefinierade förutsättningar.
(e) Uppdatera givarens råsignal DCS med en automatisk justering utifrån referensmätningen och givarens mätvärde. Justering görs med en rekursiv minsta kvadrat-algoritm med glömskefaktor (*exponential decay*). Justeringen av den råa givarsignalen utförs med hjälp av en s.k. brytpunktskurva, som är en kontinuerlig men bitvis linjär kurva (se vänstra delen i Figur 5.5). Antalet segment (brytpunkter) kan definieras av användaren för varje givare, vilket gör det möjligt att kompensera för brister i givarens linjäritet. Den här typen av kurva kallas också för *a linear spline function* (Ramsay & Silverman, 2005). Den justerade signalen lagras i styrsystemet och är därmed tillgänglig för processreglering.



Figur 5.5 En SCADA-skärmdump som visar gränssnittet för autojustering av en givare för suspenderat material. Givaren mäter två medier med olika egenskaper, direkt utfällt vatten (datakluster nära ursprunget och under 10 mg/L) och försedimenterat vatten (kluster i mitten till höger, vanligtvis över 50 mg/L). De två olika egenskaperna hanteras med en icke-linjär kalibreringskurva med en brytpunkt vid 12 mg/L.

Vid Ryaverket så bidrar auto-justeringen till minskat underhållsbehov av manuell givarkalibrering som uppskattningsvis sparar 10–40 timmar per givare och år beroende på givartyp. Auto-justeringen gör att givaren kan vara kvar i processen under hela proceduren och därmed maximera mätningens drifttid. Detta minimerar också effekterna av kalibrering med artificiella medier eller medier som kanske inte är representativa för den uppmätta miljön. Genom att undersöka hur kurvorna i funktionen förändras över tiden kan fenomen som nedsmutsning, givardrift och att givaren närmar sig slutet av sin livslängd spåras och kvantifieras. Auto-justering bidrar därmed till att minimera processtörningar på grund av felaktiga givarmätningar, vilket i allmänhet förbättrar processregleringen.

Det bör noteras att den variabilitet som införs genom referensmätningen (vattenprovet) delvis reduceras genom den rekursiva korrigeringen av flera mätningar. Detta fördröjer dock även korrigering av drift och bias. En snabbare korrigering kan uppnås med hinkmetoden för referensmätningen, i kombination med en snabbare uppdatering av kalibreringskurvan.

5.7 SLUTSATSER

På många reningsverk så lagras och analyseras endast en begränsad del av den data som samlas in vid givarunderhåll på ett strukturerat sätt. Eftersom data i allt större utsträckning nyttjas av olika användare inom samma organisation, och även delas med externa partners, så kan förtroendet för givardata förbättras med hjälp av flera enkla tekniker:

- Säkerställ att omfattning och noggrannhet vid kvalitetskontroll och validering av givare är i paritet med de rutiner som tillämpas vid ackrediterade laboratorier.

- Utveckla och tillämpa rutiner för underhåll, validering och kalibrering av givare. I detta kapitel ges några mallar och exempel på hur givarvalidering bör genomföras.
- Se till att mätresultat som erhållits med referensmaterial (t.ex. kalibreringsmedia) eller referensgivare (t.ex. handhållna givare) lagras på ett enhetligt sätt och analysera dessa mätningar för att fastställa optimala intervall för rengöring, kalibrering och andra underhållsåtgärder.

TACKORD

De beskrivna valideringsrutinerna för givare baseras på en sammanställning av befintliga rutiner som används på reningsverk och av forskare runt om i världen. Ett stort tack till Ahlem Filali (INRAE, Frankrike), Angel Roblés (Universitat de València, Spanien), Conor Heffernan och Fenghua Yang (Metropolitan Water Reclamation District of Greater Chicago, IL, USA), Kris De Gussem (Aquafin, Belgien), Linda Åmand (Käppalaförbundet, Sverige) samt Sara Nilsson och Gabriel Persson (Stockholm Vatten och Avfall, Sverige) för era bidrag med valideringsrutiner.

REFERENSER

Asociación Española de Normalización (2004). UNE EN 12255-12:2004 Wastewater treatment plants – Part 12: Control and automation. 18pp.

Mašić A., Santos A. T. L., Etter B., Udert K. M. and Villez K. (2015). Estimation of nitrite in source-separated nitrified urine with UV spectrophotometry. *Water Research*, **85**, 244–254, https://doi.org/10.1016/j.watres.2015.08.031

Ramsay and Silverman (2005). Functional Data Analysis. Springer, New York, NY, 429pp.

RealTechWater Inc. (2017). www.realtechwater.com/w_p/wp-content/uploads/2021/04/Spectrum.png. accessed 12th July, 2023

S::CAN GmbH (2020). Water Quality OnLine Waste Water. https://www.s-can.at/en/download-catalogs-brochures-and-flyer/, accessed 12th July 2023. 108pp.

Thomas O. and Burgess C. (eds). (2017). UV-Visible Spectrophotometry of Waters and Wastewater. Elsevier, Cambridge, United States, 526pp.

BILAGA 5.A: STANDARDISERADE VALIDERINGSRUTINER FÖR GIVARE

I detta avsnitt beskrivs rutiner för givarvalidering av givare som typiskt används på reningsverk. Dessa rutiner kan användas direkt eller modifieras så att det passar verksamhetens mål och begränsningar. Det är viktigt att även konsultera tillverkarens rekommendationer och riktlinjer innan rutinerna tillämpas.

5.A.1 Validering av syregivare

Detta avsnitt beskriver tre sätt att validera syregivare (DO). Figur 5.A.1 visar vilka datakvalitetsmått som kan uppskattas med respektive valideringsrutin.

5.A.1.1 DO-rutin-1: Kvantifiering av bias med en handhållen referensgivare

Version och datum

- Version: 1.0
- Senast reviderad: 31 juli 2023
- Nästa granskning och revidering ska vara genomförd senast den 31 juli 2025

Syfte

Syftet med denna rutin är att kontrollera värdet på den uppmätta syrehalten hos en syregivare i dess normala miljö med ett referensvärde som genereras av en handhållen syregivare (referensmätningen). Rutinen passar bra på reningsverk som har många syregivare eftersom samma hink med referensmedia kan återanvändas för givare i närheten.

Med denna rutin går det att:

- Utvärdera mätfelet vid den aktuella driftspunkten.
- Kvantifiera bias (se Kapitel 3)

Om rutinen upprepas regelbundet över tid går det även att:

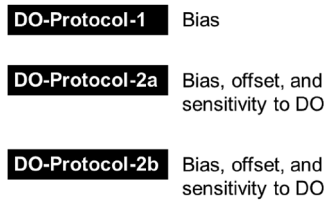
- Kvantifiera avdriften (dvs. drift av bias, se Kapitel 3)

Utrustning

- Syregivare som ska valideras
- Syregivare som används för referensmätning, det vill säga en syregivare som tidigare kalibrerats och validerats mot standardförhållanden (en noggrann kalibrering med DO-rutin 2)
- Hink (cirka 20 L)
- Omrörare

Förbrukningsvaror

- 10 L avloppsvatten som samlats intill givaren som ska valideras



Figur 5.A.1 Valideringsförfaranden för syregivare och resulterande metadata för givarens datakvalitet.

Utförande

(1) Förberedelser (1–24 timmar före validering)
  (a) Rengör den syregivaren som ska valideras.
  (b) Validera den handhållna referens-givaren med DO-rutin-2a eller DO-rutin-2b.
(2) Validering
  (a) Ta ett vattenprov (ca 10 liter) intill givaren med hinken.
  (b) Blanda vattenprovet försiktigt med omröraren (automatiskt eller manuellt) för att få en homogen blandning och undvik sedimentering av suspenderade fasta ämnen.
  (c) Inspektera syregivaren visuellt och kontrollera att den är hel och ren.
  (d) Placera både syregivaren och den handhållna DO-referensgivaren i hinken.
  (e) Se till att temperaturvärdena för båda givarna stabiliseras på samma värde (normalt 1–5 min).
  (f) Notera de stabila temperaturvärdena från båda givarna.
  (g) Se till att avloppsvattnet rörs om kontinuerligt.
  (h) Notera de stabiliserade syrehalterna på båda givarna.
  (i) Flytta tillbaka syregivaren till dess ursprungliga plats.
(3) Återställning
  (a) Häll tillbaka vattenprovet (eller återanvänd till nästa syregivare).
  (b) Rengör hink och omrörare.
  (c) Ställ tillbaka hinken och referenssyregivaren på sina normala platser.

Dataanalys och lagring

(1) Spara värdena för temperatur, syrehalt hos givaren och den handhållna referensgivaren i en databas.
(2) Skillnaden mellan syregivaren och referensgivaren är en uppskattning av givarens absoluta mätfel vid den aktuella driftsförhållanden och kan användas för att avgöra om validering med DO-rutin-2a eller DO-rutin-2b eller andra åtgärder är nödvändiga.

Övrigt

Det är tidsbesparande att återanvända vattenprovet för att validera flera syregivare vid ett tillfälle. För givare två (och efterföljande givare som valideras) så behöver enbart steg 1a och 2c–2i upprepas.

5.A.1.2 DO-rutin-2a: Tvåpunktsvalidering med vattenmättad luft

Version och datum

- Version: 1.0
- Senast reviderad: 31 juli 2023
- Nästa granskning och revidering ska vara genomförd senast den 31 juli 2025

Syfte

Denna rutin används för att validera syregivare vid två referensförhållanden, noll syrehalt och mättad (100 %) syrehalt. Referensförhållandena, det vill säga de två olika syrehalterna, kan produceras utifrån kunskap om hur syrgas löser sig i luft och vatten. För att producera mättade förhållande med 100 % syrehalt kan antingen luft (rutin-2a) eller i vatten (rutin-2b) användas. Denna rutin beskriver rutin-2a med syremättad luft.

Med denna rutin går det att:

- Kvantifiera bias, offset och lutning, beräknat utifrån två syrehalter.

Om rutinen upprepas regelbundet över tid går det även att:

- Kvantifiera avdrift (av bias, offset och lutning)

Utrustning

- Syregivare som ska valideras
- Noggrann temperaturgivare (± 0,5°C)
- Barometrisk tryckgivare (behövs enbart vid ett tillfälle)
- Konduktivitetsgivare (kan behövas om inflödet påverkas av havsvatten och stora variationer i salthalt)
- 20 L hink

Noll syrehalt i vatten (validering vid syrehalten 0 mg/L)

- 3 liters hink med kranvatten
- 10 g natriumsulfit eller motsvarande starkt reduktionsmedel
- Mättad syrehalt i luft (validering vid syrehalten 9 mg/L och 20°C)
- 1 L bägare med lock och liknande diameter som syregivaren
- Remmar för att fästa givaren i 1 L bägaren

Förbrukningsvaror

- Pappershanddukar (eller motsvarande)
- Rent kranvatten

Utförande

(1) Förberedelser (1–24 timmar före validering)
  (a) Rengör syregivaren och sätt tillbaka den i sin ursprungliga position.
  (b) Fyll 20 L hinken med 10–15 L kranvatten och placera den på platsen för valideringen. Låt kranvattnet uppnå samma temperatur som den omgivande luften.
  (c) Tillsätt 200 ml kranvatten i botten av en 1 L bägare och stäng locket. Förvara bägaren vid samma temperatur som den omgivande luften (t.ex. i samma rum eller utomhus). På så sätt säkerställs att luften är mättad vid den omgivande luftens temperatur. Skydda givaren från direkt solljus om den används utomhus.
  (d) Identifiera höjden (meter över havet) för den plats där valideringen utförs, alternativt mät barometertrycket.
  (e) Mät salthalten eller konduktiviteten på den plats där givaren är placerad (framförallt viktigt om det finns saltvatteninläckage i ledningsnätet)

(2) Validering
  (a) Skölj syregivaren med kranvatten.
  (b) Låt givaren stabilisera sig till den aktuella rumstemperaturen (utomhustemperaturen) genom att sänka ned den rena givaren 20 L hinken med rent vatten vid omgivande lufttemperatur.
  (c) Mät temperaturen i hinken med rent vatten och jämför med den temperatur som registreras av syregivaren.
  (d) Registrera båda temperaturvärdena.
  (e) Noll syrehalt i vatten
    (i) Placera den rena givaren i en 3 L hink med rent vatten och en omrörare.
    (ii) Tillsätt successivt natriumhydrosulfit (eller det valda reduktionsmedlet) tills den uppmätta syrehalten uppnår ett stabilt lågt värde nära noll.
    (iii) Registrera det uppmätta värdet i närheten av 0 mg/L.
    (iv) Registrera givarvärdet och den aktuella temperaturen som indikeras av givaren och den externa termometern. Kontrollera att temperaturerna är lika.
    (v) Jämför det registrerade värdet på syrehalten med det teoretiska värdet i tabellen över syremättnad (Tabell 5.A.1) efter korrigering för barometertryck och temperatur.

Tabell 5.A.1 Korrektionsfaktorer för koncentrationen av löst syre (DO) vid 100 % mättnad och 1 atm vid olika temperaturer.

Koncentrationer av upplöst syre vid mättnad						
	Konduktivitet (μS/cm)					
Temperatur (°C)	0	2000	4000	6000	8000	10 000
10	11,28	11,20	11,12	11,04	10,95	10,87
11	11,02	10,94	10,87	10,79	10,70	10,62
12	10,77	10,69	10,62	10,54	10,47	10,39
13	10,53	10,46	10,38	10,31	10,24	10,16
14	10,29	10,22	10,15	10,07	10,00	9,93
15	10,07	10,00	9,93	9,86	9,79	9,72
16	9,86	9,79	9,72	9,65	9,58	9,52
17	9,65	9,58	9,51	9,46	9,39	9,32
18	9,45	9,38	9,33	9,26	9,19	9,13
19	9,26	9,20	9,14	9,07	9,01	8,95
20	9,08	9,02	8,96	8,90	8,83	8,77
21	8,90	8,84	8,78	8,72	8,66	8,61
22	8,73	8,67	8,62	8,56	8,50	8,44
23	8,56	8,51	8,45	8,39	8,34	8,28
24	8,40	8,35	8,29	8,24	8,18	8,12
25	8,24	8,19	8,13	8,08	8,03	7,98

(f) Mättad syrehalt i luft
  (vi) Torka av den rena givaren med en torr pappershandduk inklusive eventuella vattendroppar på givarmembranet/locket.
  (vii) Ta bort locket från 1 L bägaren och placera givaren genom att fästa den i ett stabilt läge med remmar (eller liknande) i upprätt läge. Se till att givaren befinner sig strax ovanför vattenytan och inte kommer i kontakt med vatten.
  (viii) Låt givarsignalen stabiliseras (10–30 min). Se till att temperaturen är konstant och lika för givarkroppen som för luften under valideringen. Direkt solljus eller stark kall vind bör undvikas (i värsta fall kan det krävas validering inomhus).
  (ix) Registrera givarvärdet och den aktuella temperaturen som indikeras av givaren och lufttemperaturen av den externa termometern. Kontrollera att temperaturerna är lika.
(g) Återställ syregivaren till dess normala position.

(3) Återställning
  (a) Häll tillbaka vattenprovet (alternativt fortsätt med nästa syregivare).
  (b) Rengör hinken.
  (c) Ställ tillbaka hinken och övrig utrustning på sina normala platser.

Dataanalys och lagring

(1) Definiera tryckkorrigeringsfaktor från barometrisk mätning eller platsens höjd.
(2) Validera syregivarens temperaturkorrigering genom att beräkna skillnaden mellan den registrerade externa temperaturen och den som registreras av givaren.
(3) Kontrollera att en korrekt salthaltskorrigering tillämpas för mätningarna i processen.
(4) Beräkna skillnaden mellan syregivarens värde och det teoretiska värdet under referensförhållandena, justerat för temperatur, tryck och salthalt.

(5) Registrera alla värden.
(6) Validera syregivaren genom att uppskatta dess lutning och offset med hjälp av linjär regression för noll och mättad syrehalt.
(7) Registrera den beräknade lutningen och förskjutningen.

Övrigt

Det är tidsbesparande att återanvända vattenprovet för att validera flera syregivare vid ett tillfälle. För givare två (och efterföljande givare som valideras) så behöver enbart steg 1a och 2a–2 g upprepas.

5.A.1.3 DO-Procedur-2b: Tvåpunktsvalidering med syremättat vatten

Version och datum

- Version: 1.0
- Senaste översyn och revidering: 31 juli 2023
- Nästa granskning och revidering ska vara genomförd senast den 31 juli 2025

Syfte

Denna rutin används för att validera syregivare vid två referensförhållanden, noll syrehalt och mättad (100 %) syrehalt. Referensförhållandena, det vill säga de två olika syrehalterna, kan produceras utifrån kunskap om hur syrgas löser sig i luft och vatten. För att producera mättade förhållande med 100 % syrehalt kan antingen luft (rutin-2a) eller i vatten (rutin-2b) användas. Denna rutin beskriver rutin-2b med syremättat vatten.

Med denna rutin går det att:

- Kvantifiera bias, offset och lutning, beräknat utifrån två syrehalter.

Om rutinen upprepas regelbundet över tid går det även att:

- Kvantifiera avdrift (av bias, offset och lutning)

Utrustning

- Syregivare som ska valideras
- Noggrann temperaturgivare (± 0,5°C)
- Barometrisk tryckgivare (behövs enbart vid ett tillfälle)
- Konduktivitetsgivare (kan behövas om inflödet påverkas av havsvatten och stora variationer i salthalt)
- Luftkompressor eller tillgång till tryckluft
- Noll syrehalt i reducerat vatten (validering vid 0,0 mg DO/L)
  - 3 litershink med kranvatten
  - 10 g natriumsulfit eller motsvarande starkt reduktionsmedel
- Syremättat vatten (validering vid 9 mg DO/L och 20°C)
  - Luftmunstycke eller liknande för att lufta en 5 L hink med tryckluft.
  - Omrörare

Förbrukningsvaror

- Pappershanddukar (eller motsvarande)
- Kranvatten

Utförande

(1) Förberedelser (1–24 timmar före validering)
  (a) Rengör syregivaren och sätt tillbaka den i dess ursprungliga position igen.

(b) Fyll 3 L och 5 L hinkar med rent vatten och placera dem på platsen för valideringen. Låt det rena vattnet få samma temperatur som den omgivande luften.
(c) Identifiera höjden (meter över havet) för den plats där valideringen utförs, alternativt mät barometertrycket.
(d) Mät konduktiviteten på den plats där givaren är placerad.

(2) Validering
(a) Skölj syregivaren med rent vatten
(b) Låt syregivaren stabilisera sig vid aktuell rumstemperatur (utomhustemperatur) genom att sänka ner den i 5 L-hinken med rent vatten och samma temperatur som omgivningen.
(c) Mät temperaturen i hinken med rent vatten och jämför med den temperatur som registreras av syregivaren.
(d) Registrera båda temperaturvärdena.
(e) Noll syrehalt i vatten
(i) Placera den rena givaren i 3 L-hinken med kranvatten och en omrörare.
(ii) Tillsätt successivt natriumsulfit (eller det valda reduktionsmedlet) tills den uppmätta syrehalten uppnår ett stabilt lågt värde nära noll.
(iii) Registrera det uppmätta värdet kring 0 mg/L.
(iv) Jämför det registrerade värdet med det teoretiska värdet i tabellen över syremättnad (Tabell 5.A.1) och korrigera för barometertryck och temperatur.
(f) Syrehalt i mättat vatten
(i) Identifiera den syrehalt som fås vid 100 % mättnad vid aktuell temperatur (steg 2c) och tryck (Tabell 5.A.1).
(ii) Skölj rent syregivaren med kranvatten
(iii) Låt syregivaren stabilisera sig till aktuell rumstemperatur (utomhustemperatur) genom att sänka ner den i 5 L-hinken.
(iv) Lufta successivt vattnet i 5-litershinken med hjälp av tryckluft tills mätningen stabiliseras vid eller över den identifierade gränsen för 100 % syremättnad - sluta lufta.
(v) Låt syrehaltsmätningen stabilisera sig (3–10 min) under omrörning.
(vi) Registrera syrehalten vid 100 % mättnad och temperaturen.
(g) Återställ syregivaren till dess normala mätposition.

(3) Återställning
(a) Häll tillbaka vattenprovet (eller återanvänd till nästa syregivare).
(b) Rengör hink och omrörare.
(c) Ställ tillbaka hinken och referenssyregivaren på sina normala platser

Dataanalys och lagring

(1) Definiera tryckkorrigeringsfaktorn från barometrisk mätning eller platsens höjd.
(2) Validera syregivarens temperaturkorrigering genom att beräkna skillnaden mellan den registrerade externa temperaturen och den som registreras av givaren.
(3) Kontrollera att salthaltskorrigering är korrekt för förhållandena i processen.
(4) Beräkna skillnaden mellan syregivaren och det teoretiska värdet under referensförhållandena, justerat för temperatur, tryck och salthalt.
(5) Registrera alla värden, både uppmätta och teoretiska, i en databas.
(6) Validera syregivaren genom att uppskatta dess lutning och offset med hjälp av linjär regression för noll och mättade syrehalter.
(7) Registrera den beräknade lutningen och förskjutningen.

Övrigt

Det är tidsbesparande att återanvända vattenprovet för att validera flera syregivare vid ett tillfälle. För givare två (och efterföljande givare som valideras) så behöver enbart steg 1a och 2a–2 g upprepas.

5.A.1.4 Beskrivande information för validering av syregivare

Detta avsnitt ger exempel på information som kan ingå som beskrivande information i en valideringsrutin för syregivare.

Syfte med syrehaltsmätningen

Givaren används för att mäta och kontrollera syrehalten i aktivslamprocessen. Samtidigt kan temperaturen registreras, eftersom denna sekundära mätning ingår för att kompensera för temperaturens inverkan på syrehalten.

Mätprinciper

Det finns två vanliga typer av syregivare: elektrokemiska och fluorescerande. Båda typerna mäter syrgasens partialtryck, som omvandlas till en syrehalt:

- *Elektrokemisk (polarografisk):* Elektrokemiska syregivare består av ett gasgenomsläppligt membran, en elektrolyt, en anod och en katod. När syremolekyler diffunderar genom membranet oxideras anoden av syret, vilket skapar en ström som är proportionell mot syrets partialtryck. En initial spänning måste appliceras för att övervinna aktiveringsenergin mellan anoden och katoden, det vill säga polarisationsenergin.
- *Optisk (fluorescerande):* Fluorescerande syregivare använder fluorescens för att mäta syrgasens partialtryck. En fluorescensgivare består av ett fluorescerande färgämne som täcker givarspetsen, en ljuskälla med specifika våglängder (vanligtvis LED-ljus) och en fotodetektor för registrering av det emitterade fluorescenta ljuset. Färgämnet absorberas (exciteras) av en ljuskälla och avger därefter ljus med en längre våglängd (lägre energi) som fluorescerande ljus. När syre är närvarande vid det fluorescerande färgämnet fördröjs det återutsända fluorescensljuset i proportion till närvaron av syremolekyler på givaren.

Mätstörande faktorer

Syrehalten mäts genom partialtrycket av syrgas i lösningen. Därför kommer faktorer som påverkar syrgasens partialtryck också att påverka syrehaltsmätningen. Följande faktorer kan påverka mätningen:

- *Mätvinkel:*
  - Givarens vinkel i förhållande till vattenytan påverkar mätningen. Exempelvis kan luftbubblor eller fasta partiklar ge upphv till avvikande mätvärden.
  - *Korrigerande motåtgärder:* Om inte annat anges av tillverkaren bör givaren placeras med membranet eller den optiska delen nedåt och med en vinkel på 45° mot vattenytan.
- *Luftbubblor:*
  - Luftbubblor kan ge en oregelbunden och förhöjd syrehalt.
  - *Korrigerande motåtgärder:* Justera givaren till en skarpare kontaktvinkel med vattenytan eller placera givaren på en annan plats.
- *Barometriskt tryck:*
  - Atmosfärstrycket har en proportionell inverkan på syrgasens partialtryck. Väderrelaterade tryckförändringar är i storleksordningen 950–1050 hPa, vilket har en mindre inverkan (<5 %) på avläsningen. Under givarkalibreringen kan dock barometertrycket ha en stor inverkan och bör kompenseras, särskilt om reningsverket ligger högre än havsytan.
  - *Korrigerande motåtgärder:* Korrigera för barometertryck under kalibreringen. Till exempel motsvarar en anläggning på 1000 m över havet 899 hPa, jämfört med 1013 hPa vid havsnivå, och behöver därför korrigeras med 11 %.
- *Biofilmspåväxt:*
  - Syregivarens kontamineras av biofilmspåväxt vilket förhindrar syrgasdiffusion i membranet och ökar svarstiden.
  - *Korrigerande motåtgärder:* Rengör givaren och se över bästa rengöringsmetod.

- *Salthalt:*
  - Salthalten minskar syrets löslighet och därmed syrets partialtryck. Avloppsnät som påverkas av havsvatten eller bräddavlopp kan påverka salthalten och i sin tur syrehaltsmätningen.
  - *Korrigerande motåtgärder:* Justera för vanlig konduktivitet under kalibreringen och övervaka stora konduktivitetsförändringar i inflödet.
- *Polarisationsström* (gäller elektrokemiska givare):
  - En initial ström behövs för polarisering av givaren (dvs. för att initiera den ström som drivs av de reducerade syremolekylerna). Av denna anledning måste en elektrokemisk syregivare vara i drift med ström i minst 30 minuter innan den ger tillförlitliga data.
  - *Korrigerande motåtgärder:* Se till att (a) signalerna bedöms som opålitliga i 30 minuter efter påslagning eller (b) signaldata endast lagras när givaren har varit påslagen i minst 30 minuter.
- *Temperatur:*
  - Lösligheten av syre i vatten är starkt temperaturberoende. Av denna anledning korrigeras syrehaltsmätningen automatiskt av en intern temperaturgivare. Om givarkroppen har en annan temperatur än det omgivande vattnet (eller luften), eller om temperaturgivaren är felaktig, kommer syrehaltsmätningen i sin tur att bli felaktig (se Tabell 5.A.1).
  - *Korrigerande motåtgärder:* Se till att givarens temperatur stabiliseras före mätningarna. Validera den interna temperaturgivaren.
- *Direkt solljus* (gäller fluorescerande givare):
  - Direkt solljus kan påverka såväl mätningen som givarens temperatur.
  - *Korrigerande motåtgärder:* Skydda givaren från solljus under valideringen.

Formel för tryckkorrigering

$$\text{Tryckkorrigeringsfaktor} = p_{\text{Val}}/p_{\text{Atm}}$$

där p_{Val} är barometertrycket på platsen för valideringen och p_{Atm} är atmosfärstrycket vid havsnivån, dvs. 1 atm.

Tabell för korrigering av temperatur och konduktivitet på syrets löslighet.
Se Tabell 5.A.1.

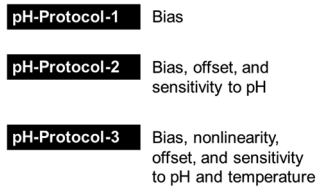
5.A.2 Procedurer för validering av pH-givare

Detta avsnitt beskriver tre sätt att validera pH-givare. Figur 5.A.2 visar vilka olika datakvalitetsmått som kan uppskattas med respektive valideringsrutin.

5.A.2.1 pH-rutin-1: Enpunktsvalidering av pH-givare vid aktuell driftspunkt

Version och datum

- Version: 1.0
- Senaste översyn och revidering: 31 juli 2023
- Nästa granskning och revidering ska vara genomförd senast den 31 juli 2025



Figur 5.A.2 Valideringsförfaranden för pH-givare och resulterande metadata för givarens datakvalitet.

Syfte

Syftet med denna är att kontrollera pH-värdet hos en pH-givare i dess normala miljö med ett referens-pH-värde som produceras av en handhållen pH-givare som referensmätning.

Med denna rutin går det att:

- Utvärdera mätfelet vid den aktuella driftspunkten.
- Kvantifiera bias (se Kapitel 3)

Om rutinen upprepas regelbundet över tid går det även att:

- Kvantifiera avdriften (dvs. drift av bias, se Kapitel 3)

Denna rutin utvärderar inte:

- Olinjäriteter, offset, lutning (känslighet för pH och/eller temperatur), givardynamik eller varians (se Kapitel 3)

Utrustning

- pH-givare som ska valideras (pH^{test}), ansluten till display, datalogger eller transformator
- Referensgivare pH-givare (pH^{ref}), rengjord och verifierad på laboratorium med pH-rutin-2 eller pH-rutin-3)
- Två 4L-hinkar (eller större)
- Mjuk trasa
- Omrörare

Förbrukningsvaror

- 4 L avloppsvatten, uppsamlat vid givarens position
- 4 L kranvatten

Utförande

(1) Beredning av provlösning:
  (a) Samla in ett prov med avloppsvatten med en av 4 L-hinkarna
  (b) Starta omröraren och se till att vattenprovet blandas till en homogen lösning utan sedimentering av suspenderat material.
(2) Rengör givaren:
  (a) Fyll den andra 4 L-hinken med kranvatten
  (b) Rengör pH-givaren som ska valideras med en trasa och kranvatten
(3) Placera test och referensgivaren i hinken med avloppsvatten
(4) Kryssa i den enhet som används vid mätningen med ett X:
  (a) Temperatur: [__] °C [__] °F [__] K
  (b) Elektrisk potential: [__] mV
(5) Se till att temperaturvärdena för båda givarna stabiliseras.
(6) Registrera värden:
  (a) Registrera temperaturvärdet med båda givarna
    (i) T^{test}:_____
    (ii) T^{ref}:_____
  (b) Registrera värdet för elektrisk potential med båda givarna
    (i) P^{test}:_____
    (ii) P^{ref}:_____
  (c) Registrera pH-värdena hos båda givarna (med befintlig kalibrering)
    (i) pH^{-test}:_____
    (ii) pH^{ref}:_____

(7) Häll tillbaka avloppsvattnet
(8) Flytta tillbaka pH-givaren till dess ursprungliga plats
(9) Rengör hinkar och omrörare
(10) Ställ tillbaka hinkar och referens-pH-givaren på sin normala plats

Dataanalys och lagring

(1) Spara värdena T^{test}, T^{ref}, P^{test}, P^{ref}, pH^{test}, pH^{ref} i ett kalkylblad eller en databas.
(2) Kontrollera att T^{test} och T^{ref} är likvärdiga och inom en acceptabel felmarginal (*t*.ex. 0,5 K). Om inte, upprepa försöket när alla lösningar har uppnått stabil temperatur. Om upprepade försök misslyckas, validera temperaturgivarna och vidta åtgärder för att säkerställa godtagbar temperaturmätning.
(3) Differensen mellan pH-mätningarna (pH^{test} - pH^{ref}) är en uppskattning av givarens absoluta mätfel under aktuella driftsförhållanden och kan användas för att avgöra om en mer omfattande validering är nödvändig (t.ex. med pH-rutin-2 eller pH-rutin-3).
(4) Differensen i potentialen (P^{test}, P^{ref}) är värdefull för att studera långtidseffekter på pH-givarens slitage, särskilt eftersom denna skillnad inte påverkas av givarens kalibrering.

5.A.2.2 pH-rutin-2: Tvåpunktsvalidering av pH-givare i buffert

Version och datum

- Version: 1.0
- Senaste översyn och revidering: 31 juli 2023
- Nästa granskning och revidering ska vara genomförd senast den 31 juli 2025

Syfte

Syftet med denna rutin är att registrera råsignalen från jonselektiva pH-givare under kontrollerade förhållanden. I detta fall innebär kontrollerade förhållanden att:

- Stabila och buffertlösningar med specifika pH-värden används
- Buffertlösningarnas temperaturer är stabila och mäts med samma pH-givare

Med denna rutin går det att:

- Utvärdera mätfelet vid två kontrollerade pH-värden
- Kvantifiera bias (se Kapitel 3), offset och lutning

Om rutinen upprepas regelbundet över tid går det även att:

- Kvantifiera avdriften (av bias, offset och lutning)

Denna rutin utvärderar inte:

- Olinjäriteter, temperaturkänslighet, givardynamik eller varians (se Kapitel 3)

Utrustning

- pH-givare som ska valideras, ansluten till display, datalogger eller transformator
- Tre 200 mL glasbägare
- Tre magnetomrörare med magnet
- Hinkar minst 5 L
- Mjuk trasa för rengöring av givaren

Förbrukningsvaror

- 5 L kranvatten
- Buffertlösningar med stabil temperatur, 100 mL vardera

- Kranvatten [W]
- Kalibreringslösning vid pH 4 [C4].
- Kalibreringslösning vid pH 7 [C7]

Utförande

(1) Beredning av lösningar:
  (a) Märk bägarna med W, C4 och C7
  (b) Placera varje bägare på varsin magnetomrörare
  (c) Lägg en magnet i varje bägare
  (d) Häll varje lösning (W, C4, C7) i bägaren med matchande etikett (W, C4, C7)
  (e) Starta omrörarna
(2) Ta bort givaren från dess normala position
(3) Rengör givaren:
  (a) Fyll en av hinkarna med kranvatten
  (b) Rengör givaren med vatten och trasa
(4) Placera givaren i bägaren märkt W
(5) Vänta tills givartemperaturen stabiliseras
(6) Se till att alla lösningar stabiliseras vid rumstemperatur
(7) Identifiera enheterna med ett X:
  (a) Temperatur: [__] °C [__] °F [__] K
  (b) Elektrisk potential: [__] mV [__] __
(8) Värde 0:
  (a) Registrera temperaturvärdet *T0*:_____
  (b) Registrera värdet för elektrisk potential P0:_____
(9) Stegförändring 1:
  (a) Flytta givaren från bägare W till bägare C4
  (b) Vänta 5 min
(10) Värde 1:
  (a) Registrera temperaturvärdet T1:_____
  (b) Registrera värdet för elektrisk potential P1:_____
(11) Stegförändring 2:
  (a) Flytta givaren från bägare C4 till bägare C7
  (b) Vänta 5 min
(12) Värde 2:
  (a) Registrera temperaturvärdet T2:_____
  (b) Registrera värdet för elektrisk potential P2:_____
(13) Stegförändring 3:
  (a) Flytta givaren från bägare C7 till bägare C4
  (b) Vänta 5 min
(14) Värdea 3:
  (a) Registrera temperaturvärdet T3:_____
  (b) Registrera det elektriska potentialvärdet P3:_____
(15) Stegförändring 4:
  (a) Flytta givaren från bägare C4 till bägare W
  (b) Vänta 5 min
(16) Värde 4:
  (a) Registrera temperaturvärdet T4:_____
  (b) Registrera det elektriska potentialvärdet P4:_____
(17) Rengör givaren med vatten och trasa
(18) Återställ givaren till dess normala position

(19) Återvinn/släng alla lösningar
(20) Rengör glasvaror, omrörare och magneter och lägg tillbaka dem i förrådet

Dataanalys och lagring

(1) Lagra värdena T0, T1, T2, T3, T4, P0, P1, P2, P3, P4 i ett kalkylblad eller en databas
(2) Kontrollera att T1, T2, T3 är likvärdiga och acceptabel felmarginal (t.ex. 0,5 K). Om inte, upprepa försöket när alla lösningar har uppnått stabila temperaturer.
(3) Värdet P2 är en uppskattning av givarens offset
(4) Två uppskattningar av givarens lutning vid standardtemperatur (T_s = 298,15 K) kan beräknas med hjälp av följande steg:
   (a) Sortera värdena T1, T2 och T3 i stigande ordning och välj T_a som det mittersta värdet (mediantemperaturen). Se till att T_a uttrycks i Kelvin (K)
   (b) Beräkna den första uppskattningen av givarens lutning som

$$S1 = \frac{P1 - P2}{3} \times \frac{T_s}{T_a}$$

   (c) Beräkna den andra uppskattningen av givarens lutning som

$$S2 = \frac{P3 - P2}{3} \times \frac{T_s}{T_a}$$

5.A.2.3 pH-rutin-3: Flerpunktsvalidering av pH-givare i buffertlösningar

Version och datum

- Version: 1.0
- Senaste översyn och revidering: 31 juli 2023
- Nästa granskning och revidering ska vara genomförd senast den 31 juli 2025

Syfte

Syftet med denna rutin är att registrera råsignalen från jonselektiva pH-givare under kontrollerade förhållanden. I detta fall innebär kontrollerade förhållanden att:

- Stabila och buffertlösningar med specifika pH-värden används
- Buffertlösningarnas temperaturer är stabila och mäts med samma pH-givare

Med denna rutin går det att:

- Utvärdera mätfelet vid tre kontrollerade pH-värden
- Kvantifiera bias (se Kapitel 3), icke-linjäritet, offset och lutningar (känslighet för pH och/eller temperatur)

Om rutinen upprepas regelbundet över tid går det även att:

- Kvantifiera avdrift (av bias, icke-linjäritet, offset och lutningar, se Kapitel 3)

Denna rutin utvärderar inte:

- Förändringar i givardynamik eller varians (se Kapitel 3)

Utrustning

(1) pH-givare som ska valideras, ansluten till display, datalogger eller transformator
(2) 7 st temperaturkontrollerade bägare på 200 mL

(3) 7 st magnetomrörare med magnet

Förbrukningsvaror

(1) 4 L kranvatten
(2) Sju lösningar med stabil temperatur, 100 mL vardera
  (a) Kranvatten vid standardtemperatur ($T_s = 298{,}15$ K) [W]
  (b) Kalibreringslösning vid pH 4, 7, 10 vid standardtemperatur ($T_s = 298{,}15$ K) [C4s, C7s, C10s].
  (c) Kalibreringslösning vid pH 4, 7, 10 vid givarens normala temperatur ($T_a = 288{,}15$ K) [C4a, C7a, C10a].

Utförande

(1) Beredning av buffertlösningar:
  (a) Märk bägarna med W, C4s, C7s, C10s, C4a, C7a och C10a
  (b) Placera varje bägare på en magnetomrörare
  (c) Lägg en magnet i varje bägare
  (d) Häll i buffertlösning i respektive bägare (W, C4s, C7s, C10s, C4a, C7a och C10a) med matchande etikett
  (e) Starta omröraren och kontrollera att lösningarna håller rumstemperatur
(2) Ta bort givaren från dess normala mätposition
(3) Rengör givaren:
  (a) Fyll en av hinkarna med kranvatten
  (b) Rengör givaren med vatten och trasa
(4) Placera givaren i bägare W och vänta tills givarens temperaturen stabiliserats
(5) Se till att alla lösningar har uppnått rumstemperatur
(6) Identifiera enheten för temperatur och elektrisk potential och registrera dem i Tabell 5.A.2.
(7) Värde 0:
  (a) Registrera temperaturvärdet T0:_____
  (b) Registrera värdet för elektrisk potential P0:_____
(8) Upprepa detta tills alla kalibreringslösningar har analyserats:
  (a) Stegförändring: Flytta givaren till nästa lösning i ordningen: W, C4s, C4a, C7a, C7s, C10s, C10a, W
  (b) Vänta i 5 minuter
  (c) Registrera temperaturvärdet T i Tabell 5.1

Tabell 5.A.2 Registrerade värden.

Temperatur	Potentiell
Mätningsenhet	
W	
C4	
C4a	
C7a	
C7s	
C10s	
C10a	
W	

(d) Registrera värdet för elektrisk potential P i Tabell 5.1
(9) Rengör givaren med vatten och trasa
(10) Återställ givaren till dess normala mätposition
(11) Återvinn/släng alla lösningar
(12) Rengör glasvaror, omrörare och magneter och lägg tillbaka dem i förrådet

Dataanalys och lagring

(1) Spara värdena i Tabell 5.A.2 i ett kalkylblad eller en databas.
(2) Kontrollera att alla temperaturer för medierna W, C4s, C7s och C10s är likvärdiga och inom en acceptabel felmarginal (t.ex. 0,5 K). Kontrollera att alla temperaturer för medierna C4a, C7a och C10a är likvärdiga och inom en acceptabel felmarginal. Om inte, upprepa försöket när alla lösningar har uppnått stabila temperaturer.
(3) Utför en linjär regression med temperaturvärdena och det kända pH-värdet. Använd alla data i Tabell 5.A.2 som motsvarar medierna C4s, C4a, C7a, C7s, C10s och C10a. Använd dock inte mätvärdena från vattenprovet W. Analysera värden för offset- och lutning (för temperatur och pH) och jämför med (a) tidigare erhållna värden (om sådana finns) och (b) teoretiska värden (Nernstekvationen). Bestäm om regressionens kvalitet är godtagbar genom att kontrollera residualerna som en funktion av den uppmätta potentialen, den uppmätta temperaturen och kända pH-värden. Använd denna rutin för att utvärdera om givarens beteende är tillräckligt linjärt för att vara godtagbart. Om det anses acceptabelt, använd de beräknade regressionskoefficienterna för att bestämma och registrera givarens bias, offset och lutning.

5.A.3 Valideringsrutiner för ammoniumgivare

Detta avsnitt beskriver två valideringsrutiner för ammoniumgivare (‘total ammonium nitrogen – TAN’). Figur 5.A.3 visar de datakvalitetsmått som kan uppskattas med dessa procedurer.

5.A.3.1 TAN-rutin-1: Enpunktsvalidering av ammoniumgivare vid aktuell driftspunkt

Version och datum

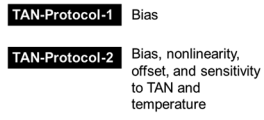
- Version: 1.0
- Senaste översyn och revidering: 31 juli 2023
- Nästa granskning och revidering ska vara genomförd senast den 31 juli 2025

Syfte

Syftet med denna rutin är att kontrollera värdet på den uppmätta ammoniumhalten hos en ammoniumgivare i dess normala miljö med ett referensvärde som produceras genom en analytisk laboratoriemätning. Rutinen har utformats för jonselektiva givare (‘ion-selective electrode – ISE’). Rutinen kan justeras till analysatorer med kolorimetrisk mätprincip, vilket beskrivs under steg 10 nedan.

Med denna rutin går det att:

- Utvärdera mätfelet vid den aktuella driftspunkten.
- Kvantifiera bias (se Kapitel 3).



Figur 5.A.3 Valideringsförfaranden för ammoniumgivare och resulterande metadata för givarens datakvalitet.

Om rutinen upprepas regelbundet över tid går det även att:

- Kvantifiera avdriften (dvs. drift av bias, se Kapitel 3)

Utrustning

- Ammoniumgivare som ska valideras
- 20 L hink
- Omrörare
- Provtagningsflaska för vattenprov av ammonium till laboratorium

Förbrukningsvaror

- 10 L avloppsvatten uppsamlat vid givarens placering
- 100 mg allyltiourea

Utförande

(1) Förberedelser (1–24 timmar före valideringen)
  (a) Rengör ammoniumgivaren som ska valideras
(2) Validering
  (a) Ta ett vattenprov på 10 liter med hinken.
  (b) Blanda vattenprovet försiktigt med omröraren till en homogen blandning utan sedimentering av suspenderade fasta ämnen.
  (c) Inspektera givaren visuellt och kontrollera att den är hel och ren.
  (d) Tillsätt allyltiourea till vattenprovet för att hämma ammoniumoxidationen till nitrit och nitrat (10 mg N/L är fullt tillräckligt och långt över hämningsnivån).
  (e) Placera ammoniumgivaren i hinken.
  (f) Se till att avloppsvattnet rörs om kontinuerligt.
  (g) Låt givarens pH-värde och temperaturvärde stabiliseras (normalt 5–10 min).
  (h) Registrera det stabiliserade pH-värdet från ammoniumgivaren.
  (i) Registrera det stabiliserade temperaturvärdet från ammoniumgivaren.
  (j) Registrera det stabiliserade ammoniumvärdet från ammoniakgivaren.
  (k) Ta ett 20 mL vattenprov från hinken för laboratorieanalys.
  (l) Återställ ammoniumgivaren till dess normala position.
(3) Återställning
  (a) Häll tillbaka vattenprovet.
  (b) Rengör hink och omrörare.
  (c) Ställ tillbaka hink och omrörare på sina normala platser.
  (d) Skicka vattenprovet (20 mL) till laboratoriet för analys med en standardmätning eller en accepterad analysmetod för referensmätning.

Dataanalys och lagring

(4) Lagra värdena för pH, temperatur och ammoniumhalt från givaren i en databas.
(5) Lagra även laboratoriets mätvärde på ammoniumhalten i samma databas.
(6) Skillnaden mellan ammoniumgivarens mätvärde och den analytiska mätningen är en uppskattning av givarens absoluta mätfel vid de aktuella driftsförhållandena och kan användas för att avgöra om ytterligare åtgärder är krävs.

Övrigt

(7) Samma vattenprov går med fördel att återanvända för att validera flera ammoniumgivaren. Då upprepas steg 1a och 2c-2l för varje ny ammoniumgivare.

(8) Automatisk insamling av ammoniumgivarens mätvärden via ett automatiserat datainsamlingssystem underlättar kvantifiering av givarens precision (se Kapitel 3).
(9) Om ammoniumgivaren har ytterligare elektroder (t.ex. K^+) för att kompensera för mätstörande effekter, lägg då till ett steg efter steg 2j för att även mäta det stabiliserade värdet som erhålls med dessa elektroder. Det kan vara av intresse att då även modifiera steg 3d så att jonkoncentrationen av den komponent som stör mätningen också kvantifieras.
(10) För att modifiera denna rutin till att en ammoniumanalysator (kolorimetrisk mätprincip), ändra steg 1a till 2k enligt följande:
    (a) Justera rengöringsproceduren enligt analysatorns manual.
    (b) I stället för att ändra givarens position kan du ändra provtagningsledningens position så att analysatorn samlar upp vatten från hinken under valideringen.
    (c) Justera väntetiderna för att ta hänsyn till att analysatorn mäter koncentrationen med fördröjning (på grund av den tid som krävs för kolorimetrisk reaktion) och med lägre frekvens (t.ex. 10-minutersintervall).

5.A.3.2 TAN-Procedure-2: procedur för validering av flerpunktsgivare för ammoniak vid aktuell driftspunkt

Version och datum

- Version: 1.0
- Senaste översyn och revidering: 31 juli 2023
- Nästa granskning och revidering ska vara genomförd senast den 31 juli 2025

Syfte

Syftet med denna rutin är att kontrollera värdet på den uppmätta ammoniumhalten hos en ammoniumgivare i dess normala miljö med ett referensvärde som produceras genom en analytisk laboratoriemätning. Rutinen har utformats för jonselektiva givare (*ion-selective electrode – ISE*). Rutinen kan justeras till analysatorer med kolorimetrisk mätprincip, vilket beskrivs under steg10 nedan.

Med denna rutin går det att:

- Utvärdera mätfelet vid fem olika ammoniumkoncentrationer
- Kvantifiera av bias (se Kapitel 3), olinjäriteter, offset och lutning

Om rutinen upprepas regelbundet över tid går det även att:

- Kvantifiera avdrift (av bias, olinjäritetert, offset och lutning)

Denna procedur kan inte användas för att:

- Kvantifiera givardynamik eller varians (se Kapitel 3)

Utrustning

- Ammoniumgivare som ska valideras
- 20 L hink
- 5 L hink
- Omrörare
- Provtagningsflaska för vattenprov av ammonium till laboratorium

Förbrukningsvaror

- 10 L avloppsvatten från givarens mätpunkt
- 1 L stamlösning av ammonium med en koncentration på 440 mg N/L
- 100 mg allyltioureapulver

Utförande

- Förberedelser (1–24 timmar före valideringen)
  (a) Rengör den ammoniumgivare som ska valideras.
- Validering
  (a) Ta ett vattenprov på 10 L med 20 L-hinken. Dekantera vid behov för att få en exakt volym.
  (b) Blanda vattenprovet försiktigt med omröraren till en homogen blandning utan sedimentering av suspenderade fasta ämnen.
  (c) Inspektera givaren visuellt och kontrollera att den är hel och ren.
  (d) Tillsätt allyltiourea för att hämma oxidationen av ammoniak (t.ex. 10 mg N/L är normalt långt över hämningsnivån).
  (e) Placera ammoniumgivaren i hinken.
  (f) Se till att avloppsvattnet rörs om kontinuerligt.
  (g) Se till att givarens pH-värde och temperaturvärde stabiliseras (normalt 5–10 min).
  (h) Ta ett 10 mL vattenprov från hinken för laboratorieanalys.
  (i) Registrera det stabiliserade pH-värdet från ammoniumgivaren.
  (j) Registrera det stabiliserade temperaturvärdet från ammoniumgivaren.
  (k) Registrera det stabiliserade värdet för ammoniumkoncentrationen från ammoniumgivaren.
  (l) Om fem mätningar har erhållits, gå till steg 2n. I annat fall, gå till steg 2 m.
  (m) Stegvis tillsats av stamlösning (upprepas fyra gånger):
    (i) Tillsätt 200 mL av stamlösningen av ammonium.
    (ii) Se till att givarens pH-värde och temperaturvärde stabiliseras (normalt 5–10 min).
    (iii) Gå tillbaka till steg 2 (i).
  (n) Återställ ammoniumgivaren till dess normala plats.
  (o) Häll tillbaka vattenprovet.
- Återställning
  (a) Rengör hink och omrörare.
  (b) Ställ tillbaka omrörare och hink på sina normala platser.
  (c) Överför det 20 ml stora vattenprovet till laboratoriet för analys med en standardmätning eller en accepterad analysmetod för referensmätning.

Dataanalys och lagring

- Lagra värdena för pH, temperatur och ammoniumkoncentration från givaren i en databas.
- Lagra även värdet på den uppmätta ammoniumhalten från laboratoriet i samma databas.
- Beräkna de förväntade koncentrationer som motsvarar respektive koncentrationsvärde för den ammoniumhalt som erhålls av givaren:

$$\text{förväntad koncentration} = \frac{\text{ammoniumhalt lab.} \times \text{provvolym} + \text{koncentration stamlösn.} \times \text{volym stamlösn.}}{\text{volym stamlösning} + \text{provvolym}}$$

- Utför en linjär regression med de förväntade ammoniumhalterna och motsvarande mätvärden från givaren. Avgör om regressionens kvalitet är acceptabel genom att granska residualerna som en funktion av de förväntade koncentrationerna. Nyttja denna inspektion till att utvärdera om givarens beteende är tillräckligt linjärt för att vara godtagbart. Om det anses acceptabelt, använd de beräknade regressionskoefficienterna för att bestämma och registrera givarens bias, offset och lutning.

Övrigt

- Samma vattenprover kan återanvändas för att validera flera ammoniumgivare. Då upprepas steg 1a och 2c–2k för varje nu ammoniumgivare.

- Automatisk insamling av ammoniumgivarens mätvärden via ett automatiserat datainsamlingssystem underlättar kvantifiering av givarens precision (se Kapitel 3).
- Om ammoniumgivaren har ytterligare elektroder (t.ex. K^+) för att kompensera för mätstörande effekter, lägg då till ett steg efter steg 2j för att även mäta det stabiliserade värdet som erhålls med dessa elektroder. Det kan vara av intresse att då även modifiera steg 3d så att jonkoncentrationen av den komponent som stör mätningen också kvantifieras. Efter steg 2 m kan stegvis tillsats av stamlösning av de förväxlande jonerna (t.ex. K^+) ingå i proceduren för att även mäta offset och lutning för dessa kompensationselektroder. Försiktighet bör iakttas så att inte nya förväxlingsfaktorer införs. Till exempel kommer höga koncentrationer av Cl- att påverka ammoniumelektrodens signal, så KCl-lösningar bör undvikas i dessa fall.
- För att modifiera denna procedur för användning med en ammoniumanalysator (kolorimetrisk mätprincip), ändra steg 1a till 2k genom att ta hänsyn till följande aspekter:
  (a) Justera rengöringsproceduren enligt analysatorns manual.
  (b) I stället för att ändra givarens position kan du ändra provtagningsledningens position så att analysatorn samlar upp vatten från hinken under valideringen.
  (c) Justera väntetiderna för att ta hänsyn till att analysatorn mäter koncentrationen med fördröjning (på grund av den tid som krävs för kolorimetrisk reaktion) och med lägre frekvens (t.ex. 10-minutersintervall).

5.A.3.3 Kontextuell information för validering av ammoniumgivare

Följande stycken är exempel på vad som kan ingå som kontextuell information i en valideringsrutin för ammoniumgivare.

Syfte med ammoniummätning

Givaren används oftast för att mäta och kontrollera ammoniumhalten i aktivslamprocessen. En pH- och temperaturmätning ingår för att kompensera för deras påverkan på elektrodpotentialen. Helst bör variationer i dessa kompensationsmätningar antingen kontrolleras eller mätas under valideringen av ammoniumgivaren.

Olika mätprinciper

Två typer av mätningar är vanliga för mätningar av ammonium: jonselektiva och kolorimetriska:

- *Jonselektiv elektrod:* En jonselektiv elektrod används för att mäta en elektrisk potential mellan två elektroder. Referenselektroden och mätelektroden är båda placerade i medier med en känd sammansättning och är båda separerade från mätmiljön med membran som gör att elektroner kan passera men begränsar överföringen av joner. Mätelektroden är konstruerad så att ammoniumjoner kan passera och ändra pH-värdet i mätelektroden. Detta resulterar i en elektrisk potential som kan mätas. I jämvikt är den elektriska potentialen proportionell mot logaritmen av koncentrationen av ammoniumhalten i den uppmätta miljön. För mätning av ammoniumkväve måste därför även pH och temperatur mätas.
- *Kolorimetrisk:* Den kolorimetriska metoden baseras på en reaktion mellan ammonium och ett reagens, vilken i sin tur ger upphov till en färgskiftning. Den resulterande förändringen i mediets ljusabsorbans är proportionell mot den ursprungliga ammoniumhalten.

Mätområde

För att välja rätt elektrod eller analysator måste man ta hänsyn till mätstörande faktorer (se nedan) och mätområdet. En jonselektiv elektor kan anses vara tillförlitlig från 1 till 1000 mg N/L. Detta är tillräckligt för övervakning och styrning av konventionella biologiska nitrifikationsprocesser, men inte tillräckligt noggrant för att övervaka av vattenkvaliteten i utgående vatten.

Mätstörande faktorer (jonselektiv elektrod)

Den uppmätta elektrodpotentialen påverkas av ammoniumhalten (NH_4^+). Halten av ammoniumjoner förändras som en funktion av ammoniumhalten, pH-värdet och temperaturen. För att mäta

ammoniumhalten måste därför elektrodpotentialen korrigeras för dessa mätstörande faktorer. Många kommersiella jonselektiva givare innehåller extra elektroder för automatisk korrigering.

Förutom pH och temperatur påverkas elektrodpotentialen även av joner med samma storlek och laddning som ammoniumjonen (NH_4^+). Den vanligaste mätstörningen orsakas av variationer i kaliumkoncentrationen (K^+) och kan bero på vägsalt som läcker in i ledningsnätet under vinterhalvåret. Vissa ammoniumgivare innehåller elektroder som är känsliga för sådana joner. Detta gör det möjligt att kompensera för deras effekt på den uppmätta elektrodpotentialen. Givarleverantören bör konsulteras om detta är ett problem. När det är tillämpligt kan ovanstående valideringsrutiner modifieras för att fånga upp mätfelet även för korrigeringselektroder.

Mätstörande faktorer (ammoniumanalysator)

Den kolorimetriska mätprincipen är mycket mer specifik för ammoniumkväve och har få eller inga mätstörande faktorer när den används inom det mätområde som anges av tillverkaren.

5.A.4 Valideringsrutin för givar som mäter total suspenderade halt

Detta avsnitt beskriver fyra rutiner för validering av givare som mäter total halt av suspenderade ämnen (TSS). Figur 5.A.4 visar de datakvalitetsmått som kan uppskattas med dessa rutiner.

5.A.4.1 TSS-rutin-1: Funktionskontroll av givaren

Version och datum

- Version: 1.0
- Senaste översyn och revidering: 31 juli 2023
- Nästa granskning och revidering ska vara genomförd senast den 31 juli 2025

Syfte

Syftet med denna rutin är att kontrollera avvikelsen från fabrikskalibreringen med en fysisk referensplatta med kända optiska egenskaper. Eftersom slammets egenskaper påverkar de uppmätta suspenderade ämnena bör denna kontroll endast betraktas som en kontroll av givarens grundfunktion före en fullständig validering.

Med denna rutin kan man:

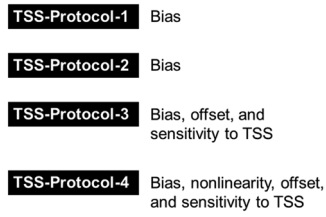
- Kvantifiera bias under standardiserade icke processmässiga förhållanden

Om rutinen upprepas regelbundet går det även att:

- Kvantifiera avdrift på grund av givarens elektronik

Utrustning

- TSS-givare ska valideras
- En fysisk TSS-referensskiva som passar den givare som ska valideras. Referensskivan är en solid bit av ett material, t.ex. plast, med fasta optiska egenskaper. Eftersom skivan måste passa i givarens mätfönster är den vanligtvis skräddarsydd av givartillverkaren och inte en generisk komponent.



Figur 5.A.4 Översikt över valideringsförfaranden för TSS-givareer och resulterande metadata för givaredatakvalitet.

Förbrukningsvaror

- Hink med varmvatten (för rengöring)
- Svamp eller borste (för rengöring)
- Utspädd syra av 1–3 % HCL (för rengöring)
- Tvättmedel (för rengöring)
- Pappershandduk

Utförande

(1) Rengör givaren med vatten, rengöringsmedel eller utspädd syra tills alla synliga rester har avlägsnats.
(2) Torka givaren, och i synnerhet mätglasfönstret, med papper.
(3) Montera den fysiska TSS-referensskivan på givaren.
(4) Låt mätningen stabiliseras vid aktuell rumstemperatur.
(5) Registrera den råa TSS-signalen. Obs: Standardvärdet för TSS ska inte användas eftersom det ändras under kalibreringen.
(6) Jämför råsignalen med föregående avläsning (eller fabriksvärde) för att uppskatta avdriften sedan senaste valideringen.

5.A.4.2 TSS-rutin-2: Kvantifiering av bias

Version och datum

- Version: 1.0
- Senaste översyn och revidering: 31 juli 2023
- Nästa granskning och revidering ska vara genomförd senast den 31 juli 2025

Syfte

Syftet med denna rutin är att validera TSS-värdet för en TSS-givare i dess normala mätmiljö med en handhållen referensmätning. Denna rutin är framför allt lämplig på reningsverk med många TSS-givare eftersom samma hink med avloppsvatten kan återanvändas för flera givare.

Med denna rutin kan man:

- Kvantifiera bias beräknad från två TSS-givare

Om rutinen upprepas regelbundet går det även att:

- Kvantifiera avdrift (bias).

Utrustning

- TSS-givare ska valideras
- En handhållen TSS-givare som fungerar som referens, nyligen (inom 1 vecka) kalibrerad och validerad med TSS-rutin-3 eller TSS-rutin-4
- 20 L hink
- Omrörare
- Ett mörkt tygstycke (behövs om TSS-givaren är placerad utomhus)

Förbrukningsvaror

- 10 L avloppsvatten från TSS-givarens mätpunkt

Utförande

(1) Förberedelser (upp till 1 vecka före valideringen)
  (a) Validera och/eller kalibrera den handhållna referens-TSS-givaren med hjälp av TSS-rutin-3 eller TSS-rutin-4.

(2) Validering
  (a) Inspektera TSS-givaren visuellt och kontrollera att glasfönstren som mäter absorbansen är hela och rena.
  (b) Rengör TSS-givaren som ska valideras (utelämna rengöring om den aktuella mätningen med potentiellt nedsmutsad givare ska valideras).
  (c) Ta ett vattenprov på 10 liter med 20 L-hinken.
  (d) Blanda vattenprovet försiktigt med omröraren för att få en homogen blandning utan sedimentering av suspenderade ämnen.
  (e) Placera den handhållna TSS-givaren i hinken och rör den försiktigt runt i en cirkelrörelse tills mätvärdet stabiliseras.
  (f) Täck över hinken med den mörka duken när du validerar utomhus.
  (g) Registrera värdet för referens-TSS-givaren.
  (h) Blanda vattenprovet försiktigt med omröraren för att få en homogen blandning utan sedimentering av suspenderade ämnen.
  (i) Placera TSS-testgivaren i hinken.
  (j) Låt TSS-avläsningen stabiliseras samtidigt som försiktigt rör runt i avloppsvattnet. Observera att detta kan ta en tid (flera minuter) om en lång tidskonstant för lågpassfiltret används i givarens digitala kommunikationsbox.
  (k) Täck över hinken med den mörka duken när du validerar utomhus.
  (l) Registrera det stabiliserade TSS-värdet.
  (m) Återställ TSS-givaren till dess normala mätposition.
  (n) Häll tillbaka vattenprovet (eller fortsätt med nästa TSS-givare och spara in steg 2b–2g).
  (o) Rengör hink och omrörare.
  (p) Ställ tillbaka omrörare, hink och den handhållna referens-TSS-givaren på sin normala plats.

Dataanalys och lagring

(1) Skillnaden mellan medelvärdet från referens-TSS-givaren (steg 2e–2g) och TSS-givaren (steg 2l) är en uppskattning av det absoluta mätfelet för givaren under aktuella driftsförhållanden (antingen rengjord eller i befintligt skick).

5.A.4.3 TSS-rutin-3: Tvåpunktsvalidering vid noll susphalt och en vald koncentration

Version och datum

- Version: 1.0
- Senaste översyn och revidering: 31 juli 2023
- Nästa granskning och revidering ska vara genomförd senast den 31 juli 2025

Syfte

Syftet med denna rutin är att validera TSS-värdet för en TSS-givare vid noll och den vanliga uppmätta TSS-koncentrationen (ca 2000–4000 mg/L i aktivslambassängen eller 5000–12 000 mg/L i returslammet). Proceduren kan användas för att validera både fast monterade och handhållna TSS-givare.

Med denna rutin går det att:

- Kvantifiera bias, offset och lutning, beräknat från två TSS-koncentrationer

Om rutinen upprepas regelbundet över tid går det även att:

- Kvantifiera drift (av bias, offset och lutning)

Utrustning

- TSS-givare ska valideras
- 2st 20 L hinkar

- 1st 10 L hink
- Omrörare
- Rengöringsutrustning: borste och svamp
- Skyddsutrustning: skyddsglasögon och skyddshandskar (vid rengöring med syra)
- Ett mörkt tygstycke (om TSS-givaren är placerad utomhus)

Förbrukningsvaror

- 5 L avloppsvatten från givarens mätpunkt
- 5 L kranvatten eller destillerat vatten
- Pappershandduk
- Utspädd saltsyra (1–3 % HCL) för rengöring
- Flaskor för vattenprov för laboratoriemätningar av suspenderade ämnen. Använd 250 ml-flaskor om inte laboratoriet använder något annat.

Utförande

(1) Förberedelser (upp till 24 timmar före valideringen)
  (a) Fyll 10-litershinken med kranvatten eller destillerat vatten och låt den avluftas inomhus. Luftbubblor från kranvattnet kan påverka avläsningen och den fyllda hinken bör få lufta i minst 4 timmar, men helst i 24 timmar.
  (b) Fyll en av 20-litershinkarna med 5 liter varmt vatten (ska användas för rengöring).

(2) Validering
  (a) Rengör och registrera givarens värde vid noll TSS.
    (i) Rengör TSS-givaren som ska valideras med rent vatten.
    (ii) Inspektera TSS-givaren visuellt och kontrollera att glasfönstren som mäter absorbansen är hela och rena.
    (iii) Placera den rengjorda TSS-givaren i 10 L-hinken med avluftat rent vatten och låt mätningen stabiliseras.
    (iv) Om mätvärdet är över noll, rengör givaren igen med en syralösning för att avlägsna kalcium och andra utfällningar.
    (v) Skölj givaren med rent vatten.
    (vi) Upprepa steg ii och försäkra dig om att givaren mäter noll, annars upprepa steg (iii) och (iv) tills avläsningen ligger mellan 0 och 10 mg/L.
    (vii) Registrera TSS-värdet för rent vatten, dvs noll susphalt.
  (b) Givarvärde vid aktuell driftspunkt.
    (i) Ta ett vattenprov på 10 liter med 20-litershinken invid givarens mätpunkt (eller med önskad koncentration).
    (ii) Blanda vattenprovet försiktigt med omröraren för att få en homogen blandning utan sedimentering av suspenderade ämnen.
    (iii) Placera TSS-givaren i hinken och rör försiktigt runt den i en cirkelrörelse tills mätvärdet stabiliseras.
    (iv) Täck över hinken med ett mörkt tyg när du validerar utomhus.
    (v) Registrera mätvärdet.
  (c) Referensmätning av TSS.
    (i) Blanda vattenprovet försiktigt med omröraren för att få en homogen blandning utan sedimentering av suspenderade ämnen.
    (ii) Ta ett vattenprov (med 250 mL vattenprovflaska) och skicka det till ett laboratorium för analys.
  (a) Återställ TSS-givaren till dess normala mätplats.
  (b) Häll tillbaka vattenprovet.
  (c) Rengör hink och omrörare.

Dataanalys och lagring

(1) Lagra givarens mätvärden i kranvatten och vid aktuell driftspunkt i en databas.
(2) Lagra referensmätvärdet för TSS från laboratoriet i samma databas.
(3) Beräkna avvikelsen i offset- och lutning med hjälp av linjär regression och de registrerade värdena.
(4) Spara värden på offset- och lutning tillsammans med mätvärdena som erhållits i steg 2a–2c som valideringsmetadata.

5.A.4.4 TSS-rutin-4: Flerpunktsvalidering av susphaltsgivare

Version och datum

- Version: 1.0
- Senaste översyn och revidering: 31 juli 2023
- Nästa granskning och revidering ska vara genomförd senast den 31 juli 2025

Syfte

Syftet med denna rutin är att validera mätsignalen från optiska TSS-givare vid tre olika TSS-koncentrationer genom att successivt späda ut en koncentrerad stamlösning.

Med denna rutin kan man:

- Kvantifiera bias, olinjäriteter, offset och lutning

Om rutinen upprepas regelbundet går det även att:

- Kvantifiera avdrift (av bias, offset, lutning och olinjäriteter)

Utrustning

- TSS-givare ska valideras
- 1 L bägare
- 4 st 20 L-hinkar
- 1 st 10 L-hink
- Omrörare
- Rengöringsutrustning: borste och svamp
- Skyddsutrustning: skyddsglasögon och skyddshandskar (vid rengöring med syra)
- En mörk trasa (när TSS-givaren valideras utomhus)

Förbrukningsvaror

- 30 L (två hinkar med 15 L vardera) avloppsvatten med en koncentration i mitten av givarens önskade mätområde.
- 3 flaskor för vattenprov för laboratoriemätningar av suspenderade ämnen. Använd 250 ml-flaskor om inte laboratoriet använder något annat.

Utförande

(1) Förberedelser
  (a) Förbered en lösning med noll TSS (upp till 24 timmar före valideringen). Fyll 10-litershinken med kranvatten eller destillerat vatten och låt den avluftas inomhus. Luftbubblor från kranvattnet kan påverka avläsningen och den vattenfyllda hinken bör få lufta ut i minst 4 timmar, men helst i 24 timmar.
  (b) Fyll en 20 liters hink med minst 15 liter kranvatten (som ska användas för att späda ut vattenproverna).

(c) Förbered TSS-stamlösningen (upp till 1 timme före valideringen).
  (i) Fyll två 20 L hinkar med 15 L avloppsvatten vardera med en koncentration i mitten av det önskade valideringsintervallet.
  (ii) Låt slammet sedimentera i 30–60 minuter.
(d) Förbered för rengöring av givaren genom att fylla den återstående 20 L hinken med 5 L varmt vatten.
(e) Förbered den koncentrerade (150 %) stamlösningen.
  (i) Dekantera försiktigt ca 30 % (5 L) av klarfasen med låg TSS i båda hinkarna i steg 1c. Koncentrationen ska nu vara ca 50 % högre än i det ursprungliga vattenprovet.
  (ii) Slå samman innehållet i de båda hinkar till en fylld 20 L hink.
  (iii) Bibehåll omrörning och avlägsna överflödigt avloppsvatten tills 15 L återstår i hinken.
  (iv) Blanda vattenprovet försiktigt med omröraren för att få en homogen blandning utan sedimentering av suspenderade ämnen, ta sedan ett vattenprov för laboratorieanalys märkt som *TSS 150 %*.

(2) Validering
(a) Rengör och registrera givarens värde vid noll TSS.
  (i) Rengör TSS-givaren som ska valideras med rent vatten.
  (ii) Inspektera TSS-givaren visuellt och kontrollera att glasfönstren som mäter absorbansen är hela och rena.
  (iii) Placera den rengjorda TSS-givaren i 10 L-hinken med avluftat rent vatten och låt mätningen stabiliseras.
  (iv) Om mätvärdet är över noll, rengör givaren igen med en syralösning för att avlägsna kalcium och andra utfällningar.
  (v) Skölj givaren med rent vatten.
  (vi) Upprepa steg ii och försäkra dig om att givaren mäter noll, annars upprepa steg (iii) och (iv) tills avläsningen ligger mellan 0 och 10 mg/L.
  (vii) Registrera mätvärdet noll för rent vatten.
(b) Mät den koncentrerade stamlösningen.
  (i) Placera TSS-givaren i mitten av hinken som innehåller 150 % TSS-koncentration. Givaren ska vara 10 cm från väggar eller botten och mätningen ska stabiliseras innan den avläses.
  (ii) Blanda vattenprovet försiktigt med omröraren till en homogen blandning utan sedimentering av suspenderade ämnen.
  (iii) Registrera givarens mätvärde vid 150 % TSS av den ursprungliga TSS-koncentrationen.
(c) Bered utspädda lösningar på 1/3 och 2/3 av den koncentrerade stamlösningen.
  (i) Späd 5 liter av stamlösningen med 10 liter kranvatten (eller avloppsvatten). Denna lösning kommer att ligga mellan noll och den ursprungliga TSS-koncentrationen.
  (ii) Ta ett vattenprov för laboratorieanalys märkt som *TSS 50 %*.
  (iii) Späd de återstående 10 L med stamlösning med 5 L kranvatten. Detta ger samma koncentration som det ursprungliga provet (eller något lägre på grund av eventuella dekanterade suspenderade ämnen).
  (iv) Ta ett vattenprov för laboratorieanalys märkt som *TSS 100 %*.
(d) Mät TSS i två koncentrationer.
  (i) Placera TSS-givaren i en av hinkarna och virvla försiktigt runt i en cirkelrörelse tills avläsningen stabiliseras. Se till att givaren är minst 10 cm från väggen och hinkens botten för att undvika ljusspridningseffekter. Täck över hinken med en mörka duk vid validering utomhus.
  (ii) Registrera givarens värde.
  (iii) Upprepa proceduren för den andra hinken.

(e) Återställ TSS-givaren till dess normala plats.
(f) Häll tillbaka vattenprovet.
(g) Rengör hink och omrörare.

Dataanalys och lagring

(1) Registrera givarens mätvärden i kranvatten och vid de tre koncentrationerna i en databas.
(2) Registrera referensmätvärdena för TSS från laboratoriet i samma databas.
(3) Beräkna offset- och lutning med linjär regression utifrån de registrerade givar- och labbvärdena.
(4) Spara offset- och lutningsvärdena tillsammans med de registrerade värdena som valideringsmetadata.

Det går även att kontrollera laboratoriemätningarnas spridning när utspädningen i hinkar har utförts noggrant. Då bör laboratoriemätningarna följa utspädningen på 50, 100 och 150 % av det ursprungliga vattenprovet.

5.A.4.5 Kontextuell information om validering av suspgivare

Nedan ges exempel kontextuell information som kan ingå i en valideringsrutin av TSS-givare.

Mättekniker

TSS uppskattas med en optisk och infraröd (IR) mätning i det icke-synliga våglängen (860–890 nm). Det finns två mättekniker, där den första mäter ljusspridningen vid en viss vinkel (90° för låga halter och 135° för högre halter), och den andra typen mäter absorptionen genom en tunn spalt (180°).

Mätstörande faktorer

Dagsljus kan påverka IR-avläsningarna och givaren bör därför skyddas från direkt solljus vid mätningar utomhus.

Det ljus som givaren skickar ut kan reflekteras på en nära vägg eller en närliggande givare (särskilt för typen som baseras på ljusspridning). Av denna anledning rekommenderar de flesta tillverkare ett minsta avstånd på 10 cm till väggar eller botten på omgivande tank.

Luftbubblor kan reflektera ljus och påverka avläsningen. Därför rekommenderas avluftat vatten för nollkalibrering.

Många TSS-givare innehåller en temperaturgivare för att korrigera för stora temperaturförändringar. Det innebär att givaren måste anpassa sig till eventuella temperaturförändringar vid förflyttning, t.ex. när givaren valideras i en hink med annan temperatur än processens.

Korrelationen mellan den optiska mätningen (spritt eller absorberat ljus) och TSS kan vara icke-linjär. Eftersom denna icke-linjäritet kan variera under året är det inte möjligt att korrigera för detta i givarens mätberäkningar. Istället används flerpunktskalibrering för att anpassa sig till de lokala slamegenskaperna. Hur detta potentiella icke-linjära förhållande hanteras under kalibreringen i givaren skiljer sig åt mellan olika tillverkare, och det är därför särskilt viktigt att förstå kalibreringsproceduren i detalj för den aktuella givaren. Det bör noteras att den föreslagna valideringsrutinen med flera mätpunkter är tillräcklig för att upptäcka en felaktig kalibrering.

Allmänna rekommendationer för referensprovtagning av TSS

Det är en utmaning att mäta TSS i avloppsvatten av två skäl.

För det första är slammet heterogent och varierar kraftigt med tid och rum i en aktivslambassäng. Det är därför svårt att ta ett representativt vattenprov om målet är att uppnå en representativ koncentration av givarens omgivande förhållanden. Även om det är möjligt att placera en handhållen givare bredvid den permanenta givaren bör alltid ett delprov tas i en hink för att säkerställa att både referensgivaren och den validerade givaren mäter samma TSS-koncentration.

För det andra kan laboratorieanalyser av TSS ha systematiska fel och slumpmässiga fel vid höga TSS-koncentrationer (över 3000 mg/l). Därför är välutvecklade laboratoriemetoder (APHA och ISO) och replikatanalyser (2–3 analyser för varje vattenprov) avgörande för att uppnå hög noggrannhet vid valideringen. Det är troligt att laboratoriets rutiner inte är tillämpliga (ackrediterade) för höga TSS-koncentrationer såsom returslam. I dessa fall bör slamprovet spädas till en accepterad koncentration. Oftast görs detta med destillerat vatten eller kranvatten, även om detta kan späda ut ljusabsorberande komponenter och ge en bias. Spädning med renat avloppsvatten rekommenderas i allmänhet inte eftersom det kan innehålla suspenderade ämnen. Det är dock troligt att de upplösta ljusabsorberande komponenterna finns i liknande koncentrationer, så att utspädningen av de ljusabsorberande komponenterna kan minskas på detta sätt (även om det kräver extra mätningar av susphalten i det reande avloppsvattnet.)

Koncentrationsintervallet i aktivslambassängen varierar under dygnet och säsongen beroende på inflödet. Det kan därför vara svårt att få fram både låga och höga TSS-koncentrationer när hela mätområdet ska valideras. Två enkla knep kan användas för att få en spridning i koncentration. Om koncentrationen i vattenprovet är för hög kan det för det första spädas ut med avloppsvatten (inte kranvatten). För det andra, om vattenprovets koncentration behöver höjas, kan ett vattenprov från returslammet spädas till önskad koncentration. Även när returslammet är för utspätt (t.ex. under högflödessituationer) kan ett stort vattenprov (20 L) tas och tillåtas sedimentera, och den rena vattenfasen avlägsnas genom att försiktigt hälla eller suga ut den. Detta kan upprepas för att uppnå önskad koncentration.

I manualer för TSS-givare rekommenderas vanligen att givarspetsen hålls minst 5 cm under vattenytan samtidigt som spetsen hålls mer än 10 cm från hinkens sidor och botten. Detta kräver vanligtvis att man använder en 20 L hink som fylls med minst 10 L avloppsvatten.

5.A.5 Valideringsrutiner för givare med absorbansmätning av UV-ljus

I detta avsnitt beskrivs två valideringsrutiner för mätning av absorbansen av ultraviolett ljus i det synliga våglängdsområdet (UV/Vis) även kallat spektrometer/spektrofotometer. Figur 5.A.5 visar de datakvalitetsmått som kan uppskattas med dessa rutiner. Detta gäller för spektrometrar som mäter koncentrationer av olika ämnen i avloppsvatten, oftast genom mätning av ett multivariat ljusabsorbansspektrum. Valideringsrutinerna är dock inte utformade för UV/AOP-system (avancerade oxidationsprocesser), men kan modifieras för detta ändamål.

Syftet med denna valideringsrutin är att ge riktlinjer för att kvalitetssäkra mätningarna.

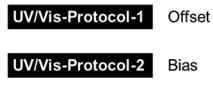
5.A.5.1 UV/Vis-rutin-1: Kontroll av givarfunktion

Version och datum

- Version: 1.0
- Senaste översyn och revidering: 31 juli 2023
- Nästa granskning och revidering ska vara genomförd senast den 31 juli 2025

Syfte

Syftet med denna rutin är att kontrollera givarens blankspektrum. Ett mätning av givarens blankspektrum görs efter vart och ett av tre rengöringssteg för att kvantifiera effekten av rengöringen och förändringar i givaren genom jämförelse med en historiska data på av blankspektra.



Figur 5.A.5 Valideringsförfaranden för spektrometrar och resulterande metadata för givarens datakvalitet.

Denna rutin gör det möjligt att:

- Kvantifiera offset i standardiserade förhållanden icke-processförhållanden

Om rutinen upprepas över tid går det även att:

- Kvantifiera avdrift av offset på grund av nedsmutsning, utfällning eller permanent skada på givarens optiska komponenter.

Utrustning

- Spektrometer (UV/Vis-givare) som ska valideras
- 10 L hink
- Mjuk trasa eller borste för rengöring
- Omrörare

Förbrukningsvaror

- 10 L destillerat vatten
- Tvättmedel (för rengöring)
- Utspädd syra 1–3 % HCL (för rengöring)
- Pappersdukar (för avtorkning av mätfönster)

Utförande

(1) Förberedelser
  (a) Avaktivera givarens automatiska rengöringsfunktioner (t.ex. luftspolning, torkare).
  (b) Loss givaren från sin normala mätpunkt.
  (c) Ta bort eventuell smuts och fasta partiklar givarens optiska mätfönster och rengör givaren med vatten och trasa, torka sedan med en mjuk papperstrasa.
  (d) Rengör tryckluftskanalerna (om sådana finns) med kranvatten och med hjälp av bomullstopps eller specialanpassade rengöringsborstar.

(2) Validering stegvis rengöring
  (a) Inspektera givaren och kontrollera att den är hel och ren.
  (b) Placera givaren i hinken med destillerat vatten.
  (c) Se till att det destillerade vattnet rörs om kontinuerligt.
  (d) Gör en mätning och spara UV/Vis-spektrumet (spektrum 1).
  (e) Ta upp givaren ur hinken.
  (f) Rengör UV/Vis med vatten, rengöringsmedel och en trasa och torka torrt med en pappersduk.
  (g) Skölj med destillerat vatten.
  (h) Placera återigen givaren i den hinken med destillerat vatten.
  (i) Se till att det destillerade vattnet rörs om kontinuerligt.
  (j) Gör en mätning och spara UV/Vis-spektrumet (spektrum 2).
  (k) Rengör givarens mätfönster med utspädd syra och torka torrt med pappersduk.
  (l) Skölj med destillerat vatten.
  (m) Placera givaren i hinken med destillerat vatten.
  (n) Se till att det destillerade vattnet rörs om kontinuerligt.
  (o) Gör en mätning och spara UV/Vis-spektrumet (spektrum 3).
  (p) Sätt tillbaka givaren på sin normala mätpunkt

(3) Rengöring och återställning
  (a) Töm hinken och rengör omröraren.
  (b) Ställ tillbaka hinken och omröraren på sin platser.
  (c) Återaktivera de automatiska rengöringsfunktioner som avaktiverats i steg 1(a).

Dataanalys och lagring

(1) Spara de tre spektra som erhållits efter respektive rengöringssteg (spektra 1–3: vatten, rengöringsmedel, syra).
(2) Spektrum 3 bör motsvara ett blankspektrum som bör vara konstant och nära noll över hela våglängdsområdet. Jämför detta blankspektrum med en historik av blankspektra (visuellt eller numeriskt) för att se om ytterligare rengöring eller underhåll kan vara motiverat eller om permanent skada har uppstått.
(3) Skillnaden mellan spektra 1 och 2 är en indikator på hur mycket material som kan avlägsnas med lätt rengöring. Kvalitativt motsvarar detta rengöring av biofilmspåväxt. I idealfallet är denna skillnad liten. Om inte, öka rengöringsfrekvensen tills skillnaden är acceptabel.
(4) Skillnaden mellan spektra 2 och 3 är en indikator på hur mycket material som endast kan avlägsnas med kemikalier, i detta fall syra. Det motsvarar effekten av att rengöra utfällningar. I idealfallet förblir denna skillnad liten. Om inte, öka frekvensen av kemisk rengöring tills skillnaden är acceptabel.

Övrigt

- Det går att upprepa proceduren för flera givare med samma vattenprov för att spara tid. Då kan steg 1a-d och 2c-k repeteras.
- Det underlättar med ett automatiserat datainsamlingssystem för alla mätvärden för att kvantifiera givarens precision (se Kapitel 3).
- Vissa tillverkare tillhandahåller ett fast referensmaterial som kan placeras i mätfönstret (se nedan). Detta kallas också för ett multifunktionellt objektglas. Detta material ersätter destillerat vatten som referensmedium, vilket ger ett snabbare och mer tillförlitligt resultat.

5.A.5.2 UV/Vis-rutin-2: Enpunktsvalidering vid aktuell driftpunkt

Version och datum

- Version: 1.0
- Senaste översyn och revidering: 31 juli 2023
- Nästa granskning och revidering ska vara genomförd senast den 31 juli 2025

Syfte

Syftet med denna rutin är att validera givarens mätvärde (t.ex. kemisk syreförbrukning (COD) eller nitrathalt) i givarens normala miljö med ett laboratoriereferensvärde.

Med denna rutin går det att:

- Utvärdera mätfelet vid den aktuella driftspunkten
- Kvantifiera förekomst av systematiska fel (se Kapitel 3)

Om rutinen upprepas över tid går det även att:

- Kvantifiera avdrift (dvs. drift av bias, se Kapitel 3)

Utrustning

- Spektrometer som ska valideras
- 10 L hink
- Omrörare
- Provtagningsflaska för labbprov.

Förbrukningsvaror

- 10 L avloppsvatten från givarens mätpunkt

- 100 mg allyltiourea pulver
- Natriumkloratsalt

Utförande

(1) Förberedelser (1–24 timmar före valideringen)
  (a) Avaktivera givarens automatiska rengöringsfunktioner (t.ex. luftspolning, torkare).
  (b) Loss givaren från sin normala mätpunkt.
  (c) Ta bort eventuell smuts och fasta partiklar givarens optiska mätfönster och rengör givaren med vatten och trasa, torka sedan med en mjuk papperstrasa.
  (d) Rengör tryckluftskanalerna (om sådana finns) med kranvatten och med hjälp av bomullstopps eller specialanpassade rengöringsborstar.
(2) Validering
  (a) Ta ett vattenprov på 10 liter med hinken.
  (b) Starta omrörning av vattenprovet och skapa en homogen blandning utan sedimentering av suspenderade fasta ämnen.
  (c) Inspektera givaren visuellt och kontrollera att den är hel och ren.
  (d) Tillsätt allyltiourea till vattenprovet för att hämma nitrifikation (t.ex. 10 mg N/L är gott och väl tillräckligt).
  (e) Tillsätt klorat för att hämma oxidationen av nitrit (t.ex. en koncentration på 20 mM).
  (f) Placera givaren i hinken. Placera givaren så att mätfönstret är minst 5 cm djupt ned samtidigt som givarspetsen är minst 10 cm från hinkens sidor och botten.
  (g) Se till att avloppsvattnet rörs om kontinuerligt.
  (h) Registrera UV/Vis-spektrumet.
  (i) Registrera spektrometerns utdata för den parameter som valideras (t.ex. COD-halt).
  (j) Ta ett 20 mL vattenprov från hinken för laboratorieanalys.
  (k) Ställ tillbaka givaren till dess normala mätpunkt.
(3) Återställning
  (a) Häll tillbaka vattenprovet.
  (b) Rengör hink och omrörare och ställ tillbaka dem på sina platser.
  (c) Analysera vattenprovet (20 ml) med en ackrediterad analysmetod på laboratoriet.

Dataanalys och lagring

(1) Spara värdena för spektrumet och givarutgången i en databas.
(2) Spara även ned laboratoriets analysresultat av vattenprovet.
(3) Skillnaden mellan givarens mätvärde och den analysresultatet är en uppskattning av givarens absoluta mätfel under aktuella driftsförhållanden och kan användas för att avgöra om ytterligare åtgärder är nödvändiga.

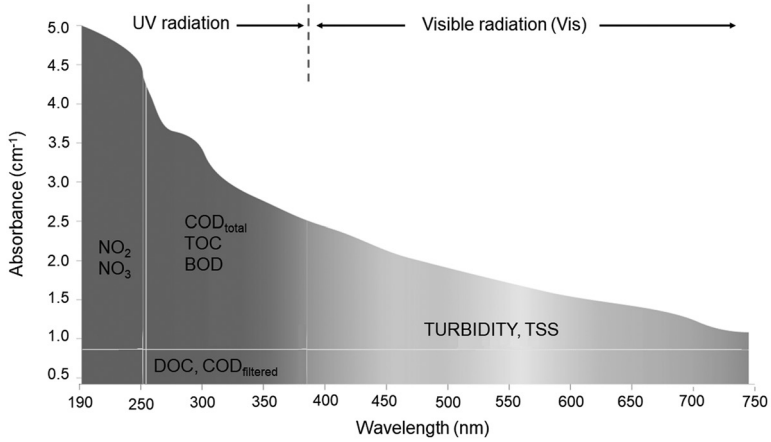
Övrigt

- Det går att upprepa proceduren för flera givare med samma vattenprov för att spara tid. Då kan steg 1a-d och 2c-k repeteras.
- Det underlättar med ett automatiserat datainsamlingssystem för alla mätvärden för att kvantifiera givarens precision (se Kapitel 3).

5.A.5.3 Kontextuell information om spektrala absorbansmätningar av UV-ljus

Spektrometrar baseras på mätning av den mängd UV/Vis-ljus som absorberas av ett vattenprov. Kommersiella givare mäter antingen en datapunkt (a) ljusabsorptionen vid en enda våglängd (t.ex. 254 nm) eller (b) ljusabsorptionen vid flera våglängder i området 200–750 nm beroende på givarleverantören. Figur 5.A.6 visar ett typiskt absorbansspektrum som funktion av våglängden.

Under idealiska mätförhållanden är mängden absorberat ljus proportionellt mot den ljusabsorberande komponents koncentration. Detta samband beskrivs av Lambert-Beers lag. I praktiken innehåller ett



Figur 5.A.6 Absorbansspektrum i det ultravioleytta synliga området med vattenkvalitetsparametrar. *Källa*: RealTechWater Inc: Bearbetad från RealTechWater Inc (2017).

vattenprov flera olika ljusabsorberande komponenter. Även om detta komplicerar givarkalibreringen förblir det teoretiska förhållandet mellan komponentkoncentrationerna och den resulterande ljusabsorptionen linjärt (Thomas & Burgess, 2017).

Om absorbansen mäts vid flera våglängder samtidigt går det att uppskatta koncentrationen av flera ämnen samtidigt med en enda datapunkt. Beroende på den exakta vattensammansättningen kan man uppskatta koncentrationerna av nedan ämnen i vatten:

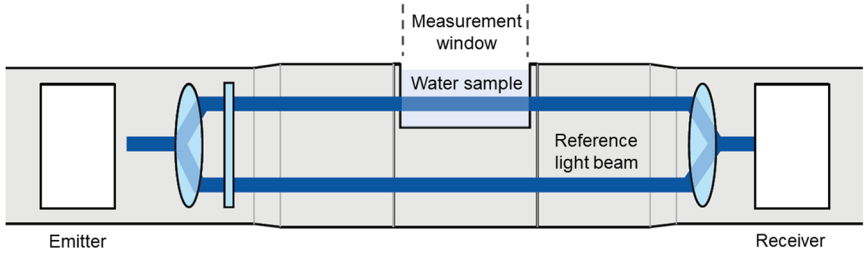
- Biokemisk syreförbrukning (BOD)
- Bensen, toluen, xylen (BTX)
- Kemisk syreförbrukning (COD)
- Löst organiskt kol (DOC)
- Nitrat (NO_3)
- Nitrit (NO_2)
- Totalt organiskt kol (TOC)
- Totalt suspenderat material (TSS)

Det exakta förhållandet mellan absorbansspektrumet och de aktuella koncentrationerna kan variera från en avloppsvattenström till en annan och kan också variera i tiden. En anledning till detta är att parametrar överlappar (t.ex. BOD, COD, DOC och TOC) motsvarar en mängd ljusabsorptionskomponenter vars individuella spektrum är olika och vars relativa koncentrationer kan förändras. Dessutom absorberar suspenderade partiklar inte bara ljus utan sprider det också. Av den anledningen kan förhållandet mellan ljusabsorptionsmätningen och de aktuella koncentrationerna bli icke-linjärt.

De flesta givartillverkare tillhandahåller modeller i den spektrofotometriska utrustningen för att korrigera absorbansmätningarna till de aktuella koncentrationerna. Beroende på vilken noggrannhet som krävs (bias och/eller precision, se Kapitel 3) kan en lokal kalibrering krävas av modellen för att ta hänsyn till platsspecifika förhållanden. Om avloppsvattnet har en unik sammansättning (t.ex. urin, Mašić *et al.*, 2015) kan en specialmodell tas fram.

5.A.5.4 Allmänna rekommendationer för validering av spektrometrar

Optisk mätning. En spektrometer består av en ljuskälla, mätfönstret (med vattenprov) och en mottagare vilket illustreras i Figur 5.A.7. En ljusstråle sänds ut genom mätfönstret genom mätfönstret med vattenprov



Figur 5.A.7 Typisk konstruktion av en online spektrometer. En ljuskälla 'emitter' producerar vitt ljus som delas upp i två strålar. Den ena strålen (överst) passerar genom ett mätfönster som innehåller vattenprovet och samlas upp vid mottagaren. Den andra strålen (nederst) når mottagaren direkt och kallas referensljusstrålen. Skillnaden i ljusintensitet mellan de två strålarna mäts för en serie våglängder, som visualiseras i ett ljusabsorptionsspektrum. *Källa*: S::CAN Bearbetad från S::CAN GmbH (2020).

samt genom en optisk referensbana som inte innehåller något vattenprov. Den relativa skillnaden mellan ljusstrålens intensitet används för att kvantifiera ljusabsorptionen vid en eller flera ljusvåglängder.

Ett rent och funktionellt mätfönster är avgörande för att givaren ska mäta korrekt. Minsta förekomst av smuts eller luftbubblor påverkar mätningarnas noggrannhet och korrekthet. Att hålla mätfönstret rent och repfritt är därför grundläggande för tillförlitliga mätningar och kärnan i underhållsarbetet.

Korrelationsbaserade mätningar. Spektrometrar kan ge korrelationsbaserade uppskattningar av olika vattenkvalitetsparametrar. Det är viktigt att de korrelationer som givaren förlitar sig på inte oavsiktligt påverkas vid hanteringen av vattenprov. Detta är särskilt viktigt för variabler som representerar koncentrationer av flera komponenter (t.ex. BOD, COD, DOC och TSS).

Det är svårt att mäta överlappande parametrar (t.ex. BOD, COD, DOC och TSS) i avloppsvatten av två anledningar. För det första representerar dessa koncentrationer flera komponenter vars individuella koncentrationer kan förändras medan den sammanlagda koncentrationen förblir densamma. I praktiken är de överlappande parametrarna heterogena och varierar kraftigt över tid och rum i en aktivslambassäng. Det är svårt att ta ett representativt vattenprov om målet är att skapa en representation av de omgivningsförhållanden som givaren vanligtvis utsätts för. Därför bör givaren alltid valideras i en välblandad hink så att absorbansmätningarna och referensmätningarna representerar samma vattenprovsammansättning.

5.A.6 Mall för valideringsrutiner

Denna mall på valideringsrutin kan användas som grund för att skräddarsy givarspecifika valideringsrutiner som saknas i den här boken. Mallen innehåller de komponenter som anses vara väsentliga för en valideringsrutin för kvalitetsbedömning av givaren och systematisk registrering av de metadata som den producerar.

5.A.6.1 Mall för valideringsrutin av givare

Version och datum

Överst i rutinen anges följande delar:

- Versionsnummer av rutin
- Senaste gången rutinen granskades och uppdaterades
- Nästa gång valideringsrutinen bör revideras

Syfte

Beskriv syftet med rutinen. Inkludera de kvalitetsmått för givardata som kan kvantifieras med stöd av valideringsrutinen (t.ex. bias, olinjäritet, offset, drift, lutning, precision, etc,).

Utrustning

Lista den utrustning som behövs för valideringen (t.ex. givare som ska valideras, referensgivare, glasvaror, blandningsutrustning). Inkludera lagringsplats om det är relevant.

Förbrukningsvaror

Lista kemikalier och förbrukningsutrustning som krävs för valideringen. Inkludera förvaringsplats om det är relevant. Beskriv eventuella förberedelser som krävs.

Utförande

Beskriv valideringen på ett stegvis sätt. Var tydlig och detaljerad. Nyanställda kan oftast bidra med insikter och kan kontrollera att valideringsproceduren är tillräckligt detaljerad och välstrukturerad.

(1) Beredning av lösning A.
(2) …

Dataanalys

Beskriv hur de registrerade valideringsdata ska bearbetas för att slutföra valideringen. Ange hur mätvärdena för datakvaliteten från givaren ska beräknas. Ange var och hur de resulterande mätvärdena och kvalitetsmåtten för givardata ska lagras.

5.A.6.2 Kontextuell information

Utöver själva valideringsproceduren kan viss bakgrundsinformation läggas till för att komplettera förståelsen av givarens syfte och praktiska användning. Listan nedan ger exempel på värdefull bakgrundsinformation:

- *Syfte med mätningen*: Beskrivning av vad givardata används till, t.ex. styrning eller övervakning av en viss process.
- *Nödvändig noggrannhet:* Vilken är den nödvändiga eller önskade noggrannheten för att uppnå mätändamålet.
- *Mätprincip*: Enkel teknisk beskrivning för att ge en intuitiv förståelse för mätprincipen för den validerade givaren.
- *Mätstörande faktorer*: Beskrivning och uppskattningar av effekterna av vanliga mätstörande faktorer som är relevanta i praktiken.
- *Metadata* för valideringsförfarandet: Information som beskriver valideringsproceduren, t.ex. version, ursprunglig utvecklare, kontakter för assistans (givarföretag/ingenjör på plats), förväntad tidsåtgång för utförande.
- *Givarens placering*: Ett foto eller en illustration är bra för att instruera ny personal om placeringen av givaren och tillhörande utrustning.
- *Lagring av metadata*: Instruktion om vilka metadata som produceras genom valideringsproceduren och som ska lagras, och på vilken plats.
- *Vanliga fel och felsökning*: Fallgropar och kända problem med specifika givare, platser och så vidare kan belysas här. Lösningar för att hantera vanliga fallgropar: Denna del kan sammanfatta instrumentteknikernas erfarenheter och kontinuerligt fyllas på med dokumentation av lokala givarerfarenheter.
- *Support och service*: Länkar och produktnummer för reservdelar/förbrukningsvaror, information om hur man får extra support från givarens underhållssystem.

BILAGA 5.B: VALIDERING AV VATTENFLÖDES-, NIVÅ-, TRYCK- OCH TEMPERATURGIVARE

5.B.1 Inledning

I praktiken så finns följande valideringstekniker för flödes, nivå, tryck och temperaturgivare:

Primär validering: Primär validering är den grundläggande elektroniska verifieringen av att ett instrument fungerar och mäter inom normala toleranser. Denna kontroll är normalt inbyggt i instrumentet och matas ut från instrumentet med jämna mellanrum med hjälp av ett s.k. HART-protokoll (*Highway Addressable Remote Transducer – HART*). Intelligenta givare använder detta för att automatiskt varna för problem eller fel på instrumentet med hjälp av det standardiserade NAMUR-protokollet. Primär validering är enkel att utföra men har begränsad nytta förutom att säkerställa korrekt funktion på instrumentet och tar inte hänsyn till processpecifika faktorer såsom nedsmutsning.

Sekundär validering: Sekundär validering motsvarar den typ av validerings som tidigare beskrivits för pH- och syregivare, d.v.s. en jämförelsen av instrumentets mätvärde med en referensmätning. Till exempel för en elektromagnetisk flödesmätare med slutet rör jämförs flödesavläsningen med en flödesmätare med flödesmätare som har kalibrerats externt och kan spåras tillbaka till antingen nationella och/eller internationella standarder från en ackrediterad testanläggning. Sekundär validering är vanligtvis mer komplicerad att utföra, men kontrollerar instrumentets noggrannhet på plats. Här utvärderas både instrumentets korrekta funktion och eventuella problem med mätpunkten.

End-to-end-testing: Den här metoden kallas ibland även *sensor-to-screen testing* och innebär att man kontrollerar att givarens mätning överförs korrekt hela vägen till processens övervakningssystem. Detta är särskilt viktigt när man använder analoga eller pulsade utgångar via ett telemetrisystem. Detta beror på att skalningen på platsen för dataöverföringen måste stämma överens med skalningen i telemetrisystemets ände av processen för att förhindra fel på grund av skillnader i instrumentets skalning.

Kalibrering: Kalibrering är validering och justering av ett instrument mot en känd standard. Detta utförs ofta i en laboratoriemiljö. Exempelvis tas elektromagnetiska flödesmätare bort från sin mätplats och monteras på en flödeskalibreringsrigg. Mätaren jämförs med en känd standard och faktorer som är programmerade i instrumentet ändras för att säkerställa korrekt drift. Kalibreringen kontrollerar inte den faktiska mätinstallationen in situ. Vissa mätningar kan kalibreras på plats beroende på instrumentets funktionalitet och utförs med hjälp av en spårbar kalibreringslösning och protokoll i instrumentet för korrigering. Detta görs ofta med hjälp av standardprotokoll för laboratorier, t.ex. en enkel standardmetod.

5.B.2 Validering av flödesgivare

Metoden för validering av flödesmätning beror på vilken mätprincip som tillämpas. Det finns ett stort antal olika mättekniker som var och en kräver olika valideringsmetoder och därmed olika typer av kompetens. I det här avsnittet beskrivs ett urval av de vanligaste mätteknikerna som används på reningsverk.

5.B.2.1 Elektromagnetiska flödeshastighetsgivare

Elektromagnetiska flödesgivare är mätare som mäter flödet i ett slutet rör. De allra flesta av den här typen av flödesmätare används när röret är fullt och mäter endast ledande vätskor. I huvudsak mäter tekniken vätskans hastighet i ett fast rör.

Det finns mättekniker som mäter i delvis fulla rör, men dessa använder i huvudsak en teknik av typen yta-hastighet genom att mäta nivån i röret. I det här fallet måste både hastigheten och nivån kontrolleras.

Det finns potentiella störningar med denna mätteknik eftersom den förutsätter en stabil flödesprofil och därför kommer allt som får flödesprofilen att avvika från en standardhastighetsprofil att resultera

Figur 5.B.1 En kraftigt igensatt rörledning uppströms en elektromagnetisk flödesmätare.

i en avvikelse. Det är därför en elektromagnetisk flödesmätare normalt sett kräver raka rör, minst fem rördiametrar uppströms och två rördiametrar nedströms. Om det finns en påverkan, t.ex. en ventil eller en kraftig rörböj, krävs normalt 15 rördiametrar uppströms och 5 rördiametrar nedströms.

Så länge en elektromagnetisk flödesmätare är korrekt installerad är det relativt lite som påverkar själva givaren och i det här fallet bör en primär validering av instrumentelektroniken räcka som valideringsteknik tillsammans med ett periodiskt end-to-end-test. Risken med igensättning av mätpunkten bör dock tas i beaktan, och bör styra behovet av sekundär validering.

Figur 5.B.1 visar påväxt omedelbart uppströms en elektromagnetisk flödesmätare där rörets diameter har minskat från 300 till 125 mm på vissa ställen under en 3-årsperiod. Om denna grad av nedsmutsning var linjär (vilket den inte var) skulle flödesmätaren registrera ett värde som var 5,76 gånger högre än det verkliga flödesmätarvärdet. I princip antog flödesmätaren att den hade en yta på 0,283 m^2 (ytan på en 300 mm elektromagnetisk flödesmätare), men i verkligheten var den reducerad till 0,05 m^2, vilket skulle öka hastigheten med 5,76 gånger.

Om det finns en risk för igensättning kommer en sekundära valideringen krävas för att detektera avvikelsen. Det här är dock ett extremt exempel som skulle kunna upptäckas relativt enkelt, men i verkligheten kan förändringar i flödesmönstret vara mycket mer subtila och svåra att upptäcka utan en sekundär validering.

Flödesmätning med flygtid (*time-of-flight*) används ofta som en sekundär valideringsteknik eftersom den är känsligare för igensättning och ett signalfel genereras normalt om igensättning börjar bli ett problem.

Trots att risken för igensättning är låg så bör sekundär validering genomföras regelbundet eftersom det ändå finns en risk för avdrift som kan påverka mätnoggrannheten. Detta är särskilt viktigt på större flödesmätare där mätfel troligtvis har stor påverkan på verksamheten.

5.B.2.2 Flödesmätning i öppen kanal

Flöde i öppna kanaler är i grunden en nivåbaserad mätteknik och därför kan nivåmätningen verifieras med en enkel nivåmätningsteknik, antingen med hjälp av en portabel nivåmätare eller tumstock.

Figur 5.B.2 Översvämmad, förvrängd och allmänt rörig mätpunkt för flöde.

Mätteknikens noggrannhet är också beroende av att den primära mätanordningen som normalt sett är en ränna eller ett utlopp från en tank damm, fungerar korrekt. För dessa typer av konstruktioner finns det några grundläggande tekniker som kan användas för att kontrollera att de fungerar korrekt:

- Visar den primära mätanordningen några tecken på deformation? Är dammluckan böjd?
- Finns det något hinder i eller i närheten av rännan eller överfallet? Störs flödesmönstret överhuvudtaget?
- Har mätanordningen fritt utlopp?

Figur 5.B.2 visar ett klassiskt exempel på flera mättekniska problem med nivåmätning. Den nedersta dammluckan en betydande deformation som på grund av dammluckans utåtgående vinkel skulle bidra till en underskattning av flödet om detta vore det enda problemet med mätanordningen. Det är dock inte det enda problemet, för flödet nedströms dammluckan har inte fritt utlopp utan flödar delvis tillbaka. Detta leder till överskattning av flödesmätningen. Utöver dessa problem finns det också ett returrör som skapar svallvågor och skum i mätkammaren, vilket påverkar ultraljudsflödesmätningen negativt.

Figur 5.B.3 illustrerar däremot en perfekt flödesprofil med fritt utflöde, vilket framgår av det spjutformade mönstret vid utloppet från rännan.

Figur 5.B.3 Ränna med fritt utlopp utan störningar.

Det som är viktigt med alla mätanordningar (rännor och utlopp) är att de är rena och inte påverkas av hinder (inklusive allt som kan förorena rännan), att de är hela och inte översvämmas av någon anledning.

Utöver detta ska den sekundära mätningen som mäter flödet genom att mäta nivån vara pålitlig och och fri från vibrationer eller rörelser. Om en ultraljudsnivåmätare används ska den skyddas från solljus eftersom det kan orsaka en förstärkning av mätningen ('solar gain') genom att temperaturen höjs, vilket påverkar ljudets hastighet.

Att den primära mätanordningen är fri från dessa störningar kan kontrolleras visuellt på några minuter utan verktyg eller hjälpmedel.

När det gäller att kontrollera noggrannheten på själva nivåmätningen, oavsett om det är ultraljud, radar eller någon annan nivåmätningsteknik, är den enklaste metoden att helt enkelt använda en tumstock och jämföra med nivågivarens mätning. Detta är inte lika exakt som att använda ett vattenpass i kombination med nivåmätning, men en tumstock är oftast tillräckligt noggrann för en snabb validering. Ett annat alternativ är att använda en kontrollplatta som förkalibreras till en given höjd och sedan använda en standardplatta för att kontrollera mätnoggrannheten (om man använder vanlig beröringsfri teknik).

För beröringsfria tekniker, som utgör majoriteten av de mätningar som utförs, så får operatören en direkt kontroll på hur långt givarens nivåmätning har drivit, som sedan kan nyttjas för att avgöra om kalibrering krävs. Ett alternativ är, såsom nämnts tidigare, att istället använda en referensmätning med antingen en flödesslinga med ackrediterad elektromagnetisk flödesmätare eller en certifierad mätning av yta-hastigheten (*area-to-velocity unit*).

5.B.2.3 Flödesgivare med area-hastighetsmätning

Flödesmätning med s.k. area-hastighet (*area-velocity*) skiljer sig från andra flödesmätningstekniker genom att det finns två obekanta variabler i flödesekvation. Den första är vattennivån som är den enda variabeln i mätningens delberäkning av arean och vattenhastigheten. Nivån är enkel att validera (som tidigare beskrivits), men själva hastighetsmätningen är svårare att validera på plats. Den kan kontrolleras med hjälp av tillsats av färg eller salt i vattenflödet, men det är för omständligt och används sällan av instrumenttekniker i praktiken. För smidig kalibrering på plats skulle någon alternativ referensmätningsteknik krävas för hastighetsmätningen, t.ex. en hastighetspistol.

Avsaknaden av en lämplig valideringsteknik för flödesmätning med area-hastighetsmetoden är ett olöst problem.

5.B.2.4 Flödesgivare med flygtidsmätning

Flödesmätning med flygtid har länge använts för att verifiera att flödesmätningen fungerar i ett slutet rör. En elektromagnetisk flödesmätning baseras på att mäta flödeshastigheten i det slutna röret. I praktiken finns det bara ett fåtal tekniker som kan användas för att kontrollera mätarens noggrannhet.

Ett exempel på en valideringsteknik visas i Figur 5.B.4. Denna teknik baseras på en hastighetsmätare som validerar hastigheten i rörledningen. Nackdelen med denna teknik är att en anslutningspunkt behöver monteras på rörledningen som kan vara relativt kostsamt. Om rörledningen är under tryck krävs dessutom ett trycksatt införingsverktyg för att föra in mätaren i röret på ett säkert sätt.

5.B.2.5 Andra tekniker för validering av flödesmätare

Det finns många andra tekniker för flödesmätning, men principen för validering av flödesmätningens noggrannhet bygger på de grundläggande principerna för validering:

- En primär validering krävs för att säkerställa att givaren fungerar korrekt.
- En sekundär validering verifierar på ett spårbart sätt att givaren mäter korrekt.
- Observation och okulär kontroll för att säkerställa att de fysiska installationerna (överfall etc) är hela och funktionella.
- Ett end-to-end-test för att säkerställa att skalning och signalöverföring är korrekt.

Figur 5.B.4 Flödesmätning med flygtid på ett 900 mm rör som verifieras av en instickshastighetsmätare (förgrunden).

5.B.3 Validering av nivågivare

Valideringsmetoderna för nivågivare är snarlika de metoder som används för validering av flöde i öppna kanaler eftersom båda teknikerna kräver validering av vattenhöjden, även om det beror mycket på vilken nivåteknik som används. Om det är en beröringsfri nivåmätning baseras valideringen på mätningen av luftspalten mellan givaren och vattennivån. Det finns både vissa fysiska protokoll och vissa mätprotokoll att ta hänsyn till.

De fysiska protokollen är baserade på installationens tillförlitlighet. Figur 5.B.5 visar en nivåmätning i en bräddvattentank. I detta fall har den ultraljudsgivare som används placerats på en gångjärnsförsedd konsol. Med slitage på gångjärnet uppstår en märkbar sänkning i mätningen, vilket naturligtvis påverkar mätnoggrannheten. Den här typen av installation är känslig för detta och måste antingen åtgärdas eller beaktas i mätosäkerheten.

I en ideal värld bör konsolerna vara stagade på minst två ställen på två olika punkter, t.ex. över en kanal. I exemplet i Figur 5.B.5 är det lämpligt med ett trepunkts stagningssystem eller ett låssystem för att säkerställa stabiliteten hos nivåmätningen. Specifikt för ultraljudsnivåmätning

Figur 5.B.5 Felaktig mätning orsakad av slitna konsoler.

Figur 5.B.6 Nivåmätning för flöde med radar (svart givare), ultraljud (blå givare med solskydd) och tryck (metallstång på höger sida) framför en ränna som fungerar som den primära mätanordningen.

bör man kontrollera att vätskenivån inte kan komma upp i nivå med mätzonen genom att installera nivåmätningen minst 350 mm ovanför överbyggnadsnivån. Om det dessutom krävs en hög noggrannhetsnivå bör ultraljudsgivaren även skyddas mot solinstrålning. Ett bra exempel på en korrekt flödesmätningsinstallation visas i Figur 5.B.6 som mäter både indirekt (via ultraljuds- och radarnivåmätning) och även genom tryckmätning.

När tryckmätning används för att mäta nivån krävs en validering av nivån. Detta kan göras med en nivåmätning eller med en indirekt metod som t.ex. validering av luftspalten mellan givaren och vattenytan, men det kan också ske med en referenstryckmätning via ett rörsystem som görs separat från huvudmätningen.

I korthet så kan följande rutiner användas för att validera tryckmätningar:

- Veckovis kontroll av att givaren är ren och att botten och ytan under givaren ytan är fri från hinder och skräp.
- Kvartalsvis kontroll av stag och upphängningsanordningar samt och en *end-to-end*-validering för att säkerställa en korrekt signalöverföring.
- Årlig kontroll av:
  - Givarens grundfunktion (primär validering).
  - Sekundär validering av själva nivåmätningen, helst med en spårbar mätteknik.
  - Om signalerna kommunicera analogt bör denna överföring kontrolleras årligen med hjälp av ett kalibrerat analogt referensinstrument. Detta bör även inkludera ett end-to-end-test som en del av valideringen.

5.B.4 Validering av tryckgivare

Tryckmätning är en vanlig mätning på flera platser i vattensystemet i både distribution, dricksvatten- och avloppsvattenrening. Varje applikation har valideringar som kan vara unika för den specifika tillämpningen. Det finns dock några kontroller som är generellt tillämpbara för alla tryckmätningsapplikationer. Det är viktigt att med en hög säkerhetsnivå, särskilt för trycksatta applikationer. Även lågtryckssystem kan vara farliga. Instrumentteknikern bör för säkerställa att mätpunkten isoleras från alla trycksystem som kan innebära en risk innan några kontroller påbörjas.

Figur 5.B.7 Tryckgivare med keramiskt membran (med tillstånd av Vega Control Systems).

Det finns det några grundläggande fysiska kontroller, som gäller för alla tryckmätningar, som bör utföras:

- Kapslingens och givarens visuella skick.
- Tryckmembranets skick - kontrollera om det finns revor eller tecken på slitage.
- Kabelgenomföringens täthet - den får inte läcka.
- Kondens i höljet - det får inte finnas någon.
- Kontrollera anslutningarna till transmitterenheten.
- Kontrollera tätningen - den får inte trycka på membranet.

Figur 5.B.7 visar en tryckgivare med ett keramiskt membran med fördelen att vara motståndskraftigt mot de flesta problem som mindre motståndskraftiga membran kan drabbas av.

Det är bra att förstå hur en tryckgivare fungerar. I princip alla tryckgivare är differenstrycksmätare eftersom avläsningarna sker i förhållande till ett referenstryck. De tre primära mätmetoderna på reningsverk är:

- Absolut tryck – trycket i förhållande till vakuum.
- Övertryck – trycket i förhållande till atmosfärstrycket.
- Differentialtryck – trycket i förhållande till ett annat, sekundärt tryck i processen.

Tryckgivare med membran har ersatt traditionella bälggivare i många reningsverksapplikationer. Membranen är platta eller koncentriskt korrugerade metall- eller keramikskivor som böjs vid olika tryckförändringar. När membranet böjs mäts böjningen med olika metoder, t.ex. kapacitivt, induktivt, piezoelektriskt, piezoresistivt eller optiskt. Böjningen ändrar den associerade elektriska egenskapen (t.ex. avståndet mellan plattorna i en kondensator, piezoelektrisk respons, loopreluktans eller optisk transmittans), som omvandlas till en tryckavläsning i instrumentet.

Membrantätningssystem består av en tryckgivare, en membrantätning och en påfyllningsvätska. De är antingen direktmonterade eller anslutna via kapillärslangar. Membrantätningen utgör en fysisk barriär mellan processen och givaren och överför även trycket genom påfyllningsvätskan i kapillärröret till givarelementet.

När en tryckivare installeras i ett system med snabba tryckförändringar eller tryckslag bör även en s.k. snubber (*snubber*) installeras för att skydda givarelementet från skador. En snubber fungerar som en stötdämpare. Den fungerar genom att en kolv drivs upp mot en öppning under snabba tryckförändringar, vilket stryper flödet och därmed dämpar de tryckstörningar som når själva tryckmembranet. Genom att ändra kolvdiametern kan snubbern dimensioneras för att passa olika viskositeter och nivåer av tryckslag.

Tryckavlastare av typen porösa skivor rekommenderas inte för reningsverksapplikationer eftersom de lätt täpps till, men de fungerar däremot i renvattenapplikationer.

De tryckavkännande elementen är vanligtvis tillverkade av rostfritt stål, nickellegering (Hastelloy) eller keramik. Om det inte finns lämpliga material för den process man vill mäta i, bör en membrantätning övervägas. Ett membrantätningssystem ger en kemisk barriär samtidigt som det överför tryck till avkänningselementet via en vätska. Typiska fyllningsvätskor är glycerin, silikon och halokarbon.

Mätelelementet ska vara åtkomligt för underhåll, med membrantätningar och kapillärrör som krävs för att montera givaren på en höjd där man kan utföra underhåll på den. Avstängningsventiler

bör finnas för att underlätta service av givaren. Om givaren är monterad högre eller lägre än processanslutningen måste en offset tillämpas på mätningen för att kompensera för höjdskillnaden. Flänsade givare bör användas för processer som är benägna att bli nedsmutsade. För applikationer med vätskor som innehåller fasta partiklar kan spolning och/eller membranisolering behövas. Membran för fasta applikationer finns tillgängliga med nästan alla tänkbara processanslutningar; anslutningarna bör vara minst 25 mm (1 in) i diameter. Membranflänsar finns tillgängliga med integrerad spolning, anslutningar eller membranförlängningar för att förlänga avkänningsmembranet in i processen.

Vid installation av tryckgivare bör det skapas förutsättningar för att validera givaren med en extern tryckmätning, antingen genom att givaren kan demonteras och valideras offline, eller genom ett förgreningsrör så att en extra extern givare kan monteras. Detta är i linje med de grundläggande valideringsprinciperna som tidigare beskrivits.

5.B.5 Validering av temperaturgivare

Temperaturgivare kalibreras sällan när de ingår som en del av en annan mätning, t.ex. i flödesgivare där det finns ett visst inslag av temperaturkompensation. Det fall där temperaturmätning ändå bör valideras är då temperaturmätningen påverkar en utsläppsvillkorad flödesmätning såsom utgående flöde på reningsverket. Det kan även vara nödvändigt med validering om temperaturmätningen påverkar en kritisk process såsom biogasproduktion.

Vid validering av temperaturgivare är det bra att ha en grundläggande förståelse för själva mätprincipen.

Det finns flera olika sätt att mäta temperatur i en online-mätapplikation, men det är tre metoder som är vanliga på reningsverk:

- Termoelement
- Resistansgivare (*Resistance temperature detector – RTD*)
- Termistorer

En annan metod, som dock är ovanlig på reningsverk, är infraröd (IR) mätning som även är en beröringsfri mätmetod. Infraröd temperaturmätning används mest av handhållna enheter och är ett exempel på en extern mätning som kan användas för att validera stationära temperaturgivare. I kommande avsnitt beskrivs dock de tre metoderna som är vanligast på reningverk.

5.B.5.1 Termoelement

Termoelement drivs av den s.k. Seebeck-effekten som innebär att en elektrisk ström orsakas av uppvärmning och värmeöverföring mellan två olika metaller. Den genererade spänningen är en icke-linjär funktion av metallernas temperatur och sammansättning.

Alla kontaktställen mellan olika metaller inducerar en Seebeckeffekt, även om vissa metallkombinationer ger en högre spänning per grad än andra. Därför måste man förhindra eller kompensera för andra övergångar än de som finns vid själva kontaktstället.

Exempelvis kan termoelementets förlängningstrådar vara tillverkade av samma metall som termoelementet eller så kan de hållas vid en känd temperatur så att hänsyn kan tas till oönskad spänning. Alla termoelementkretsar har minst två korsningar: en vid mätpunkten och den andra för att få två kopparkablar att ansluta till en voltmeter, som mäter spänningen som orsakas av själva kontaktstället.

Två kompensationsmetoder används för sekundära anslutningar: *softwire* och *hardwire*. Vid softwire-kompensering avslutas termoelementtrådarna vid ett isotermiskt block, som håller de två avslutningarna vid samma temperatur. Denna temperatur mäts av en resistansgivare, termistor eller integrerad krets. Givarens resistans mäts, omvandlas till referenstemperaturen och omvandlas sedan till en referensspänning. Denna spänning subtraheras från den som uppmäts av voltmätaren

för att erhålla termoelementets spänning, som omvandlas till den uppmätta temperaturen. Flera termoelement med samma referenstemperatur kan användas i samma mätenheten för att minska kostnaden för omvandling per termoelement.

Vid trådbunden kompensation genereras istället en spänning för att driva referensspänningen så att den blir lika med 0°C (32°F) vilket även kallas för den elektroniska ispunkten.

5.B.5.2 Resistansgivare

Resistensgivare fungerar enligt principen att det elektriska motståndet, resistiviteten, hos en metall förändras med temperaturen. Genom att mäta resistansen vid en viss punkt kan temperaturen bestämmas. Detta är den vanligaste typen av temperaturmätning att använda när en annan typ av mätning behöver temperaturkorrigeras. Ofta används en PT100-givare som i grunden är en typ av resistansgivare (se Figur 5.B.8).

En resistensgivare består av en tunn tråd eller film som bärs upp av ett substrat med motsvarande termiska utvidgningsegenskaper. Det tunna metallelementet är nödvändigt för att få tillräckligt hög resistans (normalt ca 100 Ω vid 0°C [32°F]) och med stöd för att vara tillräckligt robust för processmätning. För att mäta resistansförändringen vid den aktuella punkten måste först ledningsmotståndet från instrumentkablarna elimineras. En metod för att göra detta är att använda en konstant strömkälla och mäta spänningsändringen över elementet.

Denna teknik kräver en fyrtrådsledning eller en tretrådsledning och en s.k. Wheatstone-brygga (*wheatstone-bridge*), vilket ger ett visst fel om inte bryggan är korrigerad för den temperatur som mäts. Det kan uppstå ett s.k. självuppvärmningsfel om en ström leds genom det resistiva elementet.

Den värme som då avges av elementet kommer att påverka temperaturmätningen.

Figur 5.B.8 Flödesmätning med en separat RTD-temperaturgivare för temperaturkompensation, skyddad från solinstrålning för ökad noggrannhet.

5.B.5.3 Termistorer

Termistorer är halvledare vars resistivitet ändras som svar på temperaturförändringar. De flesta termistorers resistans minskar när temperaturen ökar.

Deras temperaturkoefficienter är högre än för resistansgivare där ett typiskt termistorelement har en resistans på 5000 Ω vid 25°C (77°F). Detta gör att fel i ledningsmotståndet är försumbara märkbara, och vanligtvis används ledningar med tre trådar. Självuppvärmningsfel kan dock vara betydande på grund av termistorns lilla massa.

Integrerade kretskortsgivare är halvledare vars resistivitet ökar när temperaturen stiger, men som i övrigt liknar termistorer. Dessa givare används inte i någon större omfattning på reningsverk, men i stället när temperaturkompensationen sker internt i en givare (t.ex. temperaturkompensationen i en ultraljudsgivare).

5.B.5.4 Installation

Alla temperaturgivare behöver skydd innan de används i en process. Termoelement, resistensgivare och termistorer installeras vanligen i ett skyddsrör som monteras så det kommer i kontakt med själva processen. Ett skyddsrör är ett tunnväggigt metallrör som är försett med ett lock i processändan så att det kan tas bort utan att processen störs. Givarna är fjäderbelastade för att det ska ha kontakt med skyddsröret för att minimera tidsfördröjningar i temperaturförändringar. Dessutom så smörjs temperaturgivaren oftast in en speciell pasta/smörja som förenklar värmeöverföringen från skyddsröret till givaren. Det är viktigt att se till att rätt pasta används för det givan temperaturområdet.

I vissa applikationer (t.ex. högtrycksapplikationer) bör man inte mäta i varken processvätskan eller genom en rörledning. Då monteras istället temperaturgivaren externt utan skyddsrör, men man måste i så fall ta hänsyn till högre mätfel och bristande responstid, särskilt om temperaturmätningen används för reglering.

I alla applikationer ska givaren skyddas från korrosiva eller eroderande miljöer eftersom fysiska förändringar i givaren gör att den förlorar sin kalibrering.

Termoelementgivare kan också förlora sin kalibrering om trådarna glödgas vid höga temperaturer. Glödgning sker när en stor temperaturförändring plötsligt inträffar (t.ex. när enheten flyttas från en varm process till en kall omgivningstemperatur). Det kan också inträffa på grund av vibrationer eller om ledningarna har installerats under spänning.

5.B.5.5 Validering och kalibrering av temperaturgivare

I praktiken kalibreras sällan temperaturgivare, om inte en givares noggrannhet ifrågasätts eller att temperaturmätningen krävs för processtyrning.

Men hur valideras en temperaturmätning? I ett laboratorium går det att jämföra med en referenstermometer alternativt med mätning i en bägare fylld med is och nästan fylld med kallt vatten. I fält däremot så används antingen en referensgivare, t.ex. en IR-termometer eller en kalibrerad resistensgivare, eller så använder man en kalibrerad spårbar temperaturkälla, t.ex. ett kalibreringsblock för temperatur. Kalibreringsblocket kan sedan skickas iväg för kalibrering varje år för att säkerställa att det fungerar korrekt. Dessa kalibreringsverktyg är dyra, men kan vara motiverade när korrekt temperaturmätning är kritiskt. Vid validering av temperaturmätning i rötkammare måste man dock vara försiktig och se till att de kalibreringsblock som används har rätt certifiering för farliga områden, s.k. EX-klassning.

doi: 10.2166/9781789065107_0161

Kapitel 6

Dataanalytiska metoder för kvalitetsbedömning av givardata

D. Aguado[1], J. Alferes[2], F. Arteaga[3], L. Belia[4], J. B. Copp[5], L. Corominas[6,7], F. Corona[8], A. Ferrer[9], H. Haimi[10,11], P. Kazemi[12], Q. H. Le[13], I. Miletic[14], M. Mulas[15], A. Robles[16], M. V. Ruano[16], S. Russo[17], O. Samuelsson[18], J. P. Steyer[19], K. Villez[20], E. I. P. Volcke[13], M. J. Wade[21,22] och J. Zambrano[23]

[1]CALAGUA – Unidad Mixta UV-UPV, Institut Universitari d'Investigació d'Enginyeria de l'Aigua i Medi Ambient – IIAMA, Universitat Politècnica de València, València, Spanien
[2]VITO – WaterClimateHub, Ostend, Belgien
[3]Universidad Católica de Valéncia, València, Spanien
[4]Primodal, Inc., Hamilton, ON, Kanada
[5]Primodal, Inc, Hamilton, ON, Kanada
[6]Katalanska institutet för vattenforskning (ICRA-CERCA), Girona, Spanien
[7]Universitetet i Girona, Girona, Spanien
[8]Högskolan för kemiteknik, Aalto-universitetet, Esbo, Finland
[9]Multivariate Statistical Engineering Research Group, Universitat Politècnica de València, València, Spanien
[10]FCG Finnish Consulting Group Ltd, Helsingfors, Finland
[11]Tekniska högskolan, Aalto-universitetet, Esbo, Finland
[12]Avdelningen för ingenjörsvetenskap och fysik, Universitat Rovira i Virgili, Tarragona, Spanien
[13]BioCo Research Group, Institutionen för grön kemi och teknik, Ghent University, Gent, Belgien
[14]inCTRL Solutions Inc, Dundas, ON, Kanada
[15]Institutionen för teknik inom teleinformatik, Federal University of Ceará, Fortaleza, Brasilien
[16]CALAGUA – Unidad Mixta UV-UPV, Departament d'Enginyeria Química, Universitat de València, Burjassot, Spanien
[17]EcoVision Lab, Fotogrammetri och fjärranalys, ETH Zürich, Schweiz
[18]IVL Svenska Miljöinstitutet, Stockholm, Sverige
[19]INRAE, LBE, Univ Montpellier, Narbonne, Frankrike
[20]Oak Ridge National Laboratory, Oak Ridge, TN, USA
[21]UK Health Security Agency, Data Analytics & Surveillance, Canary Wharf, London, Storbritannien
[22]Tekniska högskolan, Newcastle University, Cassie Building, Newcastle, Storbritannien
[23]The MathWorks AB, Knarrarnäsgatan 7, Kista 164 40, SE, Sverige

Nyckelord: Datadrivna metoder, databehandling, datavisualisering, kvalitetsbedömning, redundans, mekanistiska metoder.

Detta kapitel ger en omfattande översikt av de viktigaste datadrivna metoderna för kvalitetsbedömning av givardata som kan vara ett stöd för drift och processoptimering. Begreppet datadrivna metoder används här som ett samlingsnamn för de beräkningsmetoder och algoritmer (maskininlärning,

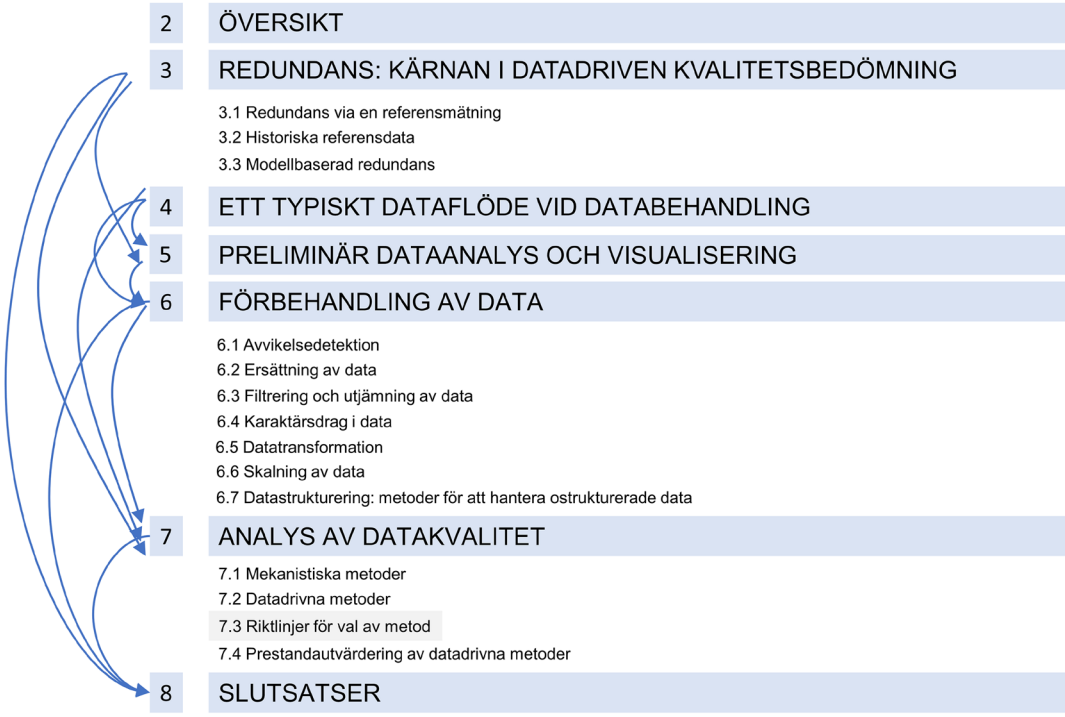
© 2025 IWAP. Detta är ett Open Access-bokkapitel som distribueras enligt villkoren i Creative Commons Attribution License (CC BY-NC-ND 4.0) som tillåter kopiering och vidaredistribution för icke-kommersiella ändamål utan derivat, förutsatt att arbetet citeras korrekt (https://creativecommons.org/licenses/by-nc-nd/4.0/). Kapitlet är från boken *Bästa möjliga data från givare i avloppsreningsprocesser*, Kris Villez, Daniel Aguado, Janelcy Alferes, Queralt Plana, Maria Victoria Ruano, Oscar Samuelsson (redaktörer).

mekanistiska modeller etc.) som på olika sätt kan användas för att tolka data. Kapitlet introducerar begreppet redundanta mätningar som ligger till grund för själva kvalitetsbedömningen i de flesta metoderna.

Kapitlet går också igenom det typiska dataflödet som består av en inledande analys, förbehandling av data och metodspecifika beräkningssteg. Vart och ett av dessa tre delsteg beskrivs i tre separata avsnitt. Utöver detta så ges även riktlinjer för att välja den lämpligaste metoden och de viktigaste prestandamåtten som används för att utvärdera dem samt en förklaring av hur man samlar in metadata utifrån datadrivna valideringsmetoder.

6.1 LÄSANVISNINGAR

Huvudsyftet med detta kapitel är introducera de mest populära algoritmerna/verktygen för datadriven kvalitetsbedömning av givardata samt val och inställning av dessa metoder utifrån ett reningsverksperspektiv. Figuren nedan visar kapitlets upplägg och beroenden mellan olika avsnitt som ett stöd för att underlätta läsningen. Det gråmarkerade avsnittet är särskilt användbart för att välja metod och upplägg av dataflöde (*data pipeline*).



6.2 ÖVERSIKT

Fördelarna med högupplösta data för drift och optimering på reningsverk blir alltmer tydliga. En effekt av detta är att fler och fler givare installeras. Det är dock vanligt att datakvaliteten förbises under beslutsprocessen med följden att de mest grundläggande frågorna om datakvalitet förblir obesvarade:

- Kontrolleras datainsamlingsprocedurerna (hämtas rätt data)?
- Kontrolleras kvaliteten på laboratoriedata?

- Kontrolleras kvaliteten på givardata?
- Är datakvalitetsförfaranden (datadrivna eller manuella) tillämpliga, på plats och används de?
- Är datakvaliteten ett problem?
- När data används, görs även datakvaliteten tillgänglig för användaren?

I takt med att datamängden ökar i både storlek och komplexitet har det blivit svårt att identifiera problem i tid, antingen manuellt eller med hjälp av enkla regler som kanske var tillräckliga tidigare när givarantalet var begränsat. För att kunna dra nytta av de investeringar som krävs för installation och underhåll av instrument är det nödvändigt att undvika insamling av data av undermålig kvalitet. Grundläggande feldetektionsmetoder kan dock misslyckas med att skilja mellan verkliga fel och typiska processvariationer på grund av en begränsad förmåga att anpassa sig till förändringar i normala datamönster. Dessutom kan förseningar i upptäckten av fel vara problematiska om data används för beslutsstöd (t.ex. i algoritmer för automatisk styrning eller manuella driftåtgärder).

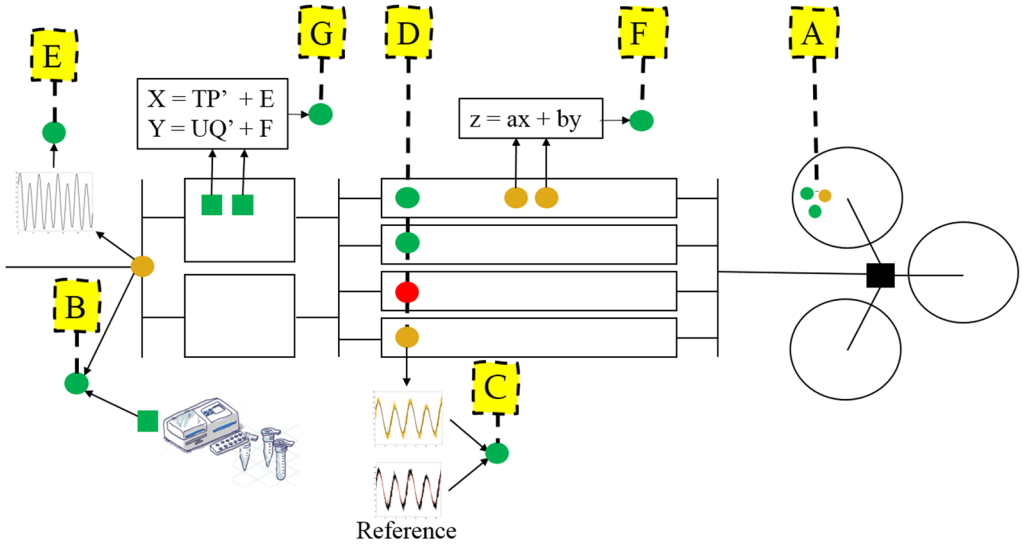
Kvalitetsbedömning av givardata blir en alltmer central uppgift för att kunna uppnå intelligent drift och optimering av reningsverk. Korrekt insamling av metadata är dock av största vikt för att framgångsrikt kunna tillämpa datadrivna metoder och för att verkligen förbättra driften på ett smart sätt. I det här kapitlet beskrivs de principer som ligger till grund för grundläggande och avancerade tekniker för bedömning av datakvalitet. Praktiska kriterier för val av teknik sammanfattas i en uppsättning allmänna riktlinjer för val av metoder för detektering av avvikande händelser och givarfel. De specifika syftena med detta kapitel är att:

- Introducera begreppet redundans som kärnan i datadriven kvalitetsbedömning.
- Definiera det typiska dataflödet för kvalitetsbedömning av data samt och belysa vikten av detta.
- Betona vikten av att titta på, och visualisera data för att bedöma av datakvaliteten.
- Beskriv de viktigaste teknikerna för förbehandling av data för att erhålla data som är lämpliga för efterföljande användning av datadrivna metoder.
- Introducera de mest populära datadrivna metoderna för kvalitetsbedömning av algoritmdata, inklusive både mekanistiska och empiriska metoder.
- Ge användbara riktlinjer för val av datadrivna metoder och dess metadatakrav, vilka indata som behövs och hur man tolkar resultaten av varje teknik.
- Beskriv de viktigaste prestandamåtten för att utvärdera datadrivna metoder.

6.3 REDUNDANS: KÄRNAN I DATADRIVEN KVALITETSBEDÖMNING

Grunden vid bedömning av datakvalitet är att jämföra givarens mätning med ett referensvärde som är en korrekt uppskattning av det sanna värdet. Termen *accepterat referensvärde* används i International Organization for Standardization (ISO, 1994) för att betona att det inte alltid är möjligt att få en perfekt referensmätning, även om dess noggrannhet åtminstone bör kvantifieras och vara tillräckligt noggrann för mätningens syfte. Det finns många typer av referensvärden som ger redundant information om givarmätningen och som därför kan användas för valideringsändamål. I detta avsnitt beskrivs tre vanliga redundanta informationskällor som kan användas som referensvärden, med varierande grad av noggrannhet: (1) redundanta mätningar, (2) historiska referensdata och (3) modellbaserade referenser. Sju exempel ges i Avsnitt 6.3.1–6.3.3 för att visa hur referensvärden och redundans kan användas för att validera reningsverksdata. De sju exemplen (a)–(g) illustreras också i Figur 6.1.

Jämförelser mellan redundanta mätningar är nära relaterade till generering av residualer, vilket är det centrala steget i feldetektion och diagnostik. Därför är allmän kunskap inom det området också tillämpligt för givarvalidering som baseras på redundanta mätningar (se t.ex. Basseville, 1993, 2001; Gustafsson, 2000; Pimentel *et al.*, 2014).



Figur 6.1 Illustration som visar de olika typerna av redundans på reningsverk. Figuren visar en förenklad reningsprocess som består av två försedimenteringsbassänger, fyra luftningsbassänger och tre eftersedimenteringsbassänger. (a) Hårdvaruredundans; sensorer som ger identiska mätningar i tid och rum. (b) Redundanta mätningar; en givarmätning jämförs med laboratoriedata för samma mätning. (c) Typiska referensdata; en mätning jämförs med historiska referensdata för att bedöma hur processens medelvärde och varians följs. (d) Rumslig redundans; identiska sensorer i flera parallella luftningsbanor. (e) Kvalitativ/mönsterredundans; historiska trenddata för inflöde används för att bestämma brusprofilen för sensorn för inflöde. (f) Teoretiska referensmodeller; en teoretisk modell som beskriver förhållandet mellan mätvariabler används för att kontrollera att både sensorn och processen överensstämmer med förväntade trender. (g) Datadrivna modeller; en representation av en Partial Least Squares soft-sensor. Cirklar = mätningar, kvadrater = indirekta referensmätningar, mätning/sensor är grön = inom förväntade gränser, orange = mindre avvikelse, röd = fel/avvikelse.

6.3.1 Redundans via en referensmätning

Den vanligaste strategin för att bedöma datakvaliteten hos sensordata (t.ex. en givarmätning) är att jämföra dem med en duplikatmätning. Denna referensmätning kan göras med samma mätteknik eller enligt en helt annan mätteknik

(A) Hårdvaruredundans

Det enklaste sättet att skapa en referensmätning är att duplicera befintlig givare med ytterligare en givare av samma typ och på så sätt uppnå hårdvaruredundans (*hardware redundancy*) (Figur 6.1A). Detta kan användas för kritiska datasignaler eller för billiga och vanliga sensorer, t.ex. tryck och temperatur. Observera att hårdvaruredundans inte nödvändigtvis kan övervaka givardrift eller igensättning om detta sker samtidigt i båda sensorerna och därefter inte observeras. Mer information om aspekter relaterade till redundanta sensormätningar finns i Haimi *et al.* (2015); Jung (2015); Nyberg och Frisk (2020); Ohmura *et al* (2019); och Villez *et al* (2016).

(B) Redundanta mätningar

En givarvalideringen baseras vanligtvis på en jämförelse mellan givarens mätning och en alternativ typ av mätning av samma ämne, vilken typiskt mäts offline (Figur 6.1B). Detta förfarande exemplifierades i Kapitel 5 (Avsnitt 5.5.2). Förutom den traditionella jämförelsen av stickprover kan veckovisa och flödesproportionella labbvattenprover också användas för att bedöma 7-dygnsmedelvärdet för

motsvarande givare. Notera att givarens mätvärden i så fall behöver omvandlas till flödesproportionellt värden för att matcha det förväntade laboratorievattenprovet för en korrekt jämförelse. Fördelen med denna metod är ett mindre provtagningsfel än för ett stickprov.

Vissa ämnen kan mätas med olika mättekniker, var och en med sina egna för- och nackdelar. Ammonium kan t.ex. mätas med en spektrofotometrisk analysator eller med jonkänsliga givare, utöver laboratorieanalys. Valideringsrutiner för givare beskrevs i Kapitel 5 (Avsnitt 5.5).

6.3.2 Referensmätningar utifrån historiska data

I detta avsnitt beskrivs fyra typer av referensmätningar som innehåller redundant information när de samlas in och analyseras över tid, vilket här benämns som referensmätningar utifrån historiska data.

(C) Typiska normalvariationer som referensdata

Typiska dygns-, vecko- och säsongsmönster finns i de flesta inflödesdata (Figur 6.1C). Kunskap om dessa mönster kan användas för att fastställa ett processmedelvärde som beskriver den normala historiska referensvariationen. Ett exempel är att uppskatta variansen runt medelvärdet, vilket indikerar dess användbarhet och precision som referensvärde.

Den största utmaningen med historiska referensvärden är att avloppsvattnets inflöde och processer förändras över tid. De historiska referensdataseten kommer att bli mindre representativa för den aktuella situationen. Detta kan delvis åtgärdas genom att processrelaterade data korreleras med förändringar i inflödet. En förutsättning för alla historiska referenser är att det finns tillräckliga metadata som beskriver de historiska förhållandena för inflödet och processen. Dessutom bör givarstörningar och felaktiga avläsningar i historiska data identifieras och tas bort för att minska bruset i typiska referensdata.

Ytterligare exempel på hur historiska data kan användas som referens för benchmarking, typiskt som prestandamått, ges av Balmér och Hellström (2012), Longo *et al.* (2016) och Samuelsson *et al.* (2017).

(D) Spatial redundans

Ett reningsverk har vanligtvis flera parallella och seriekopplade processer. Detta kan användas för att jämföra likvärdiga mätningar och identifiera avvikelser (Figur 6.1D), som vi här kallar spatial (rumslig) redundans. Utmaningen är att bedöma den förväntade skillnaden mellan t.ex. parallella behandlingslinjer, som i praktiken kan bero på belastningen och anläggningens hydraulik. Dessutom kan mätningar från seriekopplade processer innehålla information om föregående processenhet. Spatial redundans har visat sig vara användbart för att validera flödesmätningar (Villez *et al.*, 2016) och för att upptäcka avvikande luftflöden (Carlsson & Zambrano, 2016).

(E) Kvalitativ redundans

Kvalitativ redundans är en datadriven typ av redundans, som nyttjar historisk processinformation som göms i givarsignalernas mönster, som referens för det förväntade signalbeteendet. Underliggande mönster som återspeglar dynamiken i inflödet, sensorbrus eller något specifikt processbeteende kan t.ex. observeras i många mätningar (Figur 6.1E). Mönstret behöver inte ha någon fysisk innebörd för att vara användbart, utan beskriver helt enkelt hur signalen normalt beter sig. Univariat signalfiltrering (Alferes & Vanrolleghem, 2016), slamsedimenteringsprofiler (Zambrano *et al.*, 2019) och automatisk processövervakning och -styrning (Thürlimann *et al.*, 2015; Wade *et al.*, 2005) är exempel som illustrerar den breda tillämpligheten av att utnyttja historiska mönster som redundant information.

(F) Teoretiska referensmodeller

I litteraturen finns referensvärden som antingen baseras på teoretiska beräkningar eller en sammanställning av omfattande dataset som anger en referens och ett typiskt eller förväntat

intervall (Figur 6.1F). Teoretiska referensvärden är vanligast för processhastigheter, som beskriver det förväntade förhållandet mellan flera mätningar. Till exempel kan en specifik respirations- eller denitrifikationshastighet som avviker kraftigt från de teoretiska antagandena tyda på en bias i de underliggande mätningarna, snarare än än en processtörning. Standardläroböcker med fokus på reningsverk innehåller ofta referensvärden, se till exempel Henze *et al.* (2002), Loosdrecht *et al.* (2016) och Metcalf och Eddy (2004).

6.3.3 Modellbaserad redundans

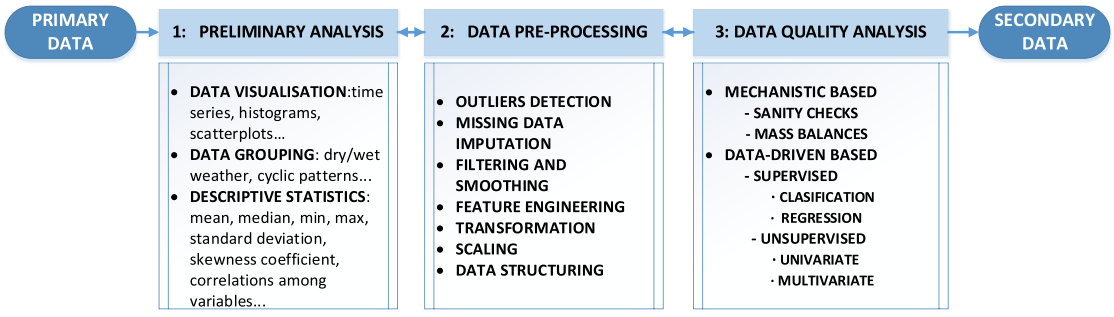
(G) Datadrivna prediktioner av referensmodeller

Det är även möjligt att nyttja modeller som specifikt tagits fram för att producera referensvärden för valideringsändamål. Det är här ingen skillnad på mekanistiska (fysikaliska modeller) eller black-box-modeller (rent datadrivna) eftersom båda modelltyperna kan producera en referensmätning genom sin prediktion, som i sin tur kan användas för givarvalidering (Figur 6.1G). En fördel med datadrivna referenser är att jämförelsen kan utföras automatiskt och i realtid när en modell väl har tagits fram. Konstruktionen av en sådan modell är fokus för kommande avsnitten i detta kapitel, som även innehåller många praktiska exempel.

6.4 ETT TYPISKT DATAFLÖDE VID DATABEHANDLING

Som tidigare framgått finns det ett behov av en strukturerad och, om möjligt, automatiserad databehandling. I denna rapport benämns en serie av databehandlingssteg för ett dataflöde (*data pipeline*). Ofta finns det en mängd olika dataflöden, var och en utformad för ett specifikt syfte. Det lämpligaste dataflödet för en viss datakälla bör väljas utifrån syftet med databehandlingen.

Ett dataflöde byggs upp genom att välja lämpliga bearbetningssteg bland delstegen i Figur 6.2, det vill säga (1) inledande analys, (2) förbehandling av data och (3) kvalitetsbedömning av data, uppdelat i metoder som bygger på fysikalisk förståelse och metoder som bygger på dataanalys och statistik (datadrivna metoder). Den inledande granskningen omfattar en rad olika metoder för att få en första förståelse av rådata, t.ex. datavisualisering (Avsnitt 6.5), dataaggregering och beskrivande statistik. Det andra steget omfattar olika förbehandlingstekniker (detektering av avvikelser, modifiering, interpolering av data, skalning, filtrering och utjämning, generering av funktioner och datastrukturering) beroende på egenskaper hos rådata och den slutliga metod som ska användas för att bedöma datakvaliteten. Det är t.ex. bara nödvändigt att ersätta förlorade eller saknade data om de nödvändiga datavärdena inte finns tillgängliga, eller att omvandla en variabel om den uppvisar en markant skev fördelning. Det sista steget i dataflödet består av en eller flera tekniker för bedömning av datakvalitet. Dessa tekniker grupperas i allmänhet i fysikaliska metoder: rimlighetskontroller



Figur 6.2 Steg som kan ingå i utvecklingen av ett dataflöde för bedömning av datakvalitet.

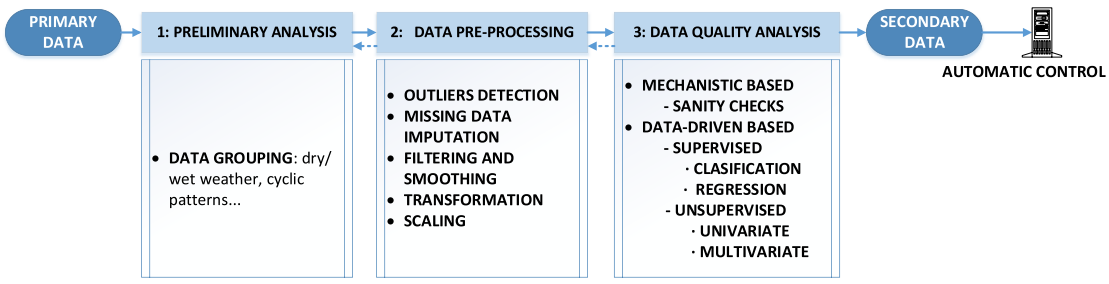
och massbalanser (Avsnitt 6.7.1), och empiriska metoder: övervakade (*supervised*) och oövervakade metoder (*unsupervised*) (Avsnitt 6.7.2).

Genomförandet av dessa tre steg kan kräva aktiva beslut av användaren (t.ex. vid visualisering av historiska data) eller ske helt automatiskt vid användning av data i realtid (t.ex. avvikelsedetektion). Dessutom behöver inte dessa steg utföras i serie, utan kan itereras på flera punkter i dataflödet. Exempelvis kan avvikande värden upptäckas i det sista steget med en metod för bedömning av datakvalitet, varefter denna information kan återföras till förbehandlingssteget, t.ex. för att fylla i luckorna med rimliga värden med en teknik för dataersättning. Dessutom krävs inte alltid alla steg som visas i Figur 6.2. Därför finns det inte ett enda dataflöde som är generiskt, utan dataflödet är fallspecifikt då det beror syftet med datakvalitetsbedömning: t.ex. automatisk kontroll, tillsyn, simulering och så vidare. Syftet avgör också om dataflödet tillåter mänsklig interaktion (t.ex. simulering) eller inte (t.ex. onlinekontroll). Figur 6.3 visar ett dataflöde som kan användas för automatisk styrning. Processens bearbetade signaldata kan användas som indata i styralgoritmen, antingen direkt (t.ex. syrekoncentrationen) eller indirekt via onlineuppskattning av en annan variabel eller parameter, t.ex. med hjälp av en datadriven modell (t.ex. artificiella neurala nätverk (ANN)).

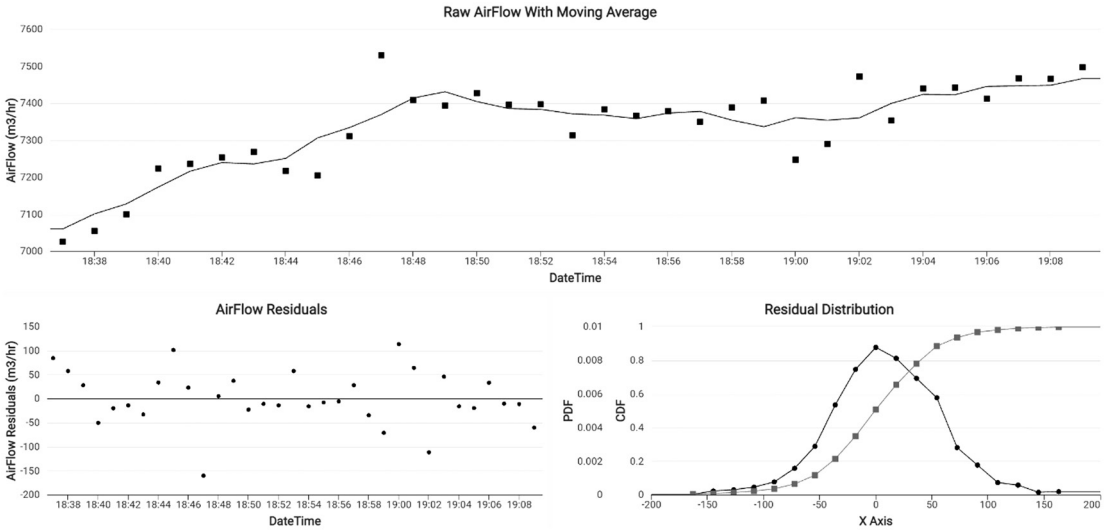
Det är få reningsverk som har implementerat automatiserad behandling av rådata i realtid såsom i Figur 6.2. I stället installeras och underhålls givare och data – både bra och dåliga – och registreras i något slags datahanteringssystem. I de flesta fall ingår inte ett databehandlingssteg vid insamlingstillfället. Istället förlitar sig dataflödet på enkel regelbaserad logik eller kriterier för dataanvändning för att kompensera för datavariabilitet och okända fel. Man måste vara medveten om att alla signaler kommer att ha variationer och fel (dvs. en signal är aldrig helt korrekt) och målet med all signalbehandling är att minimera effekterna av dessa problem när man använder data. Data av dålig kvalitet bör dock inte utelämnas utan istället förses med en annotering/etikett eller markeras (se även Kapitel 4).

Även om de flesta reningsverk inte befinner sig på denna nivå av datakvalitetsbedömning, kommer införandet av några av de steg som tidigare nämnts att driva på övergången till smart processdrift och optimering. I följande stycken beskrivs kortfattat några exempel på databehandling som för närvarande används i praktiken.

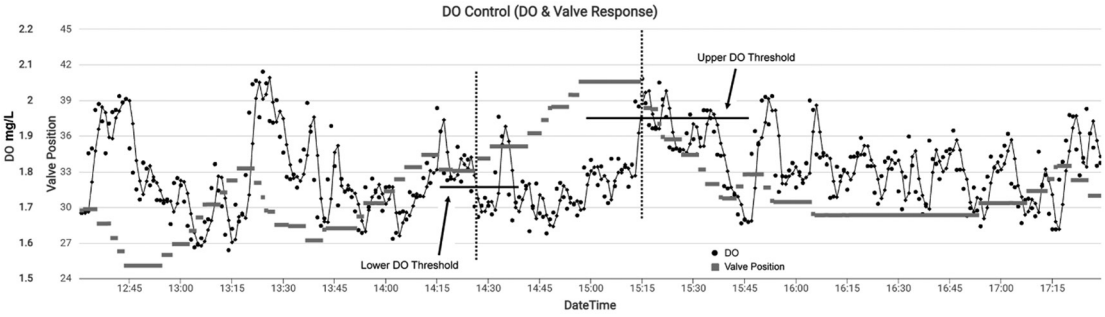
Figur 6.4 visar ett enkelt exempel på data från en luftflödessignal tillsammans med dess glidande 5-minutersmedelvärde (överst), de beräknade residualerna för samma period (nederst till vänster) och fördelningen av dessa residualer (nederst till höger). Detta exempel visar hur mycket enkla analystekniker (datavisualisering och beskrivande statistik tillsammans med ett förbehandlingssteg med filtrering) kan ge bra information om en signals variabilitet. I praktiken skulle denna analys utföras med ett större dataset under en längre tidsperiod. Här används en mindre datauppsättning i pedagogiskt syfte.



Figur 6.3 Steg som kan ingå i utvecklingen av en pipeline för bedömning av datakvalitet för automatiska kontrolländamål.



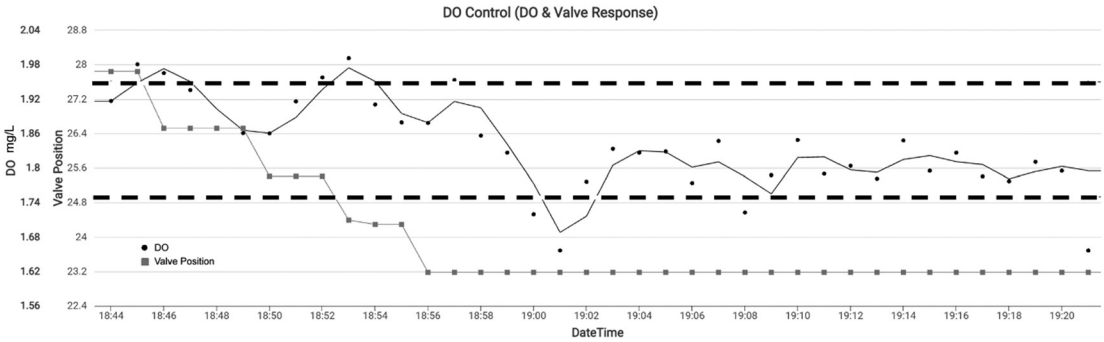
Figur 6.4 Enkel residualanalys av data från en luftflödessignal från ett reningsverk. Råsignal (överst, rutor) med 5-minuters glidande medelvärde (överst, linje). Beräknade residualer – glidande medelvärde minus råsignal (längst ned till vänster, prickar) tillsammans med kumulativa frekvensdiagram för dessa residualer (längst ned till höger, PDF [svarta prickar] och CDF [grå rutor]).



Figur 6.5 Illustrativt exempel på en syrehaltsregulator på ett reningsverk. Uppmätt syrehalt (svarta prickar), 2-min glidande medelvärde syrehalt (heldragen linje), ventilöppningsgrad (fyrkantiga prickar).

Figur 6.5 visar reglering av syrehalten på ett reningsverk där styrsignalen filtrerats för att jämna ut bruset. I det här exemplet har den uppmätta syrehalten (runda prickar) ett snabbt svar och tillhörande brus. I styrstrukturen beräknas ett 2-minuters glidande medelvärde (heldragen linje) och hårdkodade tröskelvärden definieras vid 1,75 (övre horisontella linjen) och 1,9 mg/l (nedre horisontella linjen). Ventilens respons på det fluktuerande glidande medelvärdet visar hur ventilen öppnas eller stängs när det glidande medelvärdet överskrider de två tröskelvärdena (fyrkantiga prickar).

Figur 6.5 visar att om rådata användes utan någon bearbetning, skulle alla naturliga fel och variationer i signalen påverka det resulterande reglerbeteendet. Genom att medelvärdesbilda insignalen minskas variationerna något och effekterna av en felaktig datapunkt undviks. Detta illustrerar en specifik lösning där dataanvändningen avgör i vilken grad en råsignal (och i slutändan svaret) bör modifieras före användning.

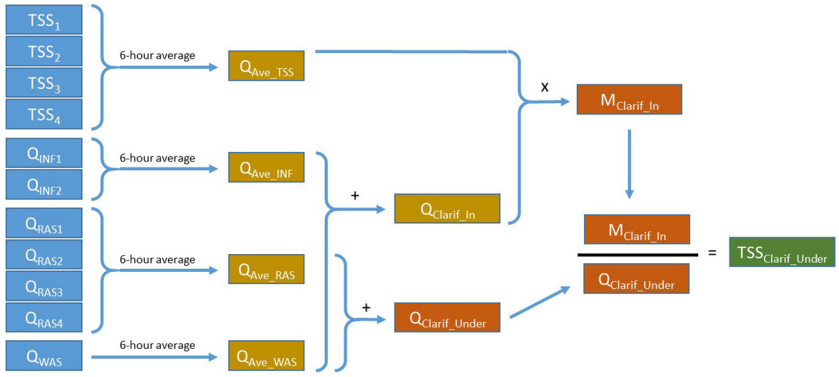


Figur 6.6 Illustrativt exempel på syrehaltsreglering med dödband och tidsbegränsat reglerbeteende (glidande syrehaltsmedelvärde (heldragen linje), uppmätt syrehalt (prickar), öppning av reglerventil (rutor med streckad linje).

Filtrering av en signal är ett vanligt sätt att jämna ut variationer, ta bort grova fel och minska brus, men det finns även andra metoder för att förbättra signalens lämplighet för ett visst ändamål. I Figur 6.6 visas två andra exempel från samma dataset. För det första visas användningen av ett dödbandsfilter där syrehalten tillåts variera mellan 1,75 och 1,9 mg/l utan korrigering i ventilöppningsgrad. Figur 6.6 visar också hur tidsstyrd dämpning där styrningen initieras först efter en definierad tidsperiod. Ventilens beteende i figuren (fyrkantig streckad linje) visar en stegvis respons, där styrningen initieras först efter en väntetid på minst 3 minuter. Det innebär att ventilen svarar, väntar i minst 3 minuter och sedan, om det behövs, svarar igen. Även om dessa begränsningar ofta används för att begränsa slitaget på ställdon och inget av dessa tillvägagångssätt i strikt mening innebär databehandling, är avsaknaden av styråtgärder i dessa två fall också ett sätt att bekräfta att variationer i data är kända och acceptabla eftersom antingen: (i) processen inte skulle påverkas nämnvärt av variationer inom detta intervall; eller (ii) signalvariationen ligger inom en rimlig nivå av förväntat fel/brus och mer information (dvs. ytterligare data) krävs för att säkerställa att förändringen i den övervakade variabeln motsvarar en tillräcklig förändring i processen.

Ett annat exempel på dämpning av en styråtgärd är när styrstrukturen kräver att ett tröskelvärde överskrids under en viss tid. Ett av skälen till att införa fördröjningar i styrningen är att skydda ställdonen från att kontinuerligt slås på och av. I det här fallet, om signalen överskrider tröskelvärdet under kortare tid än den definierade perioden, antas det att överträdelsen berodde på fel/brus eller alternativt att det var osannolikt att överträdelsen skulle påverka processen under den aktuella tidsramen. Tanken är alltså att man vet att ett genombrott endast har inträffat om tröskelvärdet överskrids under en definierad period. Det här är inte signalbehandling i sig, men det illustrerar ett vanligt tillvägagångssätt för att begränsa effekterna av variabilitet, fel och/eller brus i en signal.

I det föregående exemplet användes rullande 2-minutersmedelvärden för att styra luftningsventilernas rörelser på grund av den snabba systemdynamik som uppstår vid förändrade syrehaltsförhållanden, men andra processer i reningsverket har mycket längre tidskonstanter och kräver annan databehandling. Ett exempel är en beräkning av fasta partiklar som ingår i en slamåldersregulator (Figur 6.7) i en aktivslamanläggning med fyra parallella linjer. I det här fallet används 6-timmars medelvärden för inflödet ($Q_{\mathrm{Ave_INF}}$), TSS-koncentration (total suspenderad substans) ($Q_{\mathrm{Ave_TSS}}$), överskottsslamuttaget ($Q_{\mathrm{Ave_WAS}}$) och returslamflödet ($Q_{\mathrm{Ave_RAS}}$) för att uppskatta koncentrationen i sedimenteringsbassängens slamuttag ($TSS_{\mathrm{Clarif_Under}}$) och slutligen överskottsslammets flöde ($Q_{\mathrm{Clarif_Under}}$) för att upprätthålla en given slamhalt. Att beräkna medelvärdet för signalerna under 6 timmar är mer än tillräckligt här, eftersom tidskonstanten för den här processen är i storleksordningen dagar eller veckor och koncentrationen av fasta ämnen i flödet varierar naturligt under dagen.



Figur 6.7 Exempel på en massbalans som ingår i en slamåldersreglering med glidande medelvärde som ett filtreringsförbehandlingssteg av de råa givarsignalerna som visas i blått.

All databehandling (medelvärdesbildning, filtrering etc.) är applikationsspecifik och bör optimeras för att filtrera bort oönskad variabilitet samtidigt som den tillämpliga systemdynamiken fångas upp. Användningsfallen här har illustrerat att graden med vilken en signal bearbetas vanligtvis är en funktion av signalens egenskaper och den tidskonstant som är förknippad med dess användning.

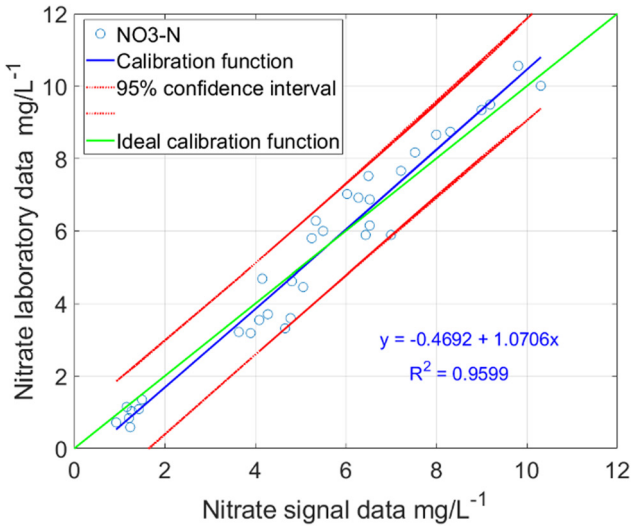
6.5 PRELIMINÄR DATAANALYS OCH VISUALISERING

Det finns många olika sätt att visa och beskriva historiska data och realtidsdata. Valet är applikationsspecifikt och bör väljas för att förmedla den information som krävs på ett kortfattat och lämpligt sätt. Typiska sätt som har använts historiskt inkluderar:

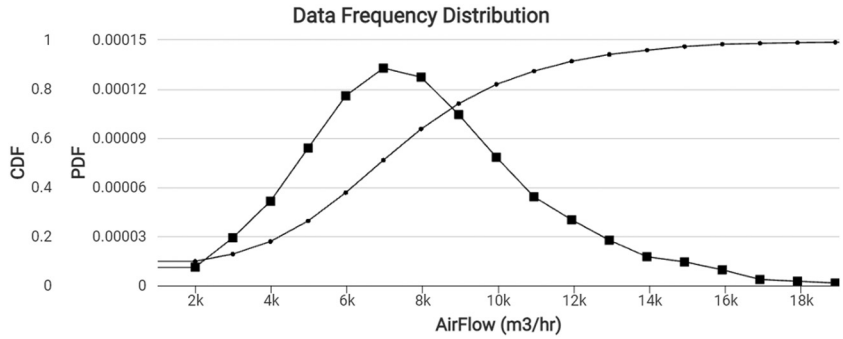
- Statistiska tabeller (som kan innehålla antal datapunkter, medelvärden, medianer, standardavvikelser, percentiler etc.)
- Stapeldiagram
- Cirkeldiagram
- Histogram
- Boxplottar
- Heat maps etc.

Oavsett tillämpning bör metoden för datavisualisering väljas så att den passar den specifika tillämpningen. Konstruktörer och operatörer kan behöva olika information från samma dataset, så den metod som används för att visualisera data för dessa grupper kan mycket väl vara olika. Nyckeln till ett framgångsrikt val är att dataanvändaren tydligt förstår den information som visas och kan extrahera den information de vill ha på ett effektivt sätt. Tidsserier, spridningsdiagram och statistiska tabeller används ofta, men det finns även andra metoder som kan vara mer lämpliga. Det bästa valet kommer att skräddarsys för applikationen och den grupp av personer som ska ta del av data.

Kontrollen av datakvaliteten inleds ofta med att plotta givardata som en funktion av motsvarande laboratoriemätningar (se även Avsnitt 6.3). Detta kan enkelt visualiseras med hjälp av ett spridningsdiagram. Överensstämmelsen mellan signaldata och laboratoriedata blir större när fler data överensstämmer med en 45° rak linje som passerar genom origo. Bedömningen kan kvantifieras genom en linjär regression, kontroll av lutningen (ju närmare 1 desto bättre överensstämmelse), skärningen med y-axeln vid origo (ju närmare 0 desto bättre) och determinationskoefficienten (R^2) för anpassningen (ju närmare 1 desto bättre). Figur 6.8 visar ett spridningsdiagram över signaldata från en nitratgivare mot laboratoriedata. Figuren visar tydligt att nitratgivaren överensstämde mycket bra



Figur 6.8 Spridningsdiagram för att bedöma kvaliteten på signaldata från en nitratgivare genom redundans via referensmätning från laboratoriedata.

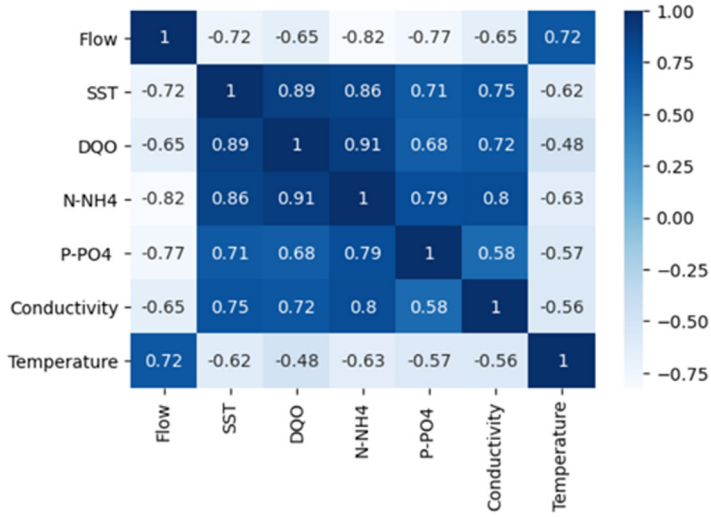


Figur 6.9 Distributionsdiagram för rådata på luftflödesmätningar för att bedöma variansen i historiska data (PDF [svarta fyrkanter] och CDF [svarta cirklar]).

med laboratoriedata eftersom den resulterande regressionen är nästan identisk med 45°-linjen (grön linje). Enkla diagram som detta används i praktiken och kan förmedla mycket information på ett mycket effektivt sätt.

När historiska data granskas och analyseras kan andra former av visualisering vara mer lämpliga. Figur 6.9 visar till exempel en frekvensfördelningsdiagram som visar fördelningen av vissa råa luftflödesdata. Detta kan t.ex. användas för att verifiera att data är normalfördelade eller för att snabbt bedöma om datafördelningen är skev.

En annan visualiseringsteknik är en s.k. *heat map*. Med den här tekniken kan värdena i en datatabell (primär- eller sekundärdata, se Kapitel 2) visualiseras som olika färger i två dimensioner. Färgvariationen i kartan underlättar tolkningen av data genom att ge visuella ledtrådar om data: hur de är grupperade eller hur de varierar över kartan. Det är ett snabbt och effektivt sätt att förmedla informationen i data till det mänskliga ögat, vilket gör det lättare att upptäcka intressanta mönster.



Figur 6.10 Heat map över korrelationerna mellan signaldata från inflödet till en anläggning för resursåtervinning av avloppsvatten.

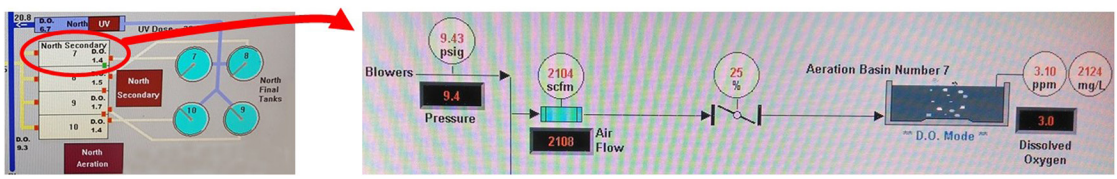
Som exempel visar Figur 6.10 en heat map över korrelationerna mellan de signaler som registrerats vid inflödet till ett reningsverk. Man kan snabbt se en negativ korrelation mellan flödet och föroreningarna i avloppsvattnet och en stark positiv korrelation mellan föroreningarna.

Idag är datavisualisering på de flesta behandlingsanläggningar ganska enkel på grund av behovet av att visa signaldata tydligt, tillförlitligt och kortfattat för att möjliggöra korrekta och snabba beslut från olika personer inom driften.

För det mesta uppnås detta med SCADA-skärmar (*Supervisory Control and Data Acquisition*) som visar det senaste värdet för viktiga processvariabler som text, ofta i kombination med siffror som visar signalvärdet som en funktion av tiden.

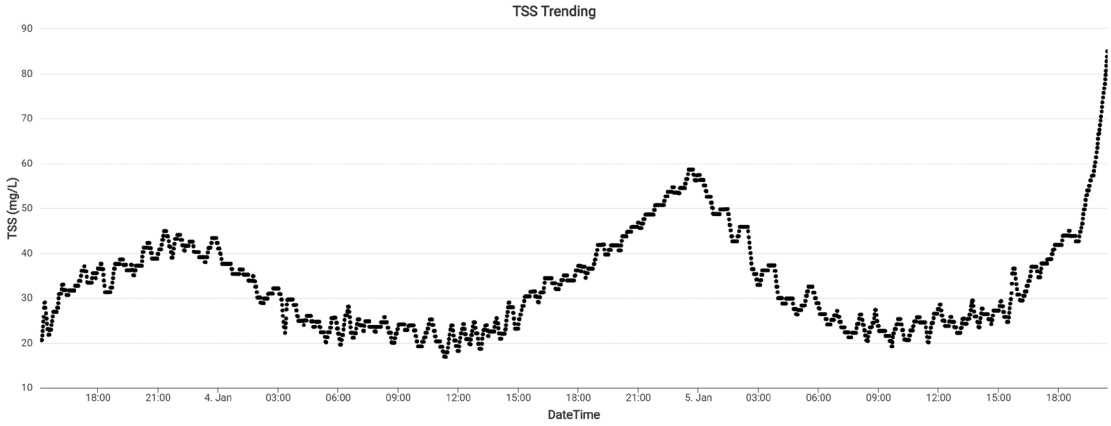
SCADA-skärmar brukar visa anläggningen i ett grafiskt användargränssnitt med flera lager, där användaren kan gå in på ett visst område i anläggningen och hämta specifik information. Man kan t.ex. vilja kontrollera luftflödet till en viss bassäng (Figur 6.11). Detta kan göras från rotskärmen genom att klicka på aktivslamobjektet, sedan på den aktuella processlinjen och sedan på själva luftningsobjektet. Vid varje steg i denna process visas olika variabler som är relaterade till den aktuella frågenivån. Det kan t.ex. räcka med att klicka på aktivslamobjektet för att visa aktuella syrehalter. Varje SCADA-applikation är anläggningsspecifik, men den grundläggande funktionaliteten brukar följa den här arkitekturen.

Själva skärmarna varierar från anläggning till anläggning och kan ha olika former och konfigurationer med en mängd olika rattar, mätare och datafält (Figur 6.12).



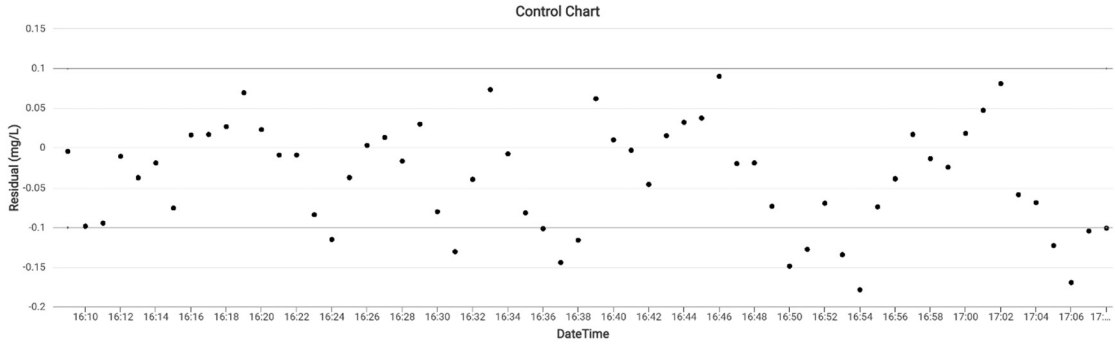
Figur 6.11 Exempel på SCADA-skärmdump.

Figur 6.12 Exempel på SCADA-skärm.



Figur 6.13 Trenddiagram för utvecklingen av TSS i utflödet från ett reningsverk.

Alternativ för att visualisera datatrender är också vanligt, t.ex. för en variabel som mäter avloppsvattnets kvalitet (t.ex. TSS). Det kan vara intressant att se det aktuella värdet, men det aktuella värdet ger inget sammanhang med avseende på det förflutna eller den förväntade omedelbara framtiden för den variabeln. Beroende på användningsfallet kan ett diagram som visar en uppmätt variabel som en funktion av tiden ge detta sammanhang (Figur 6.13). Är det aktuella beteendet normalt eller onormalt? Om det är onormalt, kan trenden användas för att i förebyggande syfte varna personalen för ett överhängande problem och proaktivt signalera för en begränsande åtgärd.



Figur 6.14 Exempel på ett kontrolldiagram som tillämpas på residualer mellan givardata och laboratoriedata.

Även andra diagramalternativ har använts med framgång. Figur 6.14 visar ett exempel på ett styrdiagram där övre och nedre gränser definieras, och överskridandet av dessa gränser kan utlösa ett larm eller en varning (se Avsnitt 6.7.2 för mer information om denna typ av diagram).

Visningen av aktuella värden ger en ögonblicksbild av processtillståndet, medan tidsseriediagram effektivt kan förmedla den senaste processdynamiken och användas som en tidig varning om processtillståndet går åt fel håll.

6.6 FÖRBEHANDLING AV DATA

Givardata från reningsverk är ofta brusiga, kan sakna värden, innehåller extremvärden, värden utanför intervallet och kan ha en skev fördelningar med mycket outliers. Den information som data förmedlar om de underliggande processerna begränsas av dessa störningar och data är inte i en idealisk form för att analyseras med datadrivna metoder. Genom förbehandlingsfasen omvandlas eller modifieras de insamlade uppgifterna till en mer lämplig form för vidare analys, vilket gör det lättare att upptäcka de underliggande mönstren. Detta är ett viktigt steg för den senare tillämpningen datadriven analys av datakvalitet (t.ex. massbalanser, maskininlärning etc.). Faktum är att detta steg vara skillnaden mellan en framgångsrik och värdelös modell.

De viktigaste motiven för att förbehandla data är:

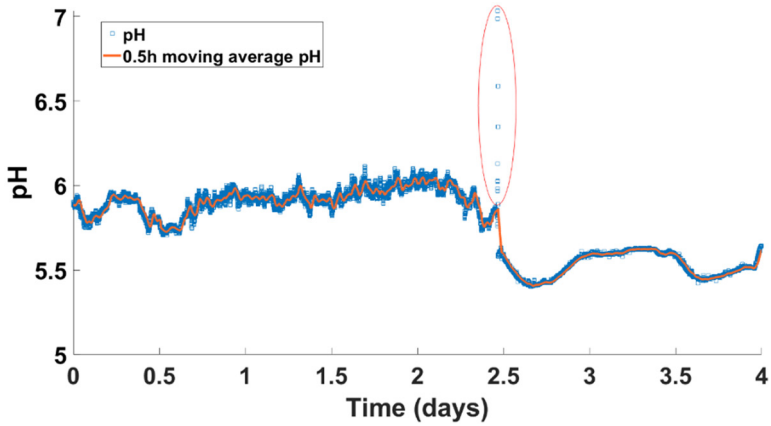
- Förbättra tolkningsbarheten av data och den information som utvinns ur dem.
- Att tillämpa metoder för bedömning av datakvalitet, som kanske inte accepterar data i sin ursprungliga form. Detta kan gälla för ostrukturerade data och dataset med förlorade data.
- För att uppnå bättre prestanda för algoritmerna. Exempelvis så konvergerar den s.k. *gradient descent*-metoden snabbare när alla variabler har en liknande skala.
- För att uppfylla viktiga modellbegränsningar eller antaganden.

Förbehandling av data omfattar många tekniker som tjänar ett eller flera av ovanstående syften. Vissa metoder används ofta (t.ex. normalisering och skalning) medan andra bara används i specifika situationer (t.ex. datastrukturering). Nästa avsnitt beskriver de mest populära teknikerna.

6.6.1 Avvikelsedetektion

6.6.1.1 Syfte

Syftet med avvikelsedetektion är att upptäcka observationer som avviker i jämförelse med resterande av observationerna i datasetet. Dessa icke-typiska observationer har ett ovanligt stort eller litet värde i förhållande till de andra mätvärdena i datasetet. Vanligtvis undersöks dessa avvikande observationer för att ta reda på vad som orsakade dem eftersom det kan bero på ett naturligt fenomen (t.ex. relaterat



Figur 6.15 pH-signaldata från en aktivslambassäng. Outliers upptäcks efter filtrering av rådata från en pH-mätning.

till reningsverkets dynamik), men det kan också vara en följd av ett fel eller en felaktig givare som därför bör uteslutas från datasetet inför vidare bearbetning.

6.6.1.2 Exempel på metoder för avvikelsedetektion

Outliers kan upptäckas på olika sätt:

- Genom visuell inspektion av de data som visas i ett trenddiagram, tillsammans med en teknisk tolkning av processen avseende förekomsten av det icke-typiska värdet.
- Genom statistiska metoder, t.ex. (a) genom att definiera nedre och övre gränser för deras detektion baserat på normala mätintervallet, eller statistiska gränser (*statistical process control*) såsom Shewhart-diagram eller och ett exponentiellt viktade glidande medelvärdet (EWMA) eller (b) med multivariata metoder.

Information om detekterade avvikande värden bör lagras genom att data förses med annoteringar (se Kapitel 4). Detta bör åtminstone omfatta den metod som använts för att upptäcka själva avvikelsen (t.ex. *EWMA*, *visuell inspektion*, lagrad som metod). Avvikande värden bör inte raderas från centraliserade databaser. Figur 6.15 visar ett exempel på detektering av avvikande värden efter ett förbehandlingssteg med filtrering med ett glidande medelvärde av givardata från en pH-givare.

6.6.2 Ersättning av data

6.6.2.1 Syfte

Ersättning av data (*data imputation*) innebär att frånvarande eller förlorade data (se Kapitel 2, Avsnitt 2.4.1.3) ersätts med ett värde som baseras på annan tillgänglig information. Denna tillgängliga information kan vara föregående och efterföljande värden från samma signal. Det kan också vara datapunkter i andra signaler som registrerats samtidigt, vilket gör att korskorrelation mellan variablerna kan utnyttjas.

Data kan saknas av olika orsaker: mätvärden som ligger utanför mätintervallet, alla observationer mäts inte av ekonomiska skäl, givarfel (Grung & Manne, 1998), kommunikationsfel, givarunderhåll, prover mäts inte vid rätt tid, data förloras på grund av allvarliga mätfel och givare med olika provtagningsperioder (Arteaga & Ferrer, 2002). Vid mätning och reglering av batchprocesser kan framtida processbeteende anses som frånvarande data (Nomikos & MacGregor, 1995).

Om man enbart analyserar de delar i datasetet där mätserier och data är kompletta, kan det leda till en betydande förlust av information.

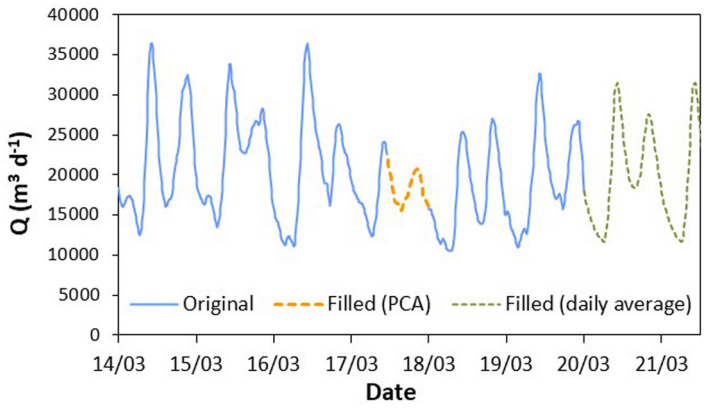
När mängden saknade data är mycket liten i förhållande till datasetets storlek (<5 %) kan det vara en rimligt att kassera ett fåtal observationer som innehåller saknade värden som inte påverkar analysen och endast leder till en liten informationsförlust. Ju större mängd data som saknas, desto större blir dock informationsförlusten, och det är detta som gör dataersättningsmetoderna så värdefulla. Dessutom är dessa metoder särskilt användbara när de data saknas i variabler som används som indata till processreglering.

6.6.2.2 Exempel på metoder för att ersätta data

I litteraturen har olika metoder föreslagits för att ersätta frånvarande eller saknade data. Valet av metod beror på flera faktorer, t.ex. datahålets längd (dvs. perioden med flera på varandra följande saknade), mönstret för de saknade data, mekanismen som ledde till att data saknades och syftet med dataanalysen. Mulder *et al.* (2018) föreslog fem ersättningsmetoder som kan användas för typiska data från reningsverk: (1) interpolering, (2) korskorrelation med andra tillgängliga variabler, (3) fylla i med motsvarande värden från variabelns genomsnittliga dagliga mönster, (4) upprepa det senaste värdet (eller värdet från föregående dag) och (5) fylla i med data från en matematisk modell (t.ex, från en inflödesgenerator för att fylla i luckorna i inflödesvariablerna, eller från en kalibrerad aktivslammodell (ASM) för att fylla i koncentrationerna i de biologiska reaktorerna eller i utflödet).

Som illustration används inkommande flöde i Figur 6.16 för att visa på möjligheten att fylla i data med hjälp av två ersättningsmetoder. Först genom korskorrelation med själva insamlade variabeln (under 2 veckor) med hjälp av en modell för principalkomponentanalys (PCA) (1), vilken tar hänsyn till autokorrelationsmönstret (dag 17 från 11:00, streckad orange linje). Metod två använder det genomsnittliga dagliga mönstret (3), beräknat med de två veckornas data (från dag 20, streckad grön linje). Mönstret för förlorade data beskriver vilka värden som är kända och vilka värden som är frånvarande eller saknas i datamatrisen X, tillsammans med sambandet, det vill säga det möjliga förhållandet mellan det okända mätvärdet och andra mätvärden vid samma tidpunkt.

Det mest problematiska fallet inträffar när en viss uppsättning saknas samtidigt (t.ex. strömbortfall för samtliga givare), eftersom detta gör det omöjligt att direkt uppskatta de parametrar som beskriver förhållandet mellan dessa variabler (Muteki *et al.*, 2005).



Figur 6.16 Inflöde till ett reningsverk under 1 vecka. Resultat av att fylla i saknade data baserat på en PCA-modell (dag 17 från 11:00, streckad orange linje) och baserat på det genomsnittliga dagliga mönstret (från dag 20, streckad grön linje).

Det är viktigt att förstå orsaken till de förlorade data. De statistiska metodernas egenskaper beror på orsaken och dess mekanismer som leder till förlorade data (Little & Rubin, 2022). Mekanismen till förlorade data påverkar hur mycket de ersatta värdena påverkar efterföljande dataanalytiska steg. Det finns tre möjliga fall:

(1) MCAR (*Missing completely at random*): Detta fall sker när det inte finns något samband mellan variablernas värden (observerade och saknade), eller sannolikheten för att de saknas. De saknade positionerna är bara ett slumpmässigt urval från datamatrisen. Detta innebär inte att mönstret i sig är slumpmässigt, utan snarare att saknade positioner inte beror på de verkliga underliggande värdena. Detta är ett brett antagande, men kan vara rimligt i vissa situationer. Det kan t.ex. vara praktiskt när vissa variabler är mycket dyra och endast mäts för en slumpmässig delmängd av dataposterna, medan de minst dyra variablerna mäts för alla dataposter (Allison, 2001). Detsamma gäller när data saknas på grund av schemalagda underhållsprotokoll för givare som inte är relaterade till processens funktion (Nelson, 2002), eller när det finns slumpmässiga givarfel eller slumpmässiga problem med överbelastning av data i kommunikationen med datalagringssystemen.

(2) MAR-fall (*Missing at random*): Detta fall innebär att sannolikheten för att en datapunkt saknas beror på andra värden som är tillgängliga, men inte på det okända värdet för den datapunkt som saknas (dvs. benägenheten för att en datapunkt saknas är inte relaterad till förlorade data, utan till andra uppgifter som faktiskt är tillgängliga). Detta fall är mindre restriktivt än MCAR. Observerade data innehåller den information som behövs för att skatta de data som saknas. MAR inträffar t.ex. när dataposter inte registreras när vissa andra processmätningar ligger utanför de säkra mätgränserna.

(3) NMAR-fall (*Not missing at random*) eller NI-fall (*Nonignorable*): Detta fall innebär att sannolikheten för att en datapunkt saknas beror på det icke observerade värdet för de datapunkter som saknas. Sannolikheten för att en datapunkt saknas är svår att utvärdera i detta fall eftersom den beror på de okända värdena för de data som saknas. Detta inträffar t.ex. i censurerade/avrundade data: när vissa mätningar ligger utanför detektionsgränsen (antingen över eller under). NMAR är problematiskt, eftersom fördelningen av saknade data för en uppsättning variabler inte följer fördelningen av de observerade mätningarna för samma uppsättning variabler i andra objekt. I detta fall krävs specialiserade metoder för att modellera mekanismen för saknade data för att få bra uppskattningar av de parametrar som är av intresse.

Ignorerbarhet (*ignorability*) innebär att vi inte behöver modellera mekanismen för saknade data som en del av skattningsprocessen. MCAR klassas som ignorerbart och MAR kan göras ignorerbart med lämplig analysmetod. I praktiken saknar vi (av naturliga skäl) kunskap om de förlorade värden och det går därför inte att bevisa något av de tidigare antagandena. Därför måste mekanismen för ett visst problem väljas subjektivt utifrån processkunskap. Den viktigaste frågan är att skilja mellan NMAR (NI-mekanism som är typisk för tillförlitlighets- och detektionsgränsanalyser) och ignorerbara mekanismer (MCAR och MAR) som är ett vanligt antagande i litteraturen. Följande avsnitt kommer att fokusera på MCAR och MAR.

I de följande avsnitten granskas flera statistiska metoder som föreslagits i litteraturen för att hantera saknade data, allt ifrån från borttagning av observationer med saknade data till avancerade ersättningsmetoder. Vissa av metoderna är ganska enkla och finns implementerade i vanliga statistikpaket, där man använder ad hoc metoder för att göra den ofullständiga datamängden fullständig (Schafer, 1997). Det finns även metoder som innebär avsevärda förbättringar jämfört med konventionella metoder, men som kräver specialkunskaper och programvara. Det är viktigt att notera att ingen av metoderna är perfekt och att den optimala lösningen på problemet med saknade data är att inte ha några förlorade data alls (Allison, 2001). Det finns alltså ett stort värde i att minimera förekomsten av frånvarande eller saknade data.

6.6.2.2.1 Fullständig datakassering

De flesta statistikpaket använder som standard att alla observationer med frånvarande eller saknade data kasseras, vilken kallas *complete case method – CC*. Det är en enkel metod, men har den stora nackdelen att mycket information förloras, särskilt i multivariata dataset, där ofullständiga datapunkter kan stå för en betydande del av hela datasetet. Informationsförlusten leder till minskad precision (vilket ökar osäkerheten i skattningarna) och bias när orsaken till databortfallet inte är slumpmässigt (MCAR) (Little & Rubin, 2002).

6.6.2.2.2 Ofullständig datakassering

Ett mindre restriktivt alternativ till CC-metoden för univariat (en signal) analys är att inkludera alla observationer där variablerna av intresse förekommer, vilken kallas *available case method – AC*. Med denna metod ändras antalet dataposter från variabel till variabel beroende på mönstret för saknade data. Detta kan dock bli problematiskt om databortfallet inte är slumpmässigt (MCAR).

Kim och Curry (1977) fann att AC var överlägset CC i simuleringar baserade på svagt korrelerade data, men detta är ett allvarligt problem med starkt korrelerade data, vilket noterades av Haitovsky (1968). Lopes och Menezes (2005) föreslår att korrelationsmatrisen korrigeras på samma sätt som i Higham (2002).

6.6.2.2.3 Imputation av data

Problemen med de två föregående metoderna (informationsförlust och bias) skapar ett behov av kompletterande metoder som på något fyller i luckorna med 'rimliga' värden för de data som saknas. Målet är generellt att fylla i de tomma positionerna i en ofullständig datamatris så att en fullständig matris kan erhållas. Ett komplett dataset öppnar upp för att använda fler dataanalysmetoder, jämfört med specialanpassade dataanalysmetoder som kan hantera förlorade data. Den metod som används för att ersätta de data som saknas bör följa variablernas interna korrelationsstruktur.

Det finns flera lättanvända metoder, t.ex. att interpolera värden för en variabel med hjälp av tidigare och efterföljande observationer, att beräkna medelvärden för variabeln vid mättillfället, eller att upprepa värden från tidigare dataposter.

Andra mer avancerade metoder bygger på att fylla i data från mer eller mindre komplexa modeller och utnyttja korrelationen med andra tillgängliga mätvariabler.

I de följande avsnitten beskrivs några av dessa metoder.

6.6.2.2.3.1 Metoder med enskilda imputeringar

En metod med en enstaka imputering ersätter ett (1) frånvarande eller saknat värde med ett (1) rimligt värde. Imputationer är oftast medelvärden eller ett slumpmässigt draget värde från en statistik fördelning som beskriver de databortfallet, vilket därmed även kräver en representativ statistik fördelning för databortfallet. Flera olika metoder för att ta fram en sådan fördelning beskrivs i detta avsnitt.

Ovillkorlig medelvärdesimputering: Varje förlorad datapunkt ersätts med medelvärdet av resterande data. Med denna metod kan skattningarna av storheter som inte är linjära i data (varianser, kovarianser eller percentiler) inte skattas korrekt (Haitovsky, 1968). Till exempel är varianser och kovarianser från de ersatta värdena lägre än i verkligheten, eftersom de imputerade värdena ligger i mitten av fördelningen. En annan nackdel med den här metoden är att den bara använder informationen för varje enskild variabel, dvs. den tar inte hänsyn till de insamlade uppgifternas multivariata karaktär och eventuella korskorrelationer. De följande metoderna löser detta problem på olika sätt.

Villkorlig medelvärdesimputering (Buck's metod): I denna metod (*Buck*, 1960) skattas först medelvärdet och kovariansmatrisen från data som hanteras med CC-metoden ovan. Sedan ersätts de förlorade värdena med skattningar från linjär regression med minstakvadrat-metoden. Det går även att använda andra regressionsmetoder såsom ridge regression (RR), principal component regression (PCR) eller partial least-squares (PLS) regression, särskilt när korrelationsmatrisen är s.k. illa-konditionerad vilket gör det svårt att beräkna dess invers.

Villkorlig medelvärdesimputering ger rimliga medelvärdesskattningar när data är normalfördelat. Men eftersom metoden inte tar hänsyn till variansen mellan olika variabler där mätningar för båda variablerna saknas, underskattar den varianserna och kovarianserna, även om denna underskattning är lägre än den som erhålls med den ovillkorliga medelvärdesimputering.

Imputering med stokastisk regression: Denna metod liknar villkorlig medelvärdesimputering, men metoden har en extra korrigering för att undvika systematiska underskattningar av variansen. Korrigering innebär att ett slumpmässigt värde adderas till den linjära skattningen, där det slumpmässiga värdet dras från en normalfördelning med medelvärdet noll och en skattad kovariansmatris för förlorade data, givet tillgängliga data.

Precis som i villkorlig medelvärdesimputering kan linjär regression ersättas av andra regressionsmetoder om kovariansmartisen är illa-konditionerad.

Imputering med Expectation maximization algorithm (EM): Tanken med EM-algoritmen är att uppskatta de parametervärden som gör att en antagen sannolikhetsfördelning för databortfallet stämmer bäst överens med tillgängliga data. Det är vanligt att anta en normalfördelning och de parametervärden, $\boldsymbol{\Theta}$, som då behöver uppskattas är medelvärde och kovarians för variablerna.

Om databortfallet orsakas enligt MAR eller MCAR så kan alla parametervärdena skattas som likelihood funktionen, $L(\boldsymbol{\Theta} \mid \mathbf{X}_{\text{obs}})$, vilken är proportionell mot sannolikheten för observerade data, $\mathbf{X}_{\text{obs}}$, generats från parameteruppsättningen, $\boldsymbol{\Theta}$:

$$L\left(\boldsymbol{\Theta} \mid \mathbf{X}_{\text{obs}}\right) \propto P\left(\mathbf{X}_{\text{obs}} \mid \boldsymbol{\Theta}\right)$$

se (Schafer, 1997) för en detaljerad förklaring.

EM-algoritmen (Little & Rubin, 2002) består av följande steg:

- Ersätt förlorade värden genom initiala uppskattningar.
- Uppskatta parametervärdena med hjälp av det färdiga datasetet.
- (E-steg) Generera ny värden som ersättning för de saknade värdena utifrån den aktuella parameteruppsättningen.
- (M-steg) Gör en ny skattning av parametrarna utifrån den fullständiga datauppsättningen i föregående E-steget.
- Upprepa stegen E och M till dess att parametervärdena konvergerar.

Det bör noteras att steg E kan drabbas av problemet med en illa-konditionerad kovariansmatris, vilket då kräver alternativa regressionsmetoder. Se Schafer (1997) för en effektiv implementering av denna algoritm.

Begränsningar i metoderna med enskilda imputeringar: När någon av ovan imputeringsmetoder har använts så är datamatrisen komplett, vilket underlättar efterföljande dataanalys. Vid en första anblick kan detta verka tillfredsställande, men det kan dock finns problem med de ersatta värdena, särskilt om kovariansstrukturen mellan olika variabler har påverkats. Imputering av medelvärden gör till exempel att de skattade varianserna och kovarianserna underskattas. Imputering av predikterade värden från regressionsmodeller tenderar däremot att överskatta korrelationerna. Dessa osäkerheter innebär att alla imputerade dataset bör hanteras restriktivt för beslutsstöd då statistiska mått såsom standardfel, *p-värden* kan vara missvisande och skilja sig från de sanna värdena.

6.6.2.2.3.2 Multipel imputering

Multipel imputering (MI) innebär att M värden skapas för varje förlorat värde, vilket ger en uppsättning datapunkter som återspeglar variationen i signaldata (Rubin, 1987). Med denna metod genereras M kompletta dataset, och vart och ett av dem kan analyseras med statistiska standardtekniker för att uppskatta medelvärde, varians etc.

MI bygger på tre huvudantaganden: en sannolikhetsmodell för det kompletta datasetet (observerade och saknade värden), ett antagande (beskrivet som en sannolikhetsfördelning) om

osäkerheten i de skattade parametervärdena för sannolikhetsmodellen samt att databortfallet kan antas vara försumbart.

Den vanligaste MI-metoden är 'data augmentation – DA' algoritmen (Tanner & Wong, 1987). Denna metod är iterativ som iterativt ersätter data och drar slutsatser om de okända parametrarna, men till skillnad från EM-algoritmen så sker detta stokastiskt. DA börjar med en slumpmässig imputering med antagna parametervärden (vanligtvis från EM-algoritmen) och erhåller sedan nya parametrar från en s.k. posteriorfördelning, från de observerade och imputerade data. Iterationerna, *t,* växlar mellan två steg:

I-steget (imputation) slumpar fram ersättningsvärden från fördelningen som bestämts av observerade data och parametervärden i föregående iteration.

P-steg (posterior fördelning) slumpar fram en ny parameteruppsättning från posteriorfördelningen som baseras på observerade och ersatta data.

Detta alternerande förfarande skapar en s.k. Markov-kedja som till slut konvergerar till en sannolikhetsfördelning (Schafer, 1997).

I praktiken bör Markovkedjans längd vara tillräckligt lång för att säkerställa konvergens och utan att påverkas av initialvärdena. Eftersom algoritmen konvergerar mot en sannolikhetsfördelning och inte en enda uppsättning värden, gör detta det ganska svårt att avgöra om konvergens faktiskt har uppnåtts (Allison, 2001). EM-algoritmens konvergenshastighet är dock en användbar indikation på konvergenshastigheten för DA.

Resultatet av en DA ger en enda, färdig datauppsättning. För att tillämpa MI kan en separat kedja av iterationer för var och en av de önskade dataseten (parallell beräkning). Alternativt så görs en sekventiell beräkning med en enda lång Markovkedja av DA-cykler. I det sistnämnda fallet behöver dock imputeringarna hanteras på ett särskilt sätt för att kunna bedöma datasetens oberoende.

6.6.2.2.3.3 Principalkomponentanalys för multivariata dataset

Datadrivna modeller baserade på principalkomponentanalys (PCA) är särskilt användbara för multivariata dataset där variablerna har stor korrelation. En PCA-modell reducerar de ursprungliga K variablerna till A s.k. latenta variabler eller huvudkomponenter (*principal components – PC*). Poängen med detta är att de nya variablerna är betydligt färre (i allmänhet med $A \ll K$), linjärt oberoende men ändå förklarar större delen av variansen av de ursprungliga variablerna.

Eftersom PCA är ett effektivt sätt att beskriva korrelationsstrukturen i ett dataset, är PCA även ett bra verktyg för att ersätta databortfall. Det finns två vanliga algoritmer för att bygga en PCA-modell från ett ofullständigt dataset: (a) den s.k. NIPALS-algoritmen (Schafer, 1997; Wold *et al,* 1983) och (b) den iterativa algoritmen (IA) (Arteaga & Ferrer, 2005; Walczak & Massart, 2001).

IA (Walczak & Massart, 2001) börjar med att fylla i de saknade uppgifterna med initiala skattningar (nollor eller andra imputeringar som t.ex. medelvärdet av de kända värdena i varje kolumn). Detta ger ett rekonstruerat dataset från vilken en PCA-modell kan anpassas. Därefter ersätts de ursprungliga saknade uppgifterna med deras prediktioner från denna PCA-modell, vilket ger ett nytt rekonstruerat dataset från vilken en ny PCA-modell kan anpassas. Denna process itereras tills de ersättningsvärdena för databortfallet konvergerar.

Utöver dessa två metoder så föreslog Folch-Fortuny ett regressionsbaserat ramverk (Arteaga *et al.*, 2020; Folch-Fortuny *et al.*, 2015) som är bättre än NIPALS och IA. Den föreslagna metoden är en anpassning av IA där ramverket baseras på den s.k. KDR-metoden (*known data regression*), som är likvärdig med EM-algoritmen. Övriga beräkningsmetoder i ramverket inkluderar olika regulariserade regressionsmetoder (*Principal component regression – PCR* och *Partial Least Squares – PLS*). Ramverket för regressionsbaserade metoder kan ses som en anpassning av IA där sättet att ersätta förlorade data görs med olika metoderna inom ramverket: KDR-metoden, KDR med PCR-metoden, KDR med PLS-metoden och TSR-metoden (*Trimmed scores regression – TSR*).

Bland de olika metoderna i ramverket presterade TSR-metoden bra för att ersätta databortfall i de flesta datastrukturer (Folch-Fortuny *et al.*, 2015). Denna metod för imputering av data är den bästa kompromissen mellan prediktionskvalitet, robusthet mot datastruktur och beräkningstid. Av de andra regressionsbaserade metoderna som presenterades erbjuder KDR-metoderna med PCR och PLS också bra lösningar, men de är mer tidskrävande eftersom de innebär att ytterligare PCR- och PLS-modeller måste anpassas.

6.6.3 Filtrering och utjämning av data

6.6.3.1 Syfte

Syftet med att jämna ut data från en signal (*smoothing*) är ta bort högfrekvent brus och istället framhäva de viktigaste trenderna i den uppmätta variabeln. Ett digitalt filter tillämpar matematiska operationer på en givarsignal för att reducera eller förstärka vissa komponenter i signalen (t.ex. processvariationer, brus).

6.6.3.2 Exempel på filtreringsmetoder

Det *glidande medelvärdet* är det enklaste och mest använda filtret. Det är lätt att förstå, tillämpa och ställa in för att få tillfredsställande resultat. Det beräknas genom att ta medelvärdet av en viss delmängd av värdena i en tidsserie för att få fram deras medelvärde. Genom filtreringen skapas en ny tidsserie med medelvärden som har lägre variabilitet (och därmed lägre brus) än den ursprungliga tidsserien. Det antal värden, N, som används för att bestämma varje glidande medelvärde (dvs. den fönsterlängd som används av filtret) måste väljas. När den underliggande processen är relativt konstant är ett stort N att föredra (vilket ger en mjukare filtrerad signal). En liten fönsterlängd gör dock att det glidande medelvärdet kan följa snabba förändringar närmare (dvs. det reagerar snabbare på förändringarna). Detta innebär i sin tur att mer brus kommer att finnas kvar i signalen.

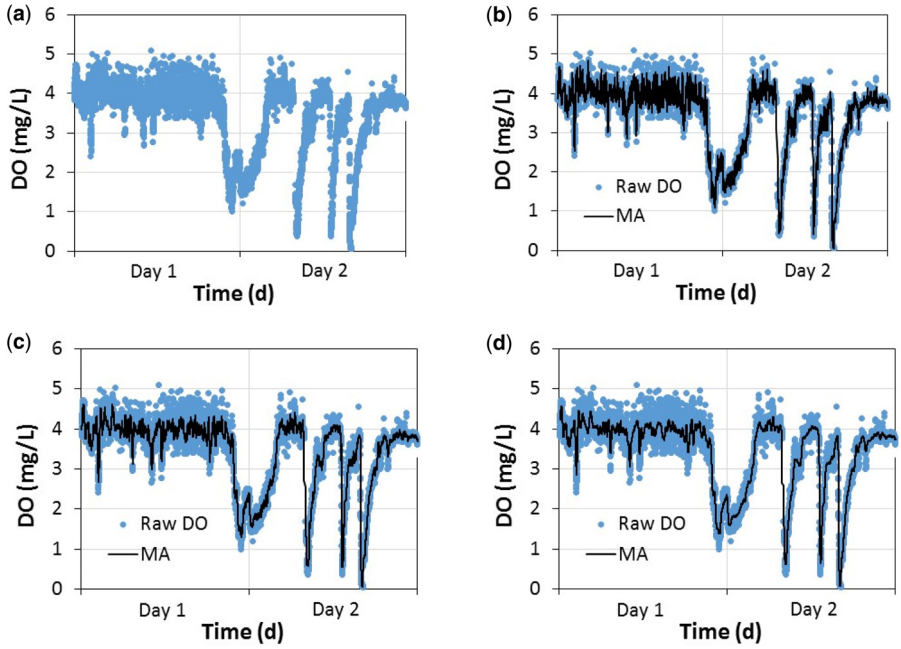
Figur 6.17 visar ett exempel på ett glidande medelvärdesfilter med olika fönsterlängder för en syrehaltsmätning.

Det glidande medelvärdet kan beräknas utifrån tidigare mätvärden (t.ex. för en fönsterlängd $n = 3$ kommer tre värden att medelvärdesbildas x_{t-2}, x_{t-1}, x_t), vilket möjliggör en realtidsberäkning av det glidande medelvärdet. För analys av historiska dataset är det bättre med ett centrerat glidande medelvärde (för tidsögonblicket t skulle de tre medelvärdena vara x_{t-1}, x_t, och x_{t+1}), vilket ger en symmetrisk medelvärdesbildning när fönsterlängden är ett udda tal. Observera att beräkningen av det glidande medelvärdet liknar tillämpningen av en rektangulär puls som filter: till exempel, för ett glidande medelvärde med $n = 3$, skulle det tillämpade filtret vara …0, 0, 1/3, 1/3, 1/3, 1/3, 0, 0, ….

Eftersom beräkningen av medelvärdet påverkas starkt av extremvärden kommer ett extremvärde att störa alla medelvärden som tas fram för det valda fönstrets längd. I de fall där extremvärden är vanliga går det istället att beräkna den glidande medianen (som alternativ till medelvärdet) som en robust skattning av fönstervärdenas genomsnitt.

För att få en ännu jämnare filtrerad signal kan de glidande medelvärdena eller de glidande medianerna itereras två eller flera gånger.

Lokal regression (local regression smoothing) använder på samma sätt ett glidande fönster, men där varje utjämnat värde bestäms utifrån en regressionsmodell som anpassas till värdena inom fönstret. För varje position i fönstret, runt fönstrets mittpunkt, beräknas en linjär funktion eller en polynomfunktion med en minsta-kvadratanpassning. Inverkan av extremvärden kan minskas genom att använda s.k. robust regression, vilket är en motsvarighet till att använda median istället för medelvärde för det glidande medelvärdesfiltret. När värdena för den signal som ska utjämnas är tydligt icke-linjära är ett annat alternativ *kernel regression* som skapar en jämn funktion och approximerar värdena lokalt inom fönstret. Det är en icke-parametrisk metod som utgår från en viktfunktion (*kernel*, vanligtvis en gaussformad viktfunktion) för att vikta varje värde i fönstret och skapa ett viktat medelvärde för varje position i fönstret. Antalet värden som ska ingår i viktningen



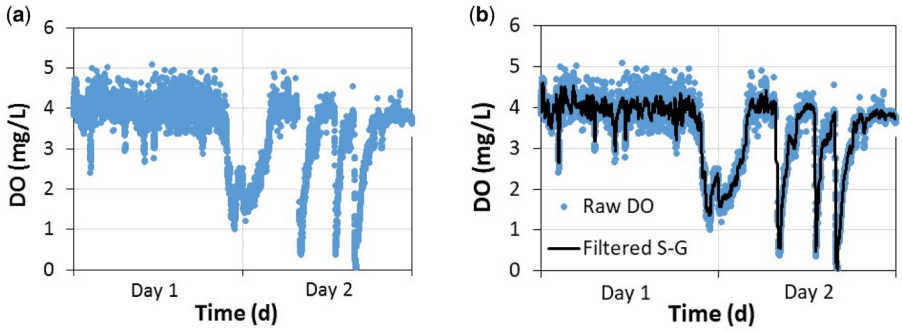
Figur 6.17 (a) Syrehalten under två dygn med 20 sekunders mätintervall tillsammans med ett överlagrat glidande medelvärde av fönsterlängd: (b) $n = 3$ (1 min glidande medelvärde), (c) $n = 15$ (5 min glidande medelvärde), och (d) $n = 45$ (15 min glidande medelvärde).

är en avvägning mellan utjämningseffekten och den fördröjning som införs i de filtrerade värden som produceras.

Exponentialfiltret utjämnar värdena i en signal genom att ta hänsyn till alla tidigare värden i tidsserien och vikta dem med minskande vikter över tid (Montgomery *et al.*, 2015). Den enklaste versionen är första ordningens exponentialfilter: $s_t = \lambda x_t + (1-\lambda)s_{t-1}$, där s_t är den utjämnade filtrerade data vid tidpunkten t, x_t det aktuella värdet för tidsserien och λ är vikten för det aktuella värdet (observationen) och bör ligga i intervallet $0 < \lambda < 1$ (även om vanliga värden som fungerar bra i praktiken ligger mellan 0,1 och 0,4). Observera att ju högre λ är desto större vikt läggs vid det aktuella värdet och därmed kommer den utjämnade utdata att följa tidsseriens beteende mer noggrant och reagera snabbare på dess förändringar. För att förbättra filtrets beteende när det finns en trend i tidsseriens värden kan exponentialfiltret appliceras rekursivt två gånger, vilket kallas andra ordningens exponentialfilter (eller dubbelexponentialfilter). På liknande sätt, när det finns ett cykliskt mönster i tidsseriens värden, kan exponentialfiltret rekursivt appliceras tre gånger, vilket kallas exponentiellt filter av tredje ordningen (eller trippelexponentiellt filter).

Savitzky-Golay-filtret är ett populärt filter som, i likhet med det glidande medelfiltret, beräknar det filtrerade värdet baserat på en ny uppsättning datapunkter. Det fungerar genom att ett polynom (av en given grad) anpassas till en ny uppsättning datapunkter med hjälp en linjär minstakvadratanpassning. En analytisk lösning finns om de sekventiella värdena i tidsserien är jämnt fördelade i tiden. Detta möjliggör snabb exekvering av filtret i realtidsapplikationer. Polynomgraden och fönsterlängden bestämmer brusreduceringsnivån och omfattningen av den filtrering som görs.

Ett polynom med lågt ordningstal kan snyggt jämna ut breda toppar som finns i originalsignalen men kan inte beskriva formen (höjd och bredd) på smala toppar i den utjämnade utsignalen (Sousa *et al.*, 2007). I Figur 6.18 visas samma exempel som i Figur 6.17 med en syrehaltsmätning i en biologisk

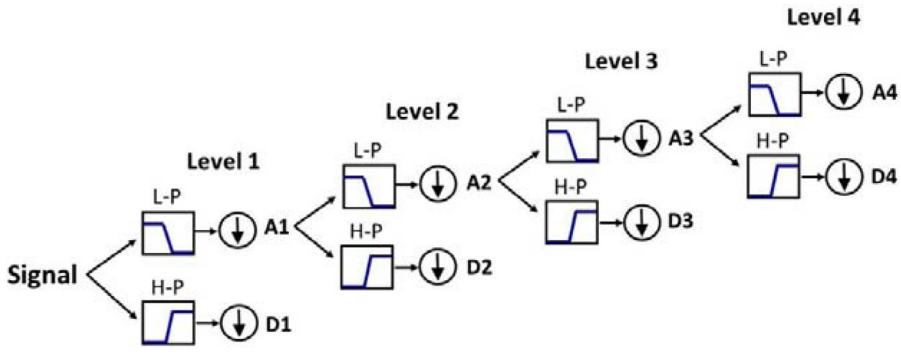


Figur 6.18 (a) Syrehalten under två dygn med 20 sekunders mätintervall. (b) Originalsignal tillsammans med överlagrad filtrerad signal från Savitzky-Golay-filter (med en fönsterlängd på 45 sampel och anpassning en andra ordningens polynomfunktion).

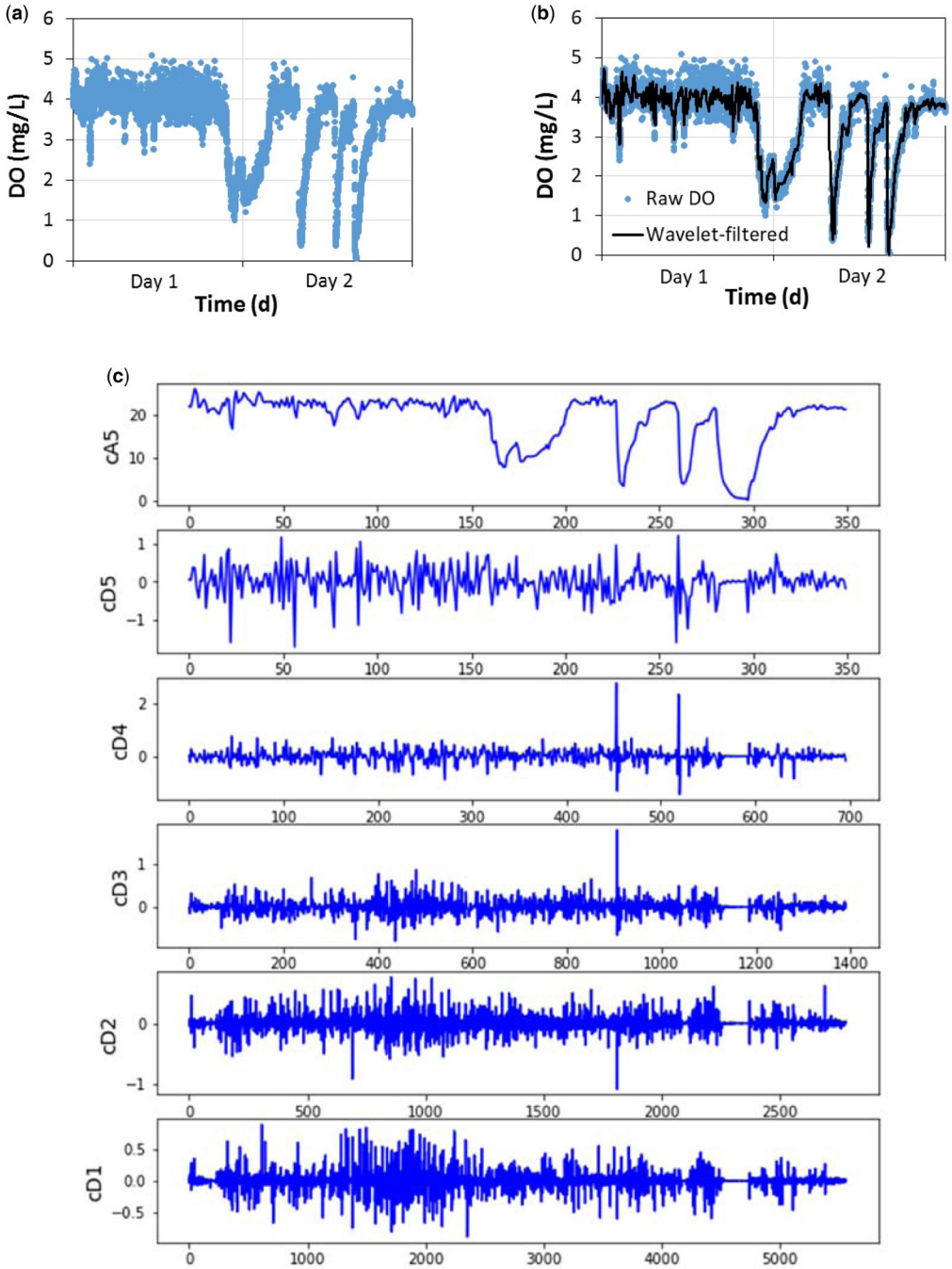
reaktor, men istället med ett Savitzky-Golay-filter med en fönsterlängd på 45 datapunkter med en andra ordningens polynomfunktion.

Wavelets (små oscillerande vågor) kan användas för utjämning, signalfiltrering, komprimering och extrahering av funktioner. En wavelet används för att dela upp originalsignalen i olika skalor med minskande detaljnivå (eller upplösning) med hjälp av en moderwavelet (*mother wavelet*) som successivt skalas och förändras i tiden. Det finns olika familjer (typer eller former) av wavelets och många av dem har visat sig vara användbara (t.ex. Haar, Daubechies, Symlets, Meyer etc.). Vanligtvis väljs wavelets empiriskt så att de har en form som är karakteristisk för det beteende som originalsignalen uppvisar. På så sätt kommer mindre av det förväntade beteendet i originalsignalen att jämnas ut.

I praktiken implementeras DWT (*Discrete Wavelet Transform*) som en sekvens av komplementära högpass- och lågpassfilter, som delar upp signalen i detaljer och approximationsdelar. Varje nivå innefattar filtrering av insignalen (eller dess approximation från föregående nivå) genom ett par komplementära filter följt av en nedsampling där vartannat värde av det filtrerade resultatet kasseras. Denna procedur (se Figur 6.19) upprepas upp till önskad precisionsnivå. Dekompositionsnivån väljs empiriskt utifrån den bearbetade signalens karaktär eller via automatiska kriterier som entropi (Mallat, 1989).



Figur 6.19 Schema för signaluppdelning upp till nivå 4. A: approximation, D: detaljer, H-P: högpassfilter och L-P: lågpassfilter.



Figur 6.20 (a) Syrehalt under två dygn med 20 sekunders mätintervall. (b) Originalsignal och waveletfiltrerad signal (efter fem nivåer av sönderdelning med wavelet Daubechies 2 – db2), (c) approximationskoefficienter (cA) och detaljkoefficienter (cD) för varje nivå av uppdelad signal.

DWT är en s.k. ortogonal basdekomposition och därmed en inverterbar operation (dvs. originalsignalen kan rekonstrueras genom att vända på beräkningen). Detaljerna representerar den högfrekventa delen av insignalen, medan approximationen (utgången från lågpassfiltret) är en utjämnad version av insignalen med föregående upplösning. Vid en given nivå är detaljerna huvudsakligen fluktuationer som beror på oönskat brus, som kan tas bort (nollställas via tröskelvärden) utan att de viktiga mönstren i originalsignalen förvrängs. Även om DWT bygger på avancerad matematik är strategin för att ta bort högfrekvent brus enkel: (1) använd DWT för att dela upp originalsignalen; (2) sätta detaljkoefficienter som är högre än det givna tröskelvärdet till noll, och (3) använda invers DWT för att beräkna den utjämnade signalen.

Figur 6.20 visar hur det tidigare filtreringsexemplet blir med en waveletfiltrerad signal med fem nivåers uppdelning med Daubechies-2-wavelets.

6.6.4 Karaktärsdrag i data

6.6.4.1 Syfte

Syftet med s.k. *feature engineering* är att omvandla primärdata till nya data som förbättrar efterföljande dataanalys, främst genom att undertrycka irrelevant information och beräkna nya variabler som anses relevanta för efterföljande analys. Att använda relevanta (dvs. informativa, meningsfulla, diskriminerande) beräkningsfunktioner är avgörande för att lyckas med datadriven analys (mönsterigenkänning, klassificering eller regression). Vid feature engineering används teknisk bakgrundskunskap (antingen om processen eller om data) för att extrahera eller skapa karaktärsdrag (*features*) såsom variabler, attribut eller deskriptorer som är relevanta för tillämpningen. Ett enkelt exempel är normalisering av energianvändningen, som kan vara mer informativ i termer av kWh/inkommande mängd kväve, jämfört med enbart själva energianvändningen.

Data från beräkningsfunktionerna kan vara betydligt mer relevanta än de ursprungliga variablerna, vilket Aguado *et al.* (2022) visade när de utvecklade en prediktiv softsensor. Detta visar på vikten av feature engineering vid praktiska AI-tillämpningar.

6.6.4.2 Exempel på metoder

Att kombinera flera olika features till ett nytt feature kan vara informativt vid datadriven modellering. Till exempel så är syrehalten i den luftade zonen på ett reningsverk viktig, men dess förhållande till det önskade börvärdet kan vara ännu mer informativ om luftningssystemets prestanda ska analyseras. På samma sätt kan elförbrukningen per dygn i termer av kWh ge insikt i reningsverkets energieffektivitet. Men för att kunna jämföra energieffektiviteten med andra verk så kan elanvändningen per volym avloppsvatten eller per antal anslutna personer vara ett bättre nyckeltal och feature.

Beräkna skillnaden mellan två givarsignaler. En differens mellan två signaler behöver förstås vara rimlig och informativa utifrån ett processperspektiv. Ruano *m.fl.* (2012) visade t.ex. att pH-differensen mellan början och slutet på den aeroba volymen gav information om nitrifikationsprocessen, och även denna pH-differens användes för reglering. Thürlimann *et al.* (2018) använde pH-differensen mellan den första och sista luftade tanken för att upptäcka toppbelastningar av ammoniumkväve, och föreslog en regelbaserad styrstrategi i kombination med kvalitativ trendanalys (som användes för att upptäcka de relativa förändringarna i den beräknade pH-differensen) som framgångsrikt testades på flera fullskaliga kommunala anläggningar.

Lägg till tidsfördröjningar, s.k. laggade värden, för att inkludera processdynamik i den datadrivna modellen. Denna beräkningsteknik skapar nya features från befintliga data och förflyttar också de beräknad features i tid. Tidsfördröjningar används ofta i tidsserieanalys och prognostisering. När ett visst feature x_t fördröjs skapas nya features helt enkelt genom att det ursprungliga värdet förflyttas önskat antal tidsenheter (n): x_{t-1}, x_{t-2}, x_{t-3}, x_{t-n}, med t.ex. ett rörligt fönster av storleken n. På så sätt kan en datadriven modell fånga hur den aktuella situationen/det aktuella beteendet (vid tidpunkten *t)* påverkas av de tidigare värdena för egenskaperna. Detta ger olika möjligheter för

att representera processens dynamik beroende på vilka features som fördröjs och hur lång denna fördröjning är. Inledningsvis kan antalet fördröjningar väljas utifrån processkunskap alternativt genom en känslighetsanalys där olika alternativ prövas och utvärderas.

Tolka kvalitativa variabler kvantitativt genom att omvandla kvalitativa data till numeriska värden. Kvalitativa mätvärden såsom on/off-läget för en pump eller en fläkt kan inte direkt användas i numeriska dataanalysmetoder. För att inkludera denna kvalitativa information i modelleringen kan man använda sig av s.k. one-hot-kodning när det inte finns någon naturlig ordning mellan kategorierna i den variabel som ska kodas. Med denna kodningsteknik skapas en ny binär variabel (kallas dummyvariabel inom statistiken) för varje möjlig kategori av den variabel som ska kodas. Om det t.ex. finns två kategorier (på/av) för en pumps tillstånd skapas två binära variabler som kodas som [1, 0] och [0, 1]. För varje tidsstämpel kommer endast en av de binära variablerna att vara 1 och den andra 0. När det finns en naturlig ordning mellan kategorierna i den kvalitativa variabel som ska kodas kan ett heltalsvärde tilldelas varje kategori. Denna kodningsteknik kallas ordinalkodning.

Diskretisering av kontinuerliga egenskaper kan ibland användas för att hjälpa den datadrivna modellen att upptäcka mönster i data som inte är lika uppenbara som när man använder kontinuerliga värden. Genom diskretiseringen omvandlas en kontinuerlig variabel till en serie diskreta eller nominella intervall. Den här tekniken kan t.ex. användas för att diskretisera koncentrationen av syrehalten i en luftningsbassäng i tre kategorier (låg, medel, hög), som senare kan kodas för att kunna integreras i den datadrivna modellen med hjälp av one-hot-kodning som [1, 0, 0], [0, 1, 0] respektive [0, 0, 1]. Observera att diskretisering vanligtvis även innebär en betydande informationsförlust.

Extrahera information från tidsstämpeln som kan vara användbar för det aktuella modelleringsändamålet, t.ex. vilken timme på dygnet det är, veckodag eller helg, helgdag eller inte, årstid osv. Dessa datum- och tidsbaserade features är intressanta för att förutsäga variabler som uppvisar ett typiskt mönster (t.ex. temperatur och inflöde).

Beräkna derivatan av primärdata, som kan ge användbar information i olika situationer. Den första derivatan av en givarsignal är signalens förändringshastighet med tiden, och den tolkas vanligtvis som lutningen (eller tangenten) till signalen vid varje tidsögonblick. Andraderivatan är derivatan av den första derivatan och ett mått på signalens krökning (dvs. förändringshastigheten för signalens lutning). Högre derivatordningar kan beräknas på liknande sätt. Derivaten kan användas för att upptäcka toppar i primärdata. Derivat har använts inom analytisk spektroskopi i flera syften (Mark & Workman, 2007): (1) för spektral diskriminering (dvs. för att framhäva små strukturella skillnader mellan nästan identiska spektra); (2) för korrigering av irrelevant bakgrundsabsorption; och (3) för spektral upplösningsförbättring (dvs. för att öka den skenbara upplösningen hos överlappande spektralband för att underlätta bestämningen av antalet band och deras våglängder). Försiktighet bör dock iakttas vid numerisk dervivering eftersom den vanligtvis försämrar signal/brusförhållandet. Eftersom derivering av en signal ökar den relativa mängden brus (signal/brusförhållandet för en derivata är nästan alltid sämre än för den ursprungliga signalen), är derivering i kombination med lämplig brusfiltrering avgörande för att optimera signal/brusförhållandet. Derivatan av primärdata från SBR (Sequencing Batch Reactors) kan beräknas för att upptäcka viktiga brytpunkter (ammoniumdipp, nitratknä, nitrattopp, syrehaltsbrytpunkt etc.) (Dutta & Sarkar, 2015) för signalprofilerna med lågkostnadssensorer (oxidationsreduktionspotential (ORP), pH, syrehalt) eftersom de är nära relaterade till de biologiska reaktionerna (nitrifikation, denitrifikation, avlägsnande av organiskt material etc.). Detta möjliggör en realtidsövervakning av längden på SBR-faserna, vilket ger bättre hantering av processavvikelser och variationer i inflödessammansättningen (Marsili-Libelli, 2006) än när traditionell SBR-styrning används som baseras på ett fast tidsschema. I kontinuerliga reningsprocesser har derivator av givarsignaler också visat sig vara användbara, särskilt när givare är benägna att driva och signalen därmed innehåller information i sina derivator (Ruano *et al.*, 2012; Thürlimann *et al.*, 2018; Wang *et al.*, 2021).

6.6.5 Datatransformation

6.6.5.1 Syfte

Transformationer är icke-linjära omvandlingar av signaldata. Som exempel kan log-transformation och kvadratrots-transformation nämnas, vilka gör det möjligt att korrigera avvikelser från normalfördelade rådata. Genom att tillämpa en matematisk funktion på värdena för en variabel är det möjligt att omforma dess fördelning. Det främsta motivet för att transformera data är att omvandla en skev fördelning till en normalfördelning, eller åtminstone mer symmetrisk fördelning (Eriksson *et al.*, 2013). Att data är normalfördelade är ett mycket vanligt antagande i många statistiska analyser, särskilt inom parametrisk statistik. Detta antagande gör det möjligt att bestämma konfidensintervall och genomföra hypotestest. Avvikelser från normalfördelningen minskar tillförlitligheten i dessa tester och gör även dataanalysen mindre effektiv. Dessa normalitetsavvikelser innebär dock inte att resultaten eller slutsatserna av de statistiska testerna blir helt ogiltiga, såvida det inte är fråga om en allvarlig avvikelse. Med en transformation av data på en given variabel går det att minska avvikelsen från normalfördelningsantagandet.

Ibland kan förekomsten av extremvärden i datasetet påverka den datadrivna modellen på ett negativt sätt. Dessa extremvärden kan snedvrida fördelningen, vilket gör att den avviker kraftigt från normalfördelningsantagandet. Mätfel, felaktiga data och störningar kan alla orsaka extremvärden. Notera att transformationer (t.ex. logaritmisk transformation eller kvadratrotstransformation) kan maskera outliers. En processexpert kan identifiera vilka extremvärden som kan betraktas som outliers och först därefter tas bort från datasetet. De extremvärden som inte är outliers och som gör att variabelns fördelning inte är normal kan därefter transformeras så att fördelningen omformas till en mer symmetrisk fördelning.

6.6.5.2 Exempel på transformationsmetoder

Den *logaritmiska transformationen* används ofta eftersom variabler med en log-normalfördelning är relativt vanliga (t.ex. koncentrationer, flödeshastigheter) som uppvisar högertyngda fördelningar (svansarna till höger – positiv skevhet). Denna typ av transformation kan tolkas som att en förändring i en logaritmisk skala innebär proportionella relativa förändringar på den ursprungliga skalan. Om variabelns ursprungliga värde är x_i skulle det transformerade värdet (x_i*) vara $x_i^* = \log(x_i)$. Eftersom log(0) är odefinierat – liksom logaritmen för alla negativa tal – bör man vid logtransformation lägga till en konstant till alla värden så att de blir positiva före transformationen, dvs. $x_i^* = \log(\text{konstant} + x_i)$.

Figur 6.21 visar histogrammet för 2 års COD-mätningar parallellt med dess logaritmiska transformation. Som figuren visar så har transformationen omvandlat den högertyngda skeva fördelningen till en mer symmetrisk fördelning.

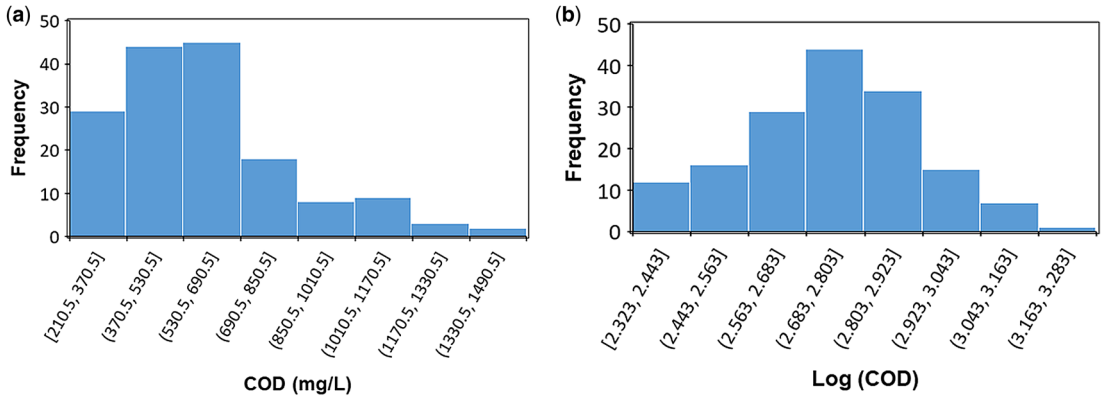
Kvadratroten och kubikroten används också för att transformera variabler som är skeva och högertyngda. Dessa transformationer kallas även *power transformations* eftersom de kan uttryckas som x_i^p, där p är den inversa exponenten (1/2 för kvadratrot, 1/3 för kubikrot). När de används komprimeras höga värden av variabeln och låga värden av variabeln blir mer utspridda, men komprimeringen och spridningen är jämnare än när den logaritmiska transformationen används.

Box-Cox är en användbar familj av transformationer som omfattar alla de tidigare nämnda transformationerna. Denna familj har en parameter (λ) som bestämmer vilken typ av transformation som ska tillämpas på data. När värdet på λ är 0 ($\lambda = 0$) är den tillämpade transformationen den naturliga logaritm, men om $\lambda \neq 0$ används en power transformation. I specialfallet $\lambda = 1$ sker ingen förändring av förändring form och de transformerade uppgifterna förskjuts bara nedåt:

$$x_i^* = \log(x_i) \text{ when } \lambda = 0 \quad \text{and} \quad x_i^* = (x_i\lambda - 1)/\lambda \text{ when } \lambda \neq 0$$

Det bör nämnas att Bickel och Doksum (1981) föreslog en liten ändring av Box-Cox-transformationen för att tillåta negativa värden på variabeln (x):$_i$

$$x_i^* = \log(x_i) \text{ when } \lambda = 0 \quad \text{and} \quad x_i^* = \text{sign}(x_i)(/x_i/\lambda - 1)/\lambda \text{ when } \lambda \neq 0$$



Figur 6.21 (a) Histogram över inkommande COD-halt (158 värden, motsvarande 2 års laboratoriemätningar). (b) Log-transformerade COD-värden för att göra fördelningen mer symmetrisk. Notera att ett histogram är ett visuellt och enkelt sätt att bedöma dess överensstämmelse med en normalfördelning.

För det mindre vanliga fallet att de insamlade variablerna uppvisar en vänstervriden fördelning (dvs. svansen sträcker sig tydligt åt vänster – negativ skevhet) kan slutligen kvadratrot, kubikrot och exponentiell omvandling användas för att omvandla dem till mer symmetriska fördelningar.

6.6.6 Skalning av data

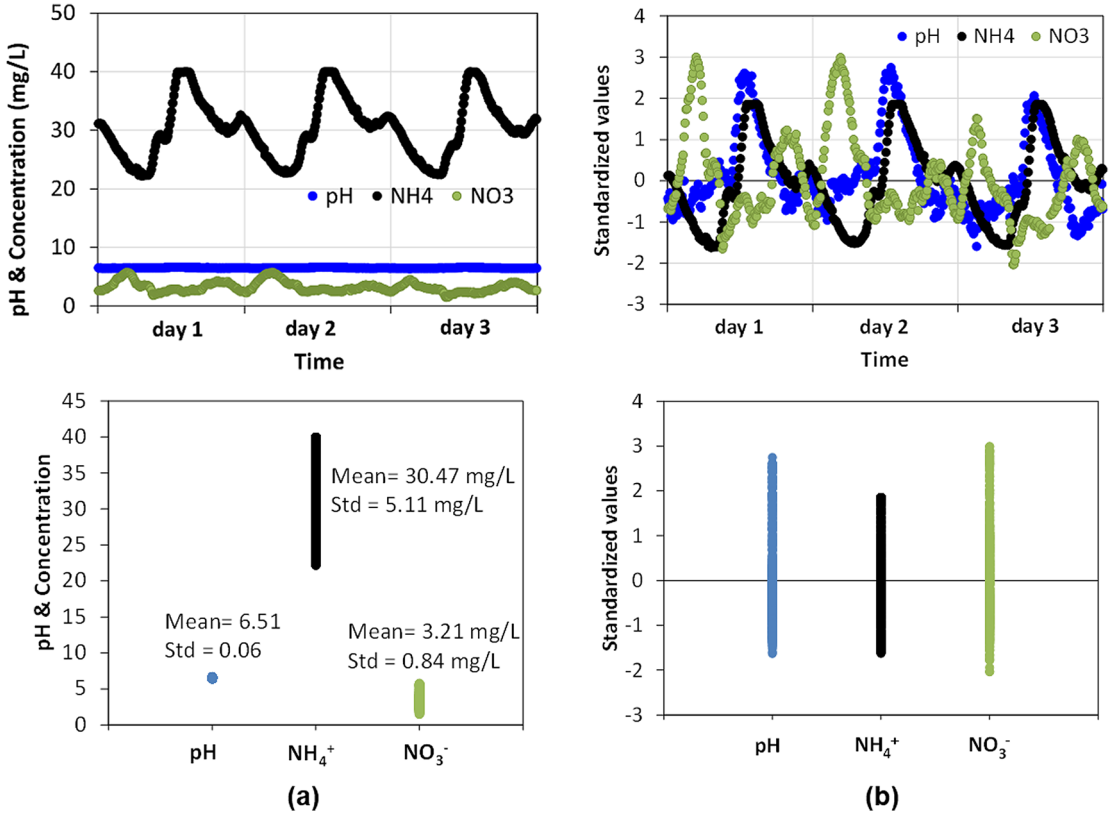
6.6.6.1 Syfte

Syftet med skalning är att ändra intervallet för en viss variabel, det vill säga att ändra dess numeriska värden men behålla deras relativa positioner/proportioner. Fördelningens form påverkas alltså inte.

Vanligtvis används datadrivna modeller för att avslöja de underliggande mönstren i data för att ge ett bättre beslutsunderlag och öka processförståelsen med övervaknings-, reglering- och prediktionsapplikationer. Därför analyseras ofta många variabler från processen samtidigt för att man ska kunna ta hänsyn till deras korrelationer och interaktioner. Dessa variabler kan ha helt olika numeriska intervall (t.ex. pH, syrehalt och flödeshastighet). Dessutom påverkas det numeriska området för en signal av den valda måttenheten. Skalning kan användas för att undvika att vissa variabler dominerar över andra på grund av deras högre numeriska intervall eller för att säkerställa att analysresultaten är okänsliga för den ursprungliga måttenheten. Många datadrivna metoder är känsliga för skalningen och fungerar inte bra när de ingående variablerna har mycket olika skalor. Skalningen av variablerna gör det möjligt att korrigera denna situation. Observera att skalningen kan reglera den inverkan som en viss variabel har på en datadriven modell. En variabel som man vet är relevant för ett visst fall kan tilldelas en högre vikt för att öka dess bidrag till den fortsatta modelleringen. Förkunskaper om olika variablers relevans kan alltså inkluderas via skalningen. Det är viktigt att notera att många dataanalysverktyg är känsliga för datainmatningens intervall, vilket innebär att varje förändring av skalan leder till en förändring av de beräknade värdena eller parametrarna.

6.6.6.2 Exempel på skalningsmetoder

Skalning till variansen ett (1) är den vanligaste och rekommenderade skalningen när man inte tar hänsyn till förkunskaper om data beaktas och då variablerna har olika mätenheter. Varje variabelvärde divideras med den totala standardavvikelsen, så att varje variabel får variansen 1. Detta kallas ibland standardisering. På så sätt har alla variabler samma chans att bidra till den fortsatta dataanalysen. Observera att de resulterande värdena inte är bundna till ett specifikt intervall.



Figur 6.22 pH-, ammoniumkväve- och nitratkoncentration i en anoxbassäng under tre dygn: (a) registrerade värden (dvs. medelvärdet för var 10:e minut) och (b) efter medelvärdescentrering och skalning till enhetsvarians.

Tillsammans med standardisering används ofta en annan skalningsteknik som kallas medelvärdescentrering (*mean centering*) för att öka tolkningsbarheten hos den datadrivna modellen. Medelvärdescentrering omvandlar alla värden till variationer kring noll i stället för runt medelvärdet. Detta gör att analysen fokuseras på den fluktuerande delen av data och tar bort förskjutningen (om det finns någon). När medelvärdescentrering och standardisering används samtidigt så subtraheras medelvärdet för varje värde på variabeln och divideras sedan med standardavvikelsen. Detta tillvägagångssätt resulterar i skalade data med noll medelvärde och variansen 1 (Figur 6.22).

Min-max-skalning (även känd som normalisering) är en annan vanlig teknik för att se till att alla variabler har samma skala. Med denna metod så subtraheras det lägsta värdet för varje variabelvärde och divideras sedan med intervallet (dvs. det högsta värdet minus det lägsta värdet). Därför ligger de skalade värdena i intervallet 0 till 1. Denna skalningsmetod används ofta med variabler som ska användas i neurala nätverk (eftersom de ofta förväntar sig ingångsvärden i intervallet 0–1). Observera att denna skalningsprocedur är känslig för extremvärden.

Blockskalning används för att undvika att en grupp naturligt relaterade variabler dominerar över andra variabler i den senare modelleringen bara på grund av det stora antalet variabler i gruppen (Eriksson *et al.*, 2013). Om vi t.ex. i en datadriven modell beaktar inflödet, syrehalten i luftningsbassängen och de spektra som produceras av en spektrofotometrisk sensor (t.ex. 221 våglängder), kan vi använda blockskalning för spektrumen för att låta de andra två variablerna (flöde och syrehalt, som kan skalas

till enhetsvarians) ha möjlighet att bidra till modellen. I blockskaleringsmetoden skalas alla variabler i varje block så att summan av deras varianser blir 1. För att uppnå detta multipliceras varje värde med inversen av kvadratroten ur antalet variabler i blocket.

Det finns även andra skalningsmetoder som inte används lika ofta i reningsverkssammanhang. Till exempel finns *Pareto-skalningen*, som använder inversen av kvadratroten ur standardavvikelsen som skalningsfaktor, och vissa metoder som använder ett storleksmått (medelvärdet eller kvadratroten ur medelvärdet) i stället för ett spridningsmått som skalningsfaktor, till exempel *nivåskalningen* eller *Poisson-skalning*.

6.6.7 Datastrukturering: metoder för att hantera ostrukturerade data

6.6.7.1 Syfte

De flesta dataanalysverktyg för *data mining*, statistisk processövervakning, maskininlärning och modellidentifiering kräver att data presenteras på ett strukturerat sätt (se Kapitel 2 för definition). I korthet så innebär det att data behöver kunna lagras i en strukturerad matris med en tydlig beskrivning av respektive kolumn/dimension. Vissa metoder kräver dessutom att data ska vara tvådimensionella. Typiskt så representerar varje rad en mättidpunkt medan varje kolumn representerar en specifik variabel på en specifik plats.

Tyvärr presenteras data inte alltid i den här typen av strukturerade matrisliknande format. Då klassificeras data som ostrukturerade data (se Kapitel 2 för definition och exempel). En naturlig steg är då att strukturera data så att de kan användas för vidare dataanalys. Följande avsnitt beskriver förbehandlingsmetoder för att strukturera data genom några exempel.

6.6.7.2 En kommentar om icke-reversibel datastrukturering

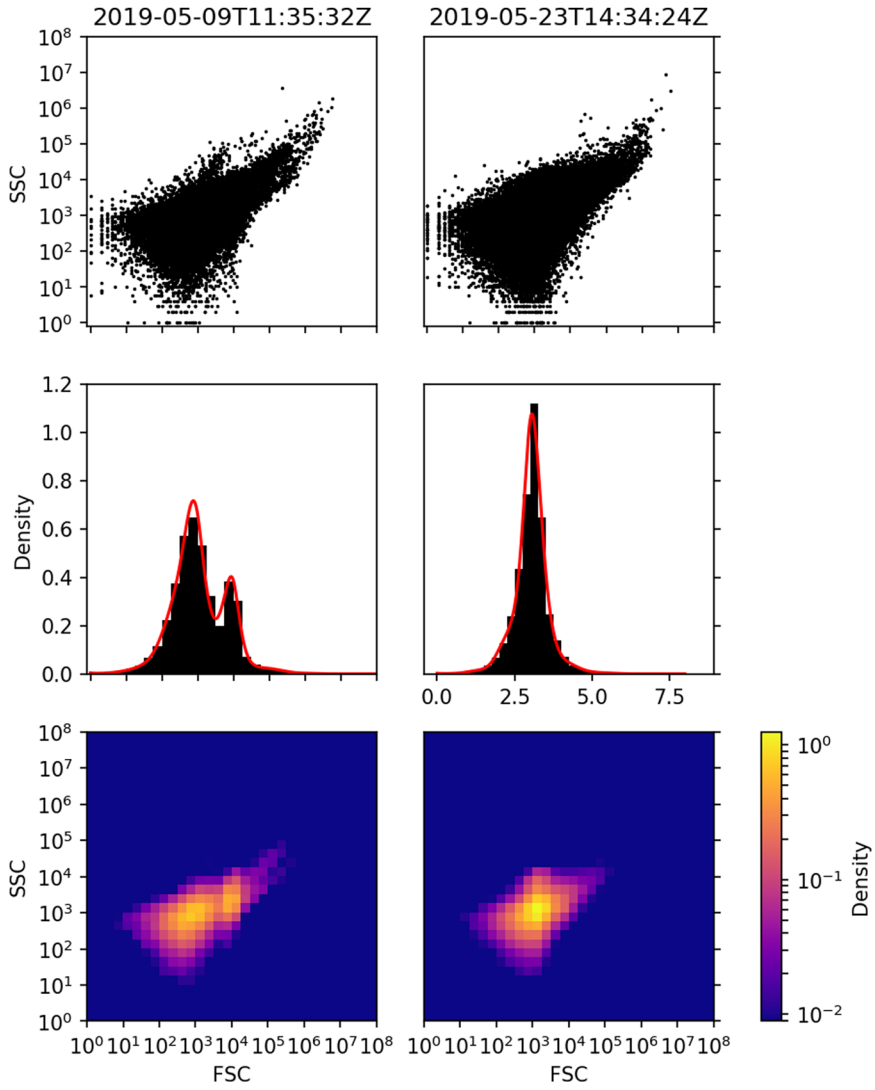
I allmänhet kan datastrukturering inte reverseras, d.v.s. att de ursprungliga ostrukturerade data inte kan rekonstrueras exakt utifrån de strukturerade uppgifterna. Datastrukturering leder därför nästan alltid till viss informationsförlust. Av den anledningen måste man därför se till att (a) lagra de ursprungliga ostrukturerade data om de anses relevanta för framtida ändamål och (b) finjustera datastruktureringsprocessen på ett sätt som gör att fördelarna (t.ex. effektiv användning och lagring av data) vägs mot informationsförlusten. Om det är möjligt och relevant bör irreversibla förbehandlingssteg, t.ex. datastrukturering, utföras före reversibla förbehandlingssteg (t.ex. centrering och skalning). Datastrukturering är ofta ett av stegen i ett dataflöde.

6.6.7.3 Metoder för att strukturera mätdata för partiklar

Antalet partiklar i vattenprover varierar vanligtvis mellan ett vattenprov och ett annat. Som Exempel 2.4 i Kapitel 2 beskrev så leder detta till ostrukturerade data när egenskaperna hos enskilda partiklar mäts. För att möjliggöra analys kan dessa ostrukturerade data omvandlas och struktureras genom (a) *binning*, (b) uppskattning av sannoliktäthetsfunktioner och (c) avståndsberäkning. Dessa datastruktureringsmetoder illustreras med data från Exempel 2.4 i Kapitel 2 i Figur 6.23. Som Kapitel 2 beskriver så innehåller dessa två dataposter mätningar av fyra variabler (fyra ljusintensiteter), vilka benämns som *forward scatter count – FSC*, *side scatter count – SSC*, fluorescein isothiocyanat och peridinin klorofyllprotein) i två vattenprover med 56 358 respektive 164 810 partiklar. De översta figurerna i Figur 6.23 visar mätningarna för två variabler (FSC och SSC). Observera att dessa mätningar ges i logaritmisk skala, vilket innebär att en transformation (Avsnitt 6.6.5) har applicerats innan datastruktureringen.

Binning innebär att mätområdet delas upp i intervall eller fack. Placeringen av dessa intervall densamma för varje datapost. Binning är jämförbart med det sätt ett histogram tas fram på.

Histogrammet kan normaliseras genom att varje summa av partiklar inom respektive intervall divideras med det totala antalet partiklar i dataposten och det aktuella intervallets bredd. Exempel för FSC-variabeln visas i mittenfigurerna i Figur 6.23 och i Tabell 6.1. Dessa histogram är beräknade med 32 lika stora intervall. När de används för alla 137 dataposter genererar detta en



Figur 6.23 Strukturering av partikelmätdata – exempel med två dataposter från en flödescytometrisignal. Vänster: första dataposten; höger: andra dataposten. Tidsstämpeln visas överst. Överst: originalmätningar för två ljusintensiteter (FSC och SSC). Mitten: normaliserat histogram (svarta staplar) och med en empiriskt sannolikhetsfördelning (röd linje) för FSC-variabeln. Nederst: tvådimensionellt normaliserat histogram för FSC- och SSC-variablerna, som här visualiseras med en heat map.

137×32-dimensionell matris. Dessa strukturerade data kan sedan enkelt analyseras med hjälp av konventionella dataanalysmetoder.

Binning kan enkelt generaliseras till det multivariata fallet. Till exempel kan mätområdet för både FSC- och SSC-variablerna delas upp i 32 lika långa intervall. Detta leder till en matris med $32 \times 32 = 1\,024$ separata intervall. De normaliserade tvådimensionella histogrammen visas i de nedre figurerna i Figur 6.23. Genom denna binning omvandlas var och en av de 137 dataposterna till

Tabell 6.1 Strukturering av partikelmätdata.

Histogram			Normaliserat histogram			Empirisk (Log-Spline) täthetsfunktion Koefficienter		
Intervall	2019-05-09 11:35:32	2019-05-23 14:34:24	Intervall	2019-05-09 11:35:32	2019-05-23 14:34:24	Parameter	2019-05-09 11:35:32	2019-05-23 14:34:24
1	17	45	1	0,001	0,001	1	−3,14	4,174
2	71	105	2	0,006	0,003	2	−2,196	−2,64
3	52	102	3	0,005	0,003	3	−2,095	−2,705
4	106	216	4	0,009	0,006	4	−0,595	−0,177
5	245	381	5	0,021	0,01	5	0,362	−0,044
6	392	762	6	0,034	0,02	6	1,84	2,169
7	766	1 284	7	0,067	0,034	7	2,747	2,54
8	1 307	2 456	8	0,114	0,066	8	3,68	4,781
9	2 572	4 762	9	0,224	0,127	9	1,279	3,416
10	4 317	8 922	10	0,376	0,239	10	3,621	1,394
11	6 579	16 302	11	0,573	0,436	11	−0,201	0,781
12	7 418	27 761	12	0,646	0,742	12	0,129	−0,599
13	6 088	41 849	13	0,53	1,119	13	−1,035	−1,639
14	3 711	24 245	14	0,323	0,648	14	−2,937	−1,167
15	2 302	9 248	15	0,2	0,247	15	−0,322	−7,481
16	4 404	5 183	16	0,384	0,139	16	−28,572	−0,946
17	3 479	2 653	17	0,303	0,071	17	−19,163	−16,196
18	812	1 325	18	0,071	0,035	18	8,666	−16,099
19	424	795	19	0,037	0,021	19	−48,311	−2,078
20	318	505	20	0,028	0,014			
21	209	298	21	0,018	0,008			
22	156	175	22	0,014	0,005			
23	94	123	23	0,008	0,003			
24	38	85	24	0,003	0,002			
25	35	36	25	0,003	0,001			
26	19	4	26	0,002	0			
27	2	3	27	0	0			
28	1	1	28	0	0			
29	0	0	29	0	0			
30	0	0	30	0	0			
31	0	0	31	0	0			
32	0	0	32	0	0			

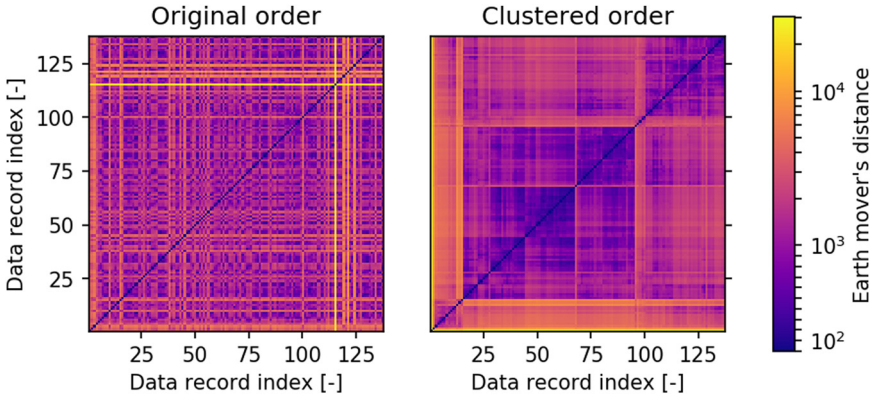
Exempel med två dataposter från en flödescytometrisignal. Tabellen visar de strukturerade data, dvs. vektorer med enhetlig längd vid beräkning av (a) ett histogram, (b) ett normaliserat histogram eller (c) koefficienterna för en log-spline-densitetsfunktion.

en sekvens av tal med enhetlig storlek. I exemplet leder det till 1 024 intervall för varje vattenprov, vilket motsvarar en $137 \times 1\,024$-dimensionell matris. De histogram som diskuteras ovan kan tolkas som en bitvis konstant täthetsfunktion. En mer förfinad funktion kan erhållas genom att använda

en bitvis linjär binning, se exempelvis Walther m.*fl.* (2009) (ekvation (6.1)) och Russo *m.*fl. (2022) (flödescytometrifallet).

Som ett alternativ till binning kan man också överväga att skatta en multivariat täthetsfunktion för varje vattenprov separat. Det är viktigt att täthetsfunktionen är parametriserad, vilket innebär att samma antal parametrar för täthetsmodellen används för att beskriva data för varje vattenprov. För detta ändamål kan man använda semiparametriska modeller, t.ex. log-spline-modeller, se Ramsay and Silverman (2005). Notera att det inte går att använda icke-parametriska täthetsfunktioner såsom den vanliga *kernel density estimation* eftersom antalet parametrar ändras som en funktion av antalet partiklar. De mellersta figurerna i Figur 6.23 visar två log-spline-modeller som båda baseras på *cubic-splines*. Den anpassade täthetsfunktionen beskriver den log-transformerade täthetsfunktionen med hjälp av 16 segment med tredjegradspolynom. Dessa polynom beskrivs i sin tur av 19 parametervärden (s.k. splinekoefficienter). Dessa spline-koefficienter listas i Tabell 6.1. I det här fallet leder denna typ av anpassning, till ett strukturerat dataset som innehåller 137 rader med 19 koefficienter vardera. Uppskattning av täthetsfunktioner kan även användas i multivariata fall (visas dock inte här). Praktiska problem med identifierbarhet kan dock uppstå i flerdimensionella fall med sex eller fler variabler.

En sista och tredje metod för att strukturera partikelmätningar består i att beräkna ett avstånd mellan varje datapostpar. I detta exempel ger det en symmetrisk 137 × 137-dimensionell avståndsmatris. Att definiera och beräkna ett avståndsmått för par av dataposter med varierande dimensioner är inte trivialt. Wasserstein-avståndet är dock ett populärt sätt att göra detta (Bigot, 2020; Frogner *et al.*, 2015). Wasserstein-avståndet är också känt som jordflyttningsavståndet ('earth (soil) moving distance – EMD' eftersom det beskriver den minsta ansträngning som krävs för att flytta en jordhög så att den matchar formen på en annan hög (Orlova *et al.*, 2016; Rubner *et al.*, 2000). För att beräkna EMD krävs i allmänhet att man löser ett storskaligt linjärt program för varje par dataposter. Detta är beräkningsmässigt svårt för generella fallet (Bigot, 2020; Frogner *et al.*, 2015) men det finns en enkel lösning för det univariata fallet. Den vänstra figuren i Figur 6.24 visar den parvisa EMD-värdet som baseras på FSC-variabeln för de 137 dataposterna i exempeldatasetet. Den 114:e dataposten (tidsstämpel: 2019-06-11 03:44:12) avviker från alla andra dataposter. Detta visar sig vara en datapost med ett extremt lågt antal partiklar (dvs. 70, se Figur 2.6 i Kapitel 2). Detta innebär att både antalet och fördelningen av FSC-mätningarna är väsentligt annorlunda. Den högra



Figur 6.24 Strukturering av partikelmätdata med EMD – exempel med 137 dataposter från en flödescytometrisignal. Vänster: parvis EMD för varje par av dataposter, i den ursprungliga ordningen (notera att den 114:e dataposten verkar vara mycket olik alla andra dataposter (i gult)). Höger: parvis EMD efter att dataposterna har ordnats enligt klusteravstånd.

figuren visar samma parvisa avstånd efter att dataposterna har ordnats enligt klustringsavstånd, efter hierarkisk klustring (med s.k. *Ward linkage*). Den blockdiagonala strukturen i den resulterande bilden tyder på att dataposterna kan grupperas i informativa grupper. Det är viktigt att notera att denna typ av klustring kräver att data struktureras genom avståndsberäkning, som illustreras här, eller genom den metod för binning eller täthetsuppskattning som beskrivs ovan.

6.6.7.4 Metoder för att strukturera textdata

Det finns flera metoder för att omvandla ostrukturerad text till ett strukturerat dataformat. En enkel metod är *bag-of-words*-metoden (Goldberg & Hirst, 2017). Med den metoden räknar man antalet gånger ett ord förekommer i ett dokument och rapporterar frekvensen för varje ord. De rapporterade frekvenserna resulterar i ett histogram för varje dokument som analyseras, med ordfrekvenser rapporterade för varje ord i något av de analyserade dokumenten. Detta leder till ett strukturerat dataformat eftersom antalet ord som beaktas är konstant. Denna *bag-of-words*-representation kan sedan användas som indata i konventionella verktyg för data mining. En mer förfinad metod kallas ordfrekvens × den inversa dokumentfrekvensen (*term frequency* × *inverse document frequency*) (Rajaraman & Ullman, 2011). Den metoden syftar till att ta hänsyn till det faktum att vissa ord, särskilt bestämd/obestämd artikel och pronomen, är vanliga i alla dokument och därför är mindre informativa om det specifika innehållet i ett visst dokument.

Utvecklingen går snabbt och alternativa maskininlärningsmodeller har utvecklats som även tar hänsyn till i vilken ordning orden förekommer i en text. Vissa av dessa modeller baseras s.k. *embeddings* eller inbäddningar. Grovt sett är en inbäddning numeriska representationer av ord eller fraser i texten som används internt i modellen. Eftersom dessa inbäddningar är numeriska gör det använda kvantitativ analys och högpresterande maskininlärningsmodeller för språkbearbetning ('Natural language processing – NLP'), inklusive översättningar. Praktiska exempel är inbäddningar från språkmodeller (Peters m.*fl*., 2018) och dubbelriktade översättningsrepresentationer från transformatormodeller (Devlin *m.fl*., 2019).

Ovanstående tekniker hör till det specialiserade området NLP. Detta är för närvarande ett mycket aktivt forskningsområde med tillämpningar inom dokumentöversättning och generativ AI. I princip kan den här typen av tekniker användas för att extrahera och navigera i information som finns i manualer, rapporter, procedurer, standarddriftsprotokoll och andra dokument som finns tillgängliga på ett avloppsreningsverk. När detta avsnitt skrevs (kring 2022–2023) fanns det inga kända tillämpningar av ovan nämnda tekniker inom vattensektorn.

6.7 ANALYS AV DATAKVALITET

Detta avsnitt ger en översikt över de viktigaste dataanalysmetoderna som kan tillämpas på reningsverk. Två huvudsakliga tillvägagångssätt presenteras: (1) mekanistiska metoder som baseras på fysikaliska modeller och (2) datadrivna metoder som bygger på statistik och empirisk dataanalys. Båda metoderna nyttjar data och genererar olika typer av information.

En mekanistisk metod bygger på vår grundläggande fysikaliska förståelse av systemet. Det kan ses som en vit låda (*white box*) där själva mekanismen för processer eller omvandlingar av indata till utdata är kända. En empirisk metod, som ibland också kallas statistisk metod, bygger däremot på observationer (t.ex. givarsignaler) snarare än teori. De kan ses som en svart låda (*black box*) där korrelationer erhålls genom analys av data. I detta fall används inte kunskap om de fysikaliska mekanismerna som driver systemen.

6.7.1 Mekanistiska metoder

6.7.1.1 Syfte och tillämpning för bedömning av datakvalitet

Mekanistiska modeller försöker beskriva orsakssamband och etablera ett mekanistiskt förhållande mellan in- och utdata. I detta avseende kan de användas för att upptäcka systematiska fel och förbättra

datakvaliteten genom att utnyttja redundant kunskap för kända samband såsom att massa, elektroner och energi bevaras. Datakorrigering med massbalanser syftar till att förbättra kvaliteten på uppmätta rådata genom att säkerställa att de data stämmer överens med kända massbalanser.

6.7.1.2 Metoder

Rimlighetskontroller och massbalanser är typiska exempel där fysikalisk systemförståelse nyttjas. I detta avsnitt presenteras kortfattat de viktigaste rimlighetskontrollerna som bör utföras. Även datakorrigering baserad på linjära massbalanser (*data reconciliation*) beskrivs med tillämpningar för att förbättra datakvaliteten, hantera frånvarande eller förlorade data och upptäcka systematiska fel. Metoden illustreras med ett exempel på en reningsprocess.

6.7.1.2.1 Univariat metod: rimlighetskontroller

Rimlighetskontroller (*sanity checks*) är snabba och grundläggande tester som gör det möjligt att bedöma om ett mätvärde är rimligt. Exempelvis så kan inte koncentrationer eller flödeshastigheter vara negativa. Den grundläggande kontrollen består alltså av att jämföra det uppmätta värdet med de numeriska kriterierna. Om det uppmätta värdet uppfyller kriterierna flaggas en kvalitetsetikett som *giltig*, annars blir kvalitetsetiketten för data *ogiltig*.

Denna typ av kontroll är lämplig att utföra innan den huvudsakliga dataanalysen påbörjas. Med hänsyn till definitionerna och de analytiska bestämningarna kan flera rimlighetskontroller utföras:

- Koncentration av alla ämnen ≥ 0
- Flödeshastigheter ≥ 0
- Totalt suspenderat material (TSS) $\geq$ *volatile suspended solids* (VSS)
- TSS i returslamflödet $\geq$ TSS i aktivslambassängen
- Total COD $\geq$ biokemisk syreförbrukning (BOD infinity) $\geq$ BOD_5
- Total COD $\geq$ filtrerad COD
- Totalkväve (TN) $\geq$ löst kväve
- Totalfosfor (TP) $\geq$ löst fosfor
- Totalt kjeldahlkväve (TKN) $\geq$ ammoniumkväve
- Totalkväve = TKN + NO_3-N + NO_2-N

När en rimlighetskontroll misslyckas ska orsaken till felet eller inkonsekvensen undersökas och motsvarande värde korrigeras (t.ex. genom att använda det sanna värdet om det hittades, eller upprepa mätningen) alternativt raderas (om ingen förklaring eller korrigering är möjlig).

För att ytterligare bedöma data kan dessutom typiska förhållanden mellan variabler beräknas och kontrolleras om de faller inom det typiska intervallet för ett vanligt reningsverk. På så sätt utförs avancerade rimlighetskontroller för att bedöma om resultaten faller inom ett rimligt intervall. Dessa kvoter kan beräknas och jämföras vid olika mätpunkter i processen: inkommande avloppsvatten, försedimenterat avloppsvatten, biobassäng etc.; TN/totalt COD; TP/totalt COD; totalt BOD_5/totalt COD; TSS/totalt COD; VSS/TSS. Typiska referensvärden för dessa kvoter finns i IWA (2012).

6.7.1.2.2 En multivariat metod: massbalanser

Massbalanser, särskilt COD- och kvävemassbalanser, är viktiga både vid dimensionering, modellering, övervakning av anläggningens prestanda och processuppföljning, näringsåtervinning och sidoströmsbehandling (Rieger *et al.*, 2010). För dessa uppgifter är det avgörande att sluta relevanta massbalanser (flöde och koncentrationer) med en vald noggrannhet, t.ex. 5–10 %.

Avvikelser från detta i massbalanserna gör att det går att upptäcka potentiella mätfel som då kan åtgärdas. Om de viktigaste processvariablerna i primärdata uppfyller alla massbalansekvationerna anses datasetet vara tillförlitligt, annars behöver nya beräkningar göras med hjälp av den tillgängliga uppsättningen massbalanser för att få korrekta sekundärdata.

Datakorrigering är en beprövad teknik för att utvärdera överensstämmelsen i insamlade data (Crowe, 1996; Ozyurt & Pike, 2004). I korthet så innebär metoden att mätningar justeras enligt ett optimeringskriterium så att de korrigerade data exakt uppfyller de givna massbalanserna (Crowe, 1996) och därför är mer exakta än de ursprungliga värdena. I detta avsnitt används linjära massbalanser för datakorrigeringen. Mer specifikt innebär den massbalansbaserade dataavstämningen att värdet på uppmätta variabler (med sanna värden x) uppskattas på nytt med hänsyn till deras uppmätta värden y och de massbalanser $h(x)$ som gäller för dessa variabler. De nya skattade variabler betecknade $\hat{x}$ är sådana att den viktade skillnaden mellan uppmätta variabler och uppskattade värden minimeras. Vikterna grupperas i matrisen W, som innehåller mätvariansen, eventuellt kompletterad med ytterligare expertkunskap om mätosäkerheten. På så sätt kan datakorrigeringsproblemet översättas till ett välkänt optimeringsproblem:

$$\left.\begin{array}{r}\min SS(\hat{x})\\ s.t.\ h(\hat{x})=0\end{array}\right\}$$
$$with\ SS(\hat{x})=(y-\hat{x})^T W(y-\hat{x}). \tag{6.1}$$

Datakorrigering används för att korrigera mätdata och för att upptäcka tillfälliga mätfel. Så kallade nyckelvariabler (*key variables*) kan antingen mätas eller ej; identifiering av dessa nyckelvariabler innebär att förbättrade uppskattningar av deras värden ($\hat{x}$) erhålls genom beräkning från andra, uppmätta variabler, givet massbalanserna $h(x)$. De korrigerade mätningarna uppfyller alla begränsningar (dvs. sluter alla massbalanser) och anses därför mer tillförlitliga och exakta jämfört med de ursprungliga värdena. När nyckelvariabler inte mäts kan det vara möjligt att beräkna optimala skattningar genom samma metod. Detta inträffar när de omätta variablerna ingår i en uppsättning oberoende massbalansekvationer vars antal är lika med eller större än antalet omätta variabler.

Typiska exempel på viktiga processvariabler på ett reningsverk som kräver hög noggrannhet är inkommande massflöden och utgående massflöde från den biologiska reningen samt massflödet av överskottsslam. Syrgasbehovet för COD- och kvävereningen är också viktiga processvariabler och måste därför normalt uppskattas.

De vanligaste massbalanserna på reningsverk är konservativa storheter såsom totala massflöden (eller volymetriska flöden, om densiteten kan antas vara konstant) eller massflöden av total COD (*mCOD*), totalkväve (*mTN*), total fosfor (*mTP*) och torrsubstans. Massbalanser kan ställas upp över enskilda och/eller kombinerade enhetsprocesser. Detta kräver att de resulterande massbalanserna innehåller alla nyckelvariabler.

Viktiga förutsättningar för en lyckad datakorrigering är tillräcklig redundans i mätningarna och att processen är i jämvikt. Redundans för variabler innebär att deras uppmätta värden också kan beräknas utifrån andra (uppmätta) variabler. På många reningsverk finns det dock ofta inte tillräckligt med uppmätta data från början, och ytterligare mätningar måste vanligtvis utföras för att säkerställa den grad av redundans som krävs för dataavstämning. Le *et al* (2018) föreslog en försöksplanering som garanterar att nyckelvariabler kan identifieras i efterföljande datakorrigering, genom att uppfylla tillräcklig redundans.

Datakorrigeringen förutsätter vidare att alla mätningar är normalfördelade kring ett nominellt värde och har en känd eller uppskattad mätosäkerhet. Eftersom massbalanserna vanligtvis är stationära är en viktig förutsättning för att tillämpa datakorrigering att anläggningen är i stationär drift under den period för vilken data används. Det får t.ex. inte förekomma något kraftigt regn eller någon förändring i slamflödet. Minst 2 veckors data eller en period som täcker en slamålder bör användas.

Innan man drar slutsatser från korrigerade data krävs en validering. Av denna anledning åtföljs datakorrigering ofta av statistiska tester för att upptäcka tillfälliga mätfel. De vanligast förekommande är det globala statistiska testet, mätvärdestestet och nodaltestet. För en mer detaljerad förklaring av datakorrigering och de tillhörande statistiska testerna, se Le *m.fl.* (2022).

6.7.1.2.2.1 Illustrativt exempel

Den studerade processen kommer från (Le *et al.*, 2022) och kan beskrivas av 16 massbalanser kring enskilda och kombinerade enhetsprocesser (Tabell 6.2). Flödesbalanserna ställdes upp kring de huvudsakliga enhetsprocesserna; gränserna för COD-, TP-, TN- och TKN-balanserna ställdes upp över kombinationer av enhetsprocesser, vilket indikeras av den streckade linjen i Figur 6.25. Mätningarna visas i Tabell 6.3. Datakorrigeringen kunde justera 22 av 32 uppmätta variabler, vilket innebär att deras korrigerade värden uppfyllde massbalanserna och deras precision förbättrades. Dessutom gjorde datakorrigeringen det möjligt att beräkna nio variabler som saknade mätningar.

Även om dessa variabler också kunde ha beräknats utifrån en vanlig massbalanser, så gav datakorrigeringen en högre precision i de korrigerade data.

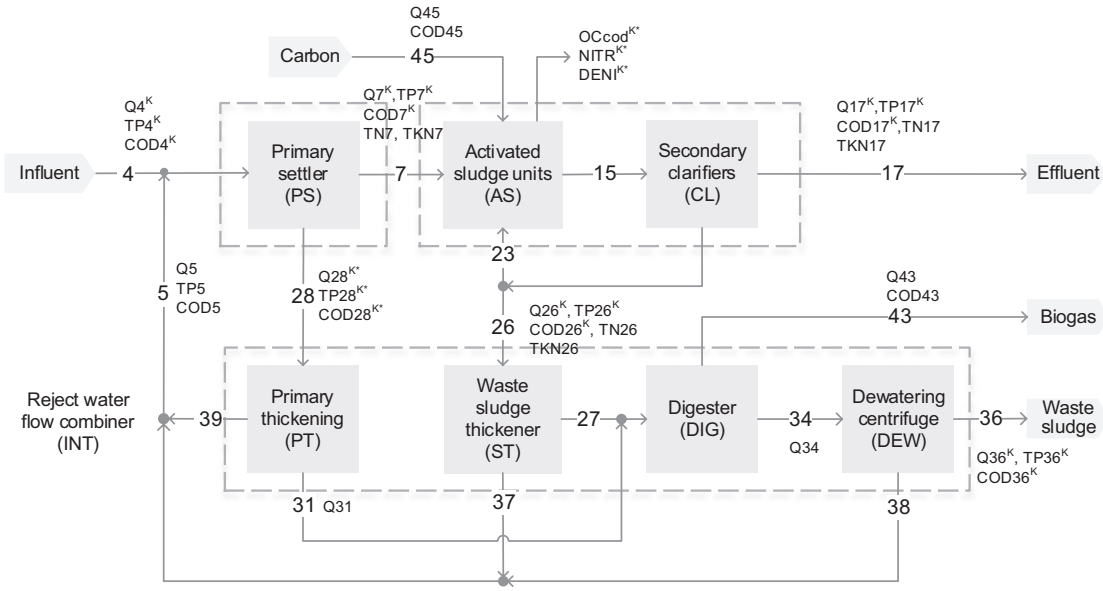
W-matrisen kan beräknas direkt från insamlade data som standardavvikelsen från mätningens medelvärde. I vissa fall kan expertkunskap om mätosäkerheten användas som indata för att bestämma W-matrisen. Ett exempel på hur man bildar en W-matris från data från en mätkampanj finns i Forouzanmehr *et al.* (2021).

Alla tre statistiska test identifierade potentiella fel i mätdata (Tabell 6.4). Det globala testet gav en övergripande indikation om att kan det finnas mätfel i datasetet. Mättestet (*measurement test*) pekade ut två specifika variabler, TP_{26} och TP_7, med misstänkt felaktiga mätvärden. Nodaltestet indikerade att det fanns fel i massbalansen runt aktivslambassängen: $mTP_7 - mTP_{17} - mTP_{26} = 0$. De korrigerade värdena för mTP_7 och mTP_{26} avvek i själva verket avsevärt från de uppmätta mätvärdena (Tabell 6.3). Strategier enligt (Narasimhan & Jordache, 2000) användes sedan för att ta bort tillfälliga mätfel. I det här fallet ledde detta till att ett fel upptäcktes i mätningen av totalfosfor (mTP_{26}), vilket i sin tur orsakade ett misstänkt fel för mTP_7 (på grund av utsmetningseffekten, *smearing effect*).

Tabell 6.2 Massbalanser för den studerade reningsprocessen.

Nr.	Enhet	Bilineär massbalans	Enhet
1	Försedimentering (PS)	$Q_4 + Q_5 - Q_7 - Q_{28}$	m/d^3
2	Aktivslambassäng (AS)	$Q7 - Q17 - Q26$	
3	Mellansedimentering (CL)	$Q26 - Q27 - Q37$	
4	Förtjockare ö.slam (ST)	$Q28 - Q31 - Q39$	
5	Förtjockare för primärslam (PT)	$Q27 + Q31 - Q34$	
6	Rötkammare (DIG)	$Q34 - Q36 - Q38$	
7	Centrifug (DEW)	$Q37 + Q38 + Q39 + Q40 - Q5$	
8	Försedimentering (PS)	$mTP_4 + mTP_5 - mTP_7 - mTP_{28}$	kg/d
9	Aktivslambassäng (AS)	$mTP_7 - mTP_{17} - mTP_{26}$	
10	Mellansedimentering (CL)	$mTP_{26} + mTP_{28} - mTP_5 - mTP_{36}$	
11	Försedimentering (PS)	$mCOD_4 + mCOD_5 - mCOD_7 - mCOD_{28}$	
12	Aktivslambassäng (AS)	$mCOD_7 + mCOD_{45} - mCOD_{17} - mCOD_{26} - OC_{cod} - 2{,}86 \times DENI$	
13	Mellansedimentering (CL)	$mCOD_{26} + mCOD_{28} - mCOD_5 - mCOD_{36} - mCOD_{43}$	
14	Aktivslambassäng (AS)	$mTN_7 - mTN_{17} - mTN_{26} - DENI$	
15	Aktivslambassäng (AS)	$mTKN_7 - mTKN_{17} - mTKN_{26} - NITR$	
16	Hela reningsprocssen	$OC_{net} - OC_{cod} - 4{,}57 \times NITR$	

Q = flöde; mTP = massflöde av totalfosfor; mCOD = massflöde av total-COD; mTN = massflöde av totalkväve; mTKN = massflöde av totalt kjeldahlkväve. Numret i den första kolumnen anger flödesnummer i procsschema; DENI = denitrifierat kväve; NITR = nitrifierat kväve; OCcod = syre som krävs för COD-reduktion; OCnet = totalt syrebehov. Massflödestermen är produkten av uppmätt flöde och motsvarande koncentration.



Figur 6.25 Flödesschema för processen i Le (2019). De variabler som anges i figuren mäts i termer av flöde och koncentrationer. De streckade linjerna representerar gränserna för massbalanserna för COD, totalfosfor (TP) och kväve (TN och TKN).

Efter att detta fel eliminerats överensstämde de uppmätta och korrigerade värdena. Detta indikerar två viktiga resultat: (1) anläggningens flödesschema var korrekt och (2) mätningarna var tillförlitliga och kunde användas vidare för andra ändamål, t.ex. modellering och processoptimering.

6.7.1.3 Potential med datakorrigering på reningsverk

Även om datakorrigering har studerats i stor utsträckning inom kemiteknik i årtionden, se t.ex. Madron *et al.* (1977); van der Heijden *et al.* (1994), har metoden hittills inte använts på reningsverk i någon större utsträckning. Det finns dock mycket att vinna på mer tillförlitliga data för vidare tillämpning vid utvärdering och optimering av processer, modellering och benchmarking. Behovet och det möjliga mervärdet av datakorrigering ökar med den nuvarande digitaliseringstrenden, som skapar ett massivt dataflöde. Paradigmskiftet från reningsverk som enbart fokuserar på vattenrening, till anläggningar som verkligen återvinner resurser i avloppsvattnet ökar fördelarna med automatiserad datakorrigering, eftersom tillförlitlig information behövs för flera olika processmål.

Flera viktiga steg har tagits mot ett automatiserat verktyg för datakorrigering. Strukturerade metoder för experimentell design och för datakorrigering har tagits fram av Le *et al.* (2018). Tillämpningen på flera fullskaliga avloppsreningsverk har visat sig vara framgångsrik (t.ex. Forouzanmehr *et al.*, 2021). Mjukvara som implementerats i MATLAB har visat sig tillräckligt flexibla för att hantera de flesta problem relaterade till datakorrigering, i alla fall offline.

För att nyttiggöra metodiken med datakorrigering krävs ytterligare utveckling av fristående mjukvara för realtidstillämpningar. Fokus bör ligga på förbehandlingsstegen av data, för vilka en modul bör skräddarsys för att samla in, förbehandla, automatiskt organisera metadata och sammantaget göra data redo för datakorrigeringsmetoden. Eftersom flödesmätningarna i en välinstrumenterad anläggning vanligtvis registreras med hög frekvens bör just flödesmätningar vara enklast att tillämpa i en onlinekorrigering. De korrigerade mätvärdena för flödeshastigheten är av största vikt vid balansering av massflöden och koncentrationer eftersom flödeshastigheter förekommer i alla relaterade balanser.

Tabell 6.3 Mätdata och korrigerade data för exemplet.

Numrering och namn i processdiagram		*Q*	mTP	mCOD	mTN	mTKN
		m³/dag	kg/dag	kg/dag	kg/dag	kg/dag
4	Inkommande	57 008 ± 3 421	446 ± 41	30 836 ± 3 084		
		56 943 ± 1 819	441 ± 33	30 796 ± 2 493		
5	Rejektvatten	6 288 ± 441	332 ± 34	9 076 ± 908		
		6 371 ± 285	332 ± 34	9 076 ± 908		
7	Försedimenterat vatten	58 101 ± 3 487	556 ± 56	21 461 ± 1 932	3 182 ± 287	3,178 ± 286
		58 177 ± 1 819	434 ± 40	21 477 ± 1 796		
15	Inflöde till sedimentering	119 575 ± 8 371				
		119 635 ± 4 089				
17	Utgående vatten	56 906 ± 2 846	112 ± 12	1882 ± 189	665 ± 67	134 ± 13
		56 895 ± 1 819	117 ± 11			
23	Returslamflöde	61 475 ± 4 304				
		61 459 ± 3 846				
26	Överskottsslam (WAS)	1 285 ± 78	193 ± 56	11 120 ± 2 224	971 ± 195	959 ± 144
		1 283 ± 77	317 ± 40	11 099 ± 2 014		
27	Förtjockat överskottsslam	133 ± 8				
		133 ± 8				
28	Primärslam	5 194 ± 364	338 ± 61	18 450 ± 3 750		
		5 137 ± 283				
31	Förtjockat primärslam	182 ± 10				
		182 ± 9				
34	Rötat slam	315 ± 23				
		315 ± 11				
36	Utgående slam	48 ± 9	312 ± 57	8 910 ± 3 030		
		49 ± 9	324 ± 34	8 949 ± 2 471		
37	Centrifug slam	1 150 ± 77				
38	Centrifug avvattning	267 ± 14				
39	Dekantat förtjockare	4 956 ± 283				
43	Biogas	6 372 ± 199		11 470 ± 230		
				11 470 ± 230		
45	Extern kolkälla			1 780 ± 89		
46	Denitrifierat kväve (DENI)				1 547 ± 353	
47	Nitrifierat kväve (NITR)				2 086 ± 321	
48	Syre för COD-reduktion (OCcod)			5 817 ± 3 121		
49	Total syreförbrukning (OCnet)			15 346 ± 3 447		

Siffror i svart = uppmätta värden (y), som beräknas som medelvärde ± standardfel för medelvärdet av uppmätta data för mätkampanjens period, grönt = korrigerade värden ($\hat{x}$); blått = beräknade värden för omätta variabler; rött = mätningar som potentiellt innehåller ett fel. Massflödestermen är produkten av uppmätt flöde och motsvarande koncentration.

Tabell 6.4 Resultat av feldetektion.

Test	Återstoden från	Graden av frihet	Kritisk p-Värde	Indikation
Globalt test	$SS(\hat{x})$ (från ekvation (6.1))	11	0,05	Fel upptäckt med p-värde = 0,0027
Mätvärdesstest	$y - \hat{x}$ (för varje enskilt uppmätt variabel)	18	0,0014	mTP7 och mTP26 med p-värde = 0,00045
Nodalt test	$h(y)$ (för varje massbalans som innehåller uppmätta variabler)	21	0,0012	mTP7 – mTP17 – mTP26 = 0 med p-värde = 0,00044

6.7.2 Datadrivna metoder

6.7.2.1 Syfte och tillämpning för bedömning av datakvalitet

Datadrivna metoder (även kallade empiriska metoder) syftar till att fastställa statistiska samband och korrelationer mellan indata och utdata. Till skillnad från mekanistiska modeller bestäms deras struktur enbart av data, och metoderna är tillämpliga på processdata från reningsverk för att bedöma datakvaliteten, prognostisera relevanta parametrar av intresse och ta fram ytterligare information som kan användas för, t.ex. modellering eller processövervakning.

Datadrivna metoder baseras på ett dataset som kan innehålla givardata, laboratoriedata och metadata som beskriver den studerade reningsprocessen. Tillämpningarna kan grovt delas in i två grupper:

- *Prediktion*: datasetet består av indata och utdata där målet är att kunna prediktera okända utdata, givet enbart indata. Det finns många tillgängliga prediktionsalgoritmer såsom neurala nätverk, *kernel-based methods* och *decision trees*.
- *Utforskning av data*: datasetet består enbart av indata och målet är att upptäcka intressanta egenskaper hos dataposterna, t.ex. kluster, underliggande samband och fördelningar. Även här finns det många tillgängliga algoritmer såsom linjär/olinjär dimensionsreduktion, *Gaussian mixture models*, *self-organizing maps* m.fl.

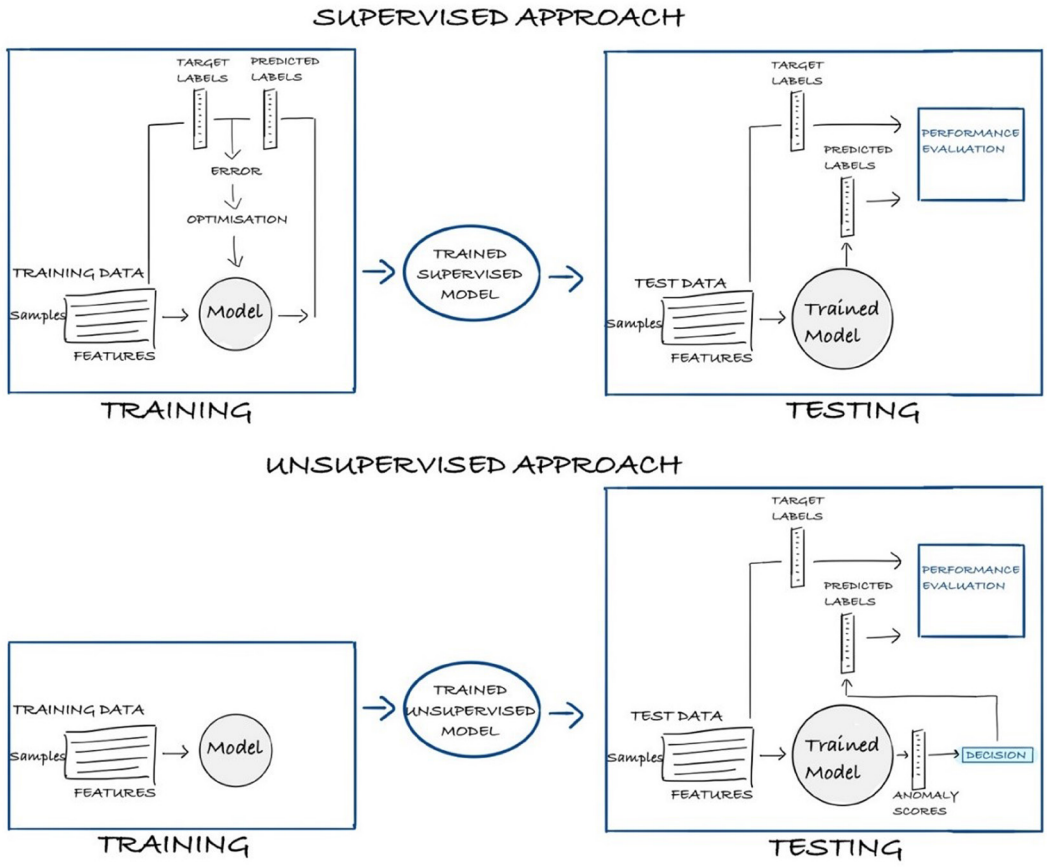
Oavsett metod så används en uppsättning med s.k. träningsdata för att bygga upp och träna modellen (in- och utdata för prediktion, och endast indata för datautforskning). Beroende på vilken typ av inlärningsmetod som används i träningsfasen kategoriseras de flesta metoderna som antingen *övervakade* (*supervised*) eller *oövervakade* (*unsupervised*), vilket illustreras i Figur 6.26. I övervakade metoder så används ett annoterat dataset med kända avvikelser/etiketter för att träna modellen, medan den oövervakade metoden inte kräver några annoteringar. I korthet så nyttjar övervakade metoder uppmärkta in- och utdata för att träna modellen till att antingen prediktera en klass (klassificering) eller värdet på en kontinuerlig variabel (regression). Oövervakade modeller används för att utforska omärkta dataset för att upptäcka dolda mönster i data. Tabell 6.5 sammanfattar de viktigaste skillnaderna mellan övervakade (*supervised*) och oövervakade (*unsupervised*) datadrivna metoder.

Så snart modellen har byggts upp valideras och testas den med ett oberoende valideringsdataset och ett testdataset. Notera att valideringsdatasetet måste innehålla annoterade data. Kvaliteten på inlärningsprocessen och metodens prestanda utvärderas sedan med olika prestandamått som beskrivs närmare i Avsnitt 6.7.4.

I det här avsnittet ges en översikt över olika datadrivna metoder som används för att utforska data och för prediktion med reningsverkstillämpningar.

6.7.2.2 Övervakade tillvägagångssätt

Övervakade metoder baseras på en uppsättning $D = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^{N}$ dataposter (Tabell 6.6) med index n. Varje datapar $(\boldsymbol{x}_n, y_n)$ består av en M-dimensionell indatavektor $\boldsymbol{x}_n$ med motsvarande utdata



Figur 6.26 Övervakade ('supervised') och oövervakade ('unsupervised') metoder för datakvalitet beroende på tillgången till annoterade data i datasetet. I figuren anges 'samples' och 'features', vilka motsvarar dataposter, respektive variabler.

y_n. Vanligtvis är komponenterna i indatavektorn $\boldsymbol{x}_n = (x_{n,1}, x_{n,2}, \ldots, x_{n,m}, \ldots, x_{n,M})$ mätvärden för lättillgängliga processvariabler, medans målvariablen och utdata y_n är de värden för en svårmätbar variabel, som antingen tillhandahålls av en domänexpert eller erhålls från dyra och tidskrävande laboratorieanalyser. De två huvudtillämpningarna för övervakad inlärning, klassificering och regression, beskrivs nedan och illustreras i Figur 6.27.

6.7.2.2.1 Klassificering

Vid klassificering antar målvariabeln y_n endast diskreta värden. Olika målvärden används för att annotera olika s.k. klasser. Begreppet klass används för att beteckna dataposter, par av indata och utdata, med gemensamma egenskaper. Annoteringar kan användas för att beskriva olika driftsförhållanden på en anläggning (t.ex. hög respektive låg belastning), liksom andra egenskaper som deras status (t.ex. normala eller avvikande förhållanden). Klassificering handlar då om att modellera förhållandet f mellan indata (mätningar) och målvariabeln (t.ex. driftförhållanden) och använda det för att förutsäga tillståndet för målvariabeln för en ny uppsättning indata. I det perspektivet handlar kvalitetsbedömning av nya data om att avgöra om en ny datapunkt tillhör en viss klass eller inte.

Tabell 6.5 Viktiga skillnader mellan övervakade och oövervakade metoder för automatiserad bedömning av datakvalitet.

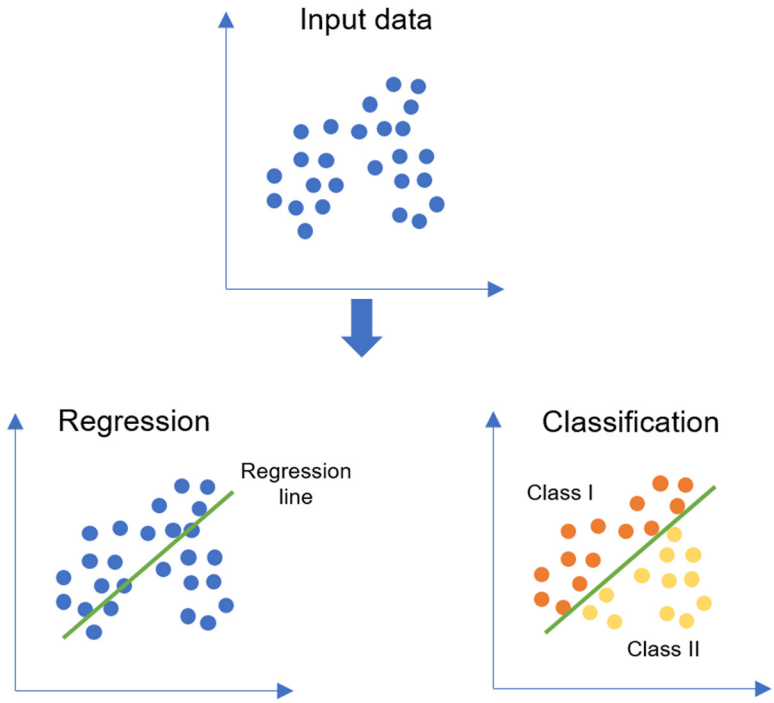
	Övervakad	Oövervakad
Annoterade data	Använder annoterade in- och utdata för att modelluppbyggnad/träning	Använder omärkta data
Syfte med dataanalys	Klassificering av data givet inlärda mönster eller prediktion av utdata	Upptäcka dolda mönster och samband i data
Handpåläggning	Kräver ofta subjektiva beslut för att annotera data på rätt sätt	Modelluppbyggnad sker automatiskt. Handpåläggning krävs för att validering utdata
Noggrannhet	Generellt högre på grund av information som fångas upp i annoteringarna	Generellt lägre på grund av omärkta data utan annoteringar
Komplexitet	Enklare men resurskrävande på grund av behovet av annoterade data. Kan vara tidskrävande att träna och kan kräva återkommande uppdateringar	Beräkningsmässigt komplext, stor mängd träningsdata krävs för att producera bra resultat
Databehov	In- och utdata	Endast indata
Lärande	Modellträning baseras på annoteringar som görs offline av av domänexperter	Modellträning kan ske i realtid
Algoritmer	Klassificering, regression	Klustring, reduktion av antalet variabler, övervakning
Avvikelsedetektion	Användaren måste ge exempel på träningsdata för varje enskild typ av avvikelse som ska upptäckas	Kan upptäcka alla typer av avvikelser, även sådana som aldrig tidigare har setts
Största utmaningen	Skapa ett komplett annoterat och funktionellt träningsdataset	Bestämma vad som anses vara acceptabelt eller typiskt (‘ground truth’)

Tabell 6.6 En uppsättning *D* av *N* datapunkter som indata och målvariabler $(x_n, y)_{\cdot n}$

	Processvariabel 1	Processvariabel 2	…	Processvariabel *m*	…	Processvariabel *M*	Målvariabel
$\boldsymbol{x}_1$	$x_{1,1}$	$x_{1,2}$		$x_{1,M}$		$x_{1,M}$	y_1
$\boldsymbol{x}_2$	$x_{2,1}$	$x_{2,2}$		$x_{2,M}$		$x_{2,M}$	y_2
	- - -	- - -		- - -		- - -	- - -
$\boldsymbol{x}_n$	$x_{n,1}$	$x_{n,2}$		$x_{n,M}$		$x_{n,M}$	y_n
	- - -	- - -		- - -		- - -	- - -
$\boldsymbol{x}_N$	$x_{N,1}$	$x_{N,1}$		$x_{N,M}$		$x_{N,M}$	y_N

Det finns många olika klassificeringsmetoder (Friedman *et al.*, 2001). Valet av metod bör baseras på tillämpningen och på egenskaperna och tillgänglighet av data. Ofta beror dock valet på användarens kunskap om olika metoder och på hur tillgängliga de är i termer av programvaruimplementeringar. Några av de mest populära klassificeringsmetoderna har reningsverkstillämpningar: *naive Bayes* (NB; Carvajal *et al.*, 2015), *logistic regression* (LR; Salman & Salem, 2012), *k-nearest neighbour* (kNN; Kim *et al.*, 2016), *random forest* (RF; Vitorino *et al.*, 2014), *support vector machines* (Fan *et al.*, 2004; Tian *et al.*, 2012) och *artificial neural networks* (ANN, Boger, 1992; Hanbay *et al.*, 2008).

Naive bayes är en generativ modell som syftar till att uppskatta sannolikheten för att ett visst in-utdatapar tillhör en viss klass, givet att indata observeras. LR, kNN, RF och ANN är s.k.



Figur 6.27 Exempel på övervakade metoder.

diskriminativa modeller som syftar till att beskriva samband för den, vanligtvis icke-linjära, funktion som kan ange vilken klass en datapost sak tilldelas. Enligt Ng och Jordan (2002) förväntas diskriminativa modeller i allmänhet ge bättre resultat än generativa modeller (a) när antagandena i de generativa modellerna inte är giltiga och (b) när stora och representativa dataset finns tillgängliga. ANN-modeller är välkända för sin flexibilitet och har visat sig kunna överträffa andra modeller, särskilt när det gäller stora datamängder. Det är dock viktigt att notera att flexibilitet kan leda till modeller som är alltför specialiserade på en viss uppsättning data, ett fenomen som ofta benämns som *overfitting*.

När klassificeringsmodellerna används i realtidsapplikationer måste de kunna hantera mätfel och annoteringsfel. De kan också kräva återkommande uppdateringar för att inkludera nya annoteringar och träningsdata i modellen. Detta beror på att nya data återspeglar förändringar i miljön och de omgivande förhållandena (t.ex. säsongsvariationer eller förändringar i avloppsvattnets sammansättning) och förändringar i styrningen av enheterna, och därmed förändringar i förhållandet mellan indata och mål. En viktig faktor att ta hänsyn till när man väljer en klassificeringsmodell är dess beräkningskomplexitet. Enkla modeller som LR och NB är vanligtvis lätta att träna och uppdatera. Omvänt är det ofta mer krävande att träna och underhålla mer komplexa modeller som ANN, eftersom deras träningsprocedur baseras på icke-konvex optimering.

Förutom de klassiska metoder som nämns ovan har nyare metoder baserade på djupa neurala nätverk, s.k. *deep learning*-metoder (Goodfellow *et al.*, 2016), blivit alltmer populära. Dessa metoder beskriver icke-linjär dynamik och komplexa interaktioner mellan variablerna (Harrou *et al.*, 2018; Mamandipoor *et al.*, 2020; Zhang *et al.*, 2018). Det är värt att nämna att dessa metoder ännu inte har tillämpats i praktiken.

6.7.2.2.2 Regression

Vid regression tillåts målvariabeln anta kontinuerliga värden till skillnad från klassificering som angav ett visst antal diskreta klasser. Men på samma sätt som i klassificering så är har modellen ett antal parametrar som behöver inlärningen anpassas (tränas) för att anpassas för att beskriva data på bästa sätt. Det enklaste regressionsproblemet är att anpassa en linje som bäst beskriver en uppsättning med tre datapunkter. Den funktion (modell) som ska bestämmas är i det här fallet ekvationen för en linje som parametriseras av lutningen och skärningen med y-axeln. Deras värden kan direkt fås linjär regression med minsta kvadratmetoden (Kay, 1998). Mer sofistikerade regressionsmetoder generaliserar denna idé till mer komplexa, icke-linjära funktioner, dessutom med flera variabler i indata. När modellen väl beskriver sambandet mellan in- och utdata kan den användas för att prediktera målvärdena direkt från indata, istället för att mäta eller bestämma dem direkt. De predikterade värden kan sedan jämföras med aktuella mätningar för att upptäcka onormala variationer och avvikelser av målvariabeln. Figur 6.28 ger ett illustrativt exempel på faktiska värden tillsammans med de förväntade värdena enligt den regressionsmodell som anpassats till de ursprungliga värdena.

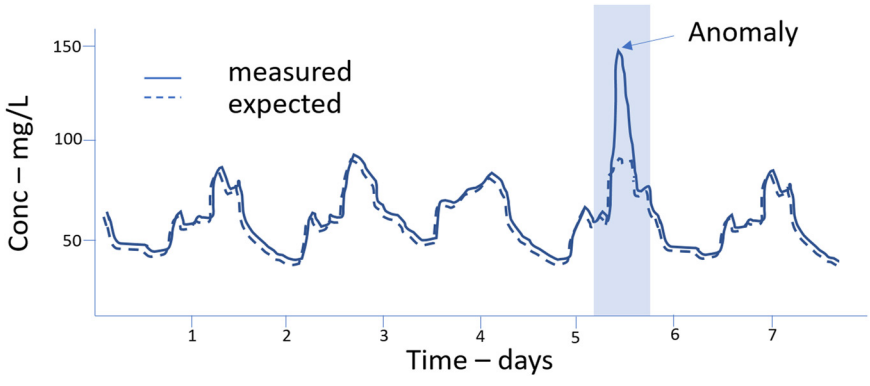
Skillnader mellan uppmätta och predikterade värden som är större än fördefinierade tröskelvärdena indikerar avvikelser.

Några av de mest populära regressionsmetoderna är linjära och olinjära regressionsmetoder, regressionsträd (RT), *random forests*, artificiella neurala nätverk, *kernelregression*, *Gaussian process regression* (GPR) och *Partial least square* (PLS, Wold *et al.*, 2001), se Friedman *et al* (2001) för en detaljerad beskrivning av dessa regressionsmetoder.

Några reningsverkstillämpningar av dessa metoder finns beskrivna i litteraturen. Granata *m.fl* (2017) uppskattade t.ex. nyckeltal för avloppsvattenrening med hjälp av RT och 'support vector regression'. Zhou *et al* (2019) predikterade inkommande dygnsflöde med hjälp av RF, givet väderleksprognoser och hushållens vattenanvändningsmönster.

Till skillnad från övervakade metoder som optimerar parametervärden är GPR en datamodelleringsteknik som beskriver sambandet mellan in- och utdata som en uppsättning slumpmässiga variabler, som inkluderar den osäkerhet som finns i utdata (Rasmussen & Williams, 2005). GPR har tillämpats på data från reningsverk, bland annat för övervakning och feldetektering på reningsverk (Samuelsson *et al.*, 2017; Zambrano *et al.*, 2019).

I likhet med de flesta regressionstekniker antas data $D = \{(x_n, y_n)\}$ $(n = 1)^N$ vara felfria, medan utdata modelleras som $y_n = f(x_n) + \text{brus}_n$. Utdata kan beskrivas som en simultan normalfördelning och gaussisk process med medelvärdesfunktion $\mu(x)$ och en *kovariansfunktion* $k(x, x)$. Kovariansfunktionen väljs av användaren på ett sådant sätt att dess element $k(x_n, x_{n'})$ uttrycker likheten mellan par av indata. Kovariansfunktionen bestämmer strukturen hos en GPR-modell



Figur 6.28 Exempel på regressionsmetoder för att upptäcka avvikelser i en processvariabel.

där man kan använda förkunskap om processen för att välja en lämplig kovariansfunktion. När specifika dataegenskaper behöver fångas upp bör en kombination av olika kovariansfunktioner övervägas (se Lloyd *et al.*, 2014). Ett urval av vanligt förekommande kovariansfunktioner finns i Samuelsson *et al.* (2017), där GPR användes för feldetektion med hjälp av mätningar av flöde och ammoniumhalten i en aktivslamprocess.

Kovariansfunktionen representeras av en matris $K(X)$ av storleken $N \times N$ vars element $k(x_n, x_{n'})$ är en vektor av okända koefficienter eller hyperparametrar p som karakteriserar funktionen k. Ett typiskt val av elementen i K är den s.k. 'squared exponential kernel') där likheten mellan två indatapar beskrivs av $k(x_n, x_{n'}) = p_1 \exp[-p_2 (||x_n - x_{n'}||)]$, med hyperparametervärdena $p = (p_1, p_2)$. En enkel metod för att välja hyperparametrar från data är att välja den vektor p vars sannolikhet maximeras, givet in- och utdata i träningsdatasetet. Vid regression är målet att uppskatta utdata y för en tidigare osedd indata x: detta ger en fördelning som i GPR är en gaussisk fördelning vars medelvärde och varians då kan bestämmas med hjälp av en matrisberäkning.

6.7.2.3 Oövervakade metoder

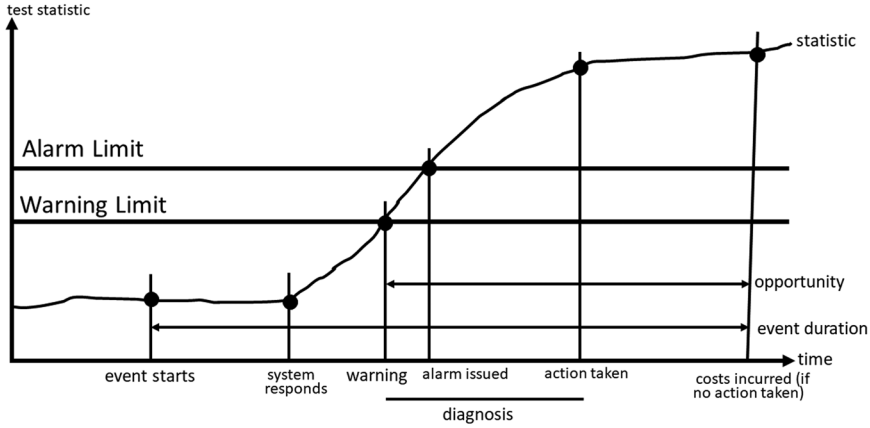
Oövervakade datadrivna metoder (*unsupervised methods*) använder enbart indata i modellbygget, d.v.s. inga annoterade utdata används i analysen. Dessa metoder kan användas för att utforska och upptäcka vissa egenskaper hos data, främst förekomsten av grupper av data med liknande egenskaper, men det går även att reducerade antalet variabler med bibehållen korrelation i data (Bishop, 2006). Först ges en översikt över de enklaste oövervakade metoderna och deras användning för univariat övervakning, och därefter beskrivs några mer avancerade metoder för *feature extraction* såsom den välkända *principal component analysis* (PCA) (Jackson, 1980) och dess tillämpning för multivariat övervakning.

6.7.2.3.1 Univariata övervakningsmetoder

Univariata övervakningsmetoder analyserar en processvariabel i taget, en signal, med syftet att detektera förändringar i den variabeln. Sådana metoder använts framgångsrikt under knappt 100 år i många branscher som en del av kvalitetskontrollen, för feldetektering i produktionsprocessen, för att förbättra affärs- och bankprocesser och för allmän anomalidetektering i processindustrin (Baklouti *et al.*, 2018; Corominas *et al.*, 2011, 2018; Newhart *et al.*, 2019; Sánchez-Fernández *et al.*, 2018; Von Sperling *et al.*, 2020). Grunderna i hur dessa metoder kan tillämpas i praktiken är ofta en del av introduktionskurserna i statistik på ingenjörsutbildningen.

Utgångspunkten för dessa metoder är att först fastställa vad som är normal drift och sedan upptäcka en förändring som avviker från detta normaltillstånd och beteckna sådana avvikelser som *avvikande* eller *felaktiga*. Detektionsförmågan skapas genom att analysera signaldata med enkla modeller. Modellerna bygger ofta på antagandet att (1) signalens medelvärde är konstant över tiden och (2) att dess spridning, eller variabilitet, inte heller förändras över tiden (Montgomery, 2019). En förändring upptäcks sedan antingen när (1) signalens medelvärde förändras eller driver bort från sin antagna fasta punkt, (2), eller när signalens varians förändras över tid, eller (3) när både medelvärdet och variansen förändras. Dessa grundläggande modeller diskuteras i följande avsnitt. Mer avancerade modeller som inte bygger på dessa antagande (t.ex. modeller som tillåter att medelvärdet varierar över tid) är också lättillgängliga, även om de ofta förbises (Box *et al.*, 2015).

De förändringar som upptäcks i en signal hittas genom att jämför värdet med olika gränser som satts upp i syfte att upptäcka avvikande händelser. Data visualiseras ofta med ett styrdiagram (*process control chart*) för att underlätta för processoperatörer och ingenjörer att tillämpa metoden. När förändringar inträffar kan användaren enkelt se att en avvikelse har uppstått, så att det går att undersöka om avvikelsen verkligen är ett fel och sedan identifiera den troliga orsaken processtörningen. En givare kan t.ex. ge avvikande mätvärden som kan upptäckas i ett styrdiagram som jämför givardata med laboratoriedata. Den operatör som tolkar styrdiagrammet får då en indikation på att vidta åtgärder, såsom att kalibrera om givaren (Thomann *et al.*, 2002). Om visualisering anses överflödigt går det att automatisera avvikelsedetektion med univariata metoder.



Figur 6.29 Grundläggande användning av styrdiagram (*källa*: Miletic *et al.*, 2008).

Ett generellt exempel på hur styrdiagram kan användas ges i Figur 6.29. I styrdiagrammet finns en övre gräns (*warning limit*). När en störning sker (*event starts*) går det en tid innan händelsen detekteras i styrdiagrammet som en varning.

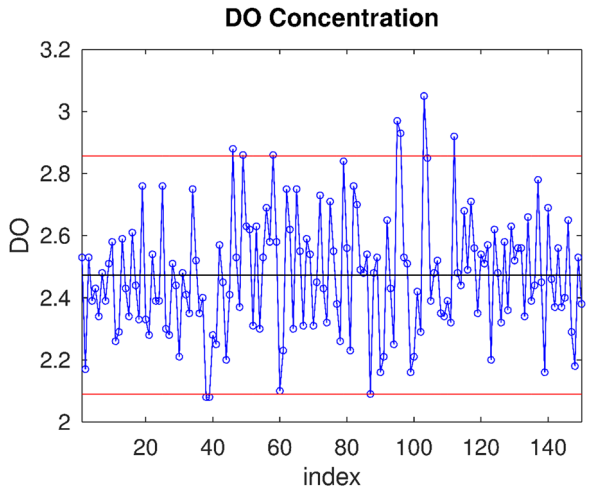
När onormala data väl har upptäckts kan åtgärder vidtas omedelbart, förhoppningsvis innan störningen orsaker negativa miljömässiga eller ekonomiska konsekvenser.

Ett styrdiagram kan användas för avvikelsedetektion för alla modeller och metoder som genererar en signal som kan tolkas utifrån förväntade normalvärden och tröskelvärden. Ett styrdiagram kan förstås även ha två gränser (ett övre och ett nedre tröskelvärde) eller två varningsgränser. I praktiken används ofta även en algoritm med s.k. *run-rules* som bestämmer hur många avvikelser som krävs för att trigga ett alarm. Det finns många varianter där några av de mest populära metoderna är de s.k. Western Electric-reglerna som till stor del utvecklats i industrin och inte akademin (Mitra, 2008). Dessa metoder syftar till att balansera reaktionsförmågan vid en avvikelse och är därför inte inriktade på teoretiska förutsättningar utan på hur de praktiskt kan implementeras och användas av andra än statistiska experter.

Univariata övervakningsmetoder bör tolkas som beräkningssystem som använder feldetekteringsmodeller för att identifiera avvikelser och stödja uppföljande motåtgärder. Förmågan att analysera fel i en signaldata (t.ex. att fastställa om ett fel beror på en givaren, en störning i den biologiska processen eller en kombination av dessa) beror på den processmodell som ligger till grund för styrdiagrammet (t.ex. ett styrdiagram med fast medelvärde). En *supervised* processmodell kan tas fram med stöd av en expert som annoterar perioder med normal drift, alternativt med stöd av en algoritm med en s.k. *bootstrapping*-metod. När processdata annoteras automatiskt är det dock viktigt att notera det krävs ett iterativt tillvägagångssätt med successiv justering/träning av styrdiagrammet och dess gränser. Denna aktivitet behöver kombineras med granskning av antalet larm och processuppföljning till dess att modellen (styrdiagrammet) beskriver det önskade driftmålet (Miletic *et al.*, 2008). För att ett tillförlitligt datainsamlings- och loggningssystem ska fungera i praktiken krävs oundvikligen periodiskt modellunderhåll (dvs. justering av styrdiagrammet) och olika uppgraderingar av datakommunikationssystemet. Dessa aktiviteter bör kontinuerligt kontrolleras och valideras i förhållande till syftet med själva styrdiagrammet. Utbildning av användare och utformning av gränssnitt kan också vara viktigt i många praktiska tillämpningar.

6.7.2.3.1.1 Shewhart styrdiagram

Det vanligaste styrdiagrammet är det som utvecklades av Walter Shewhart på Bell Labs 1931 (Newhart *et al.*, 2019) och även anses vara den äldsta statistiska tekniken för processövervakning. I Shewharts



Figur 6.30 Exempel på ett Shewhart-styrdiagram för uppmätt syrehalt (blått) på ett reningsverk. Gränsvärdena UCL och LCL visas i rött och CL i svart. Diagrammet baseras på en s.k. *t*-fördelning med ett 95 % konfidensintervall för gränserna.

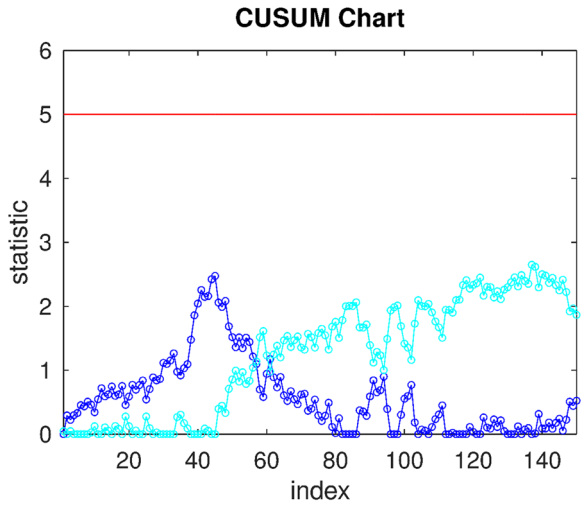
styrdiagram visualiseras dataposter x_n eller deras löpande medelvärden x_n under ett kort tidsintervall tillsammans med en mittlinje (*centre line – CL*) som motsvarar det övergripande datamedelvärdet (dvs. medelvärdet, μ_0, av medelvärdena x_n). Mittlinjen antas motsvara en konstant och normal driftpunkt, och tanken är att följa hur data avviker från den, där då stora avvikelser antas vara relaterade till onormala förhållanden. Detta görs genom att lägga till en övre kontrollgräns (UCL) och en nedre kontrollgränsen (LCL) i diagrammet, vilka beräknas genom att lägga till och dra ifrån k antal (k väljs vanligtvis till 2 eller 3) av variabelns standardavvikelser (σ) från medelvärdet, det vill säga $\mu_0 \pm 2\sigma$ (Montgomery, 2019). Om en datapost x_n (eller det löpande genomsnittet x_n) ligger över UCL eller under LCL, har en statistiskt signifikant förändring sannolikt inträffat. Bedömning av datakvaliteten görs alltså genom att jämföra data med kontrollgränserna, och om kontrollgränserna överskrids tyder det på avvikande data. Figur 6.30 visar ett exempel på hur ett Shewharts styrdiagram kan användas på syrehaltsmätningar.

Shewharts styrdiagram är starkt beroende av de kontrollgränser som används. De antas vanligtvis tillhöra en normalfördelade, stationär fördelning med oberoende data. Detta är ett restriktivt antagande som kan lättas upp för att även vara tillämpligt för processer som inte producerar normalfördelade data. Sådana metoder använder istället icke-parametriska statistiska tekniker som antar någon specifik sannolikhetsfördelning och istället ger empiriska kontrollgränser (Box *et al.*, 2015). Sådana metoder är särskilt användbara för reningsverkstillämpningar som har autokorrelerade tidsserier som inte är normalfördelade.

6.7.2.3.1.2 Styrdiagram med kumulativ summa

Ett styrdiagram med kumulativ summa (CUSUM) är en modifiering av Shewhart-diagrammet. Denna teknik utvecklades på 1950-talet och användes för att upptäcka små ihållande processförändringar. Denna och andra modifieringar av Shewhart-diagrammet är främst av akademiskt intresse och de förekommer sällan i praktisk tillämpningar. Modifieringar av Shewharts grundläggande styrdiagram som baseras på olika *run-rules* (t.ex. Western Electrics) och datafiltreringstekniker används oftare. Som ett typiskt läroboksexempel ges här en översikt över CUSUM-metodiken.

Ett CUSUM-diagram visar den kumulativa summan av tidigare och nuvarande avvikelser för dataposterna från det medelvärdet (μ_0) enligt $C_t = \sum_{n=1}^{t} (x_n - \mu_0)$.CUSUM-diagrammet kan beskrivas



Figur 6.31 Exempel på ett CUSUM-styrdiagram för syrehalten på ett reningsverk som visar C_t^+ i ljusblått och C_t^- i mörkblått. Parametrarna för diagrammet är: $K = 0{,}01$ och $h = 5{,}0$. Parametervalet gjordes för att skapa en tydlig illustration. I praktiken väljs parametervärdena för att balansera feldetektionsförmågan.

på flera sätt där det vanligaste är med ett s.k. tabellformat (som dock främst väljs av pedagogiska skäl och inte det sätt metoden implementeras i praktiken). CUSUM baseras på två värden, C_t^+ och C_t^-, som är summan av avvikelser över, respektive under, medelvärdet μ_0. De kallas för övre och nedre CUSUM:

$$C_t^+ = \max\left\{0, x_t - \left(\mu_0 + K\right) + C_{t-1}^+\right\} \tag{6.2}$$

$$C_t^- = \max\left\{0, \left(\mu_0 - K\right) - x_t + C_{t-1}^-\right\} \tag{6.3}$$

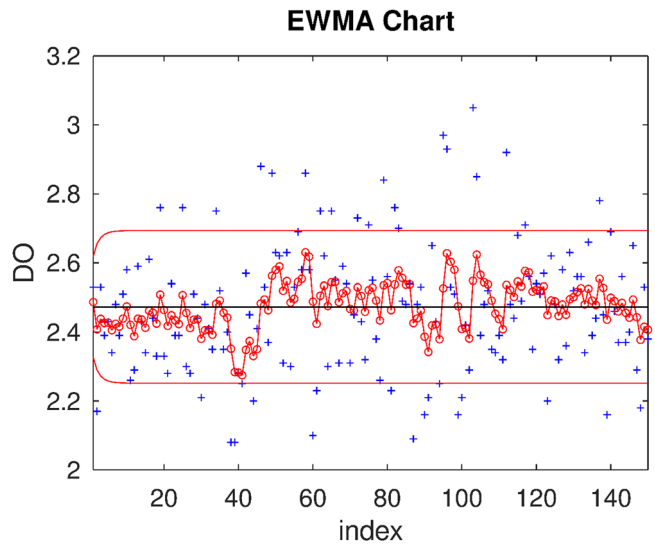
där initialvärdena sätts till noll $C_t^- = C_t^+ = 0$. Konstanten K kallas för referensvärde och väljs ofta som $K = x_1 - \mu_0 /2$, där x_1 representerar ett avvikande mätvärde man vill detektera (Ferrer-Riquelme, 2009). CUSUM-styrdiagrammet är effektivt för att upptäcka bestående förändringar, och i synnerhet när data skiftar önskat medelvärde μ_0 till ett avvikande medelvärde, x_1.

Båda statistiska måtten C_t^+ och C_t^- visualiseras i styrdiagrammet tillsammans med tröskelvärdet. Ett typiskt tröskelvärde är fem gånger standardavvikelsen för processens normala brus. Det är viktigt att välja bra tröskelvärde och K-värde för att CUSUM-metoden ska fungera korrekt.

Figur 6.31 visar ett exempel på hur ett CUSUM-styrdiagram kan tillämpas på syrehaltsmätningar från ett reningsverk.

6.7.2.3.1.3 Styrdiagram för exponentiellt viktat glidande medelvärde

Styrdiagrammet EWMA (*Exponential weighted moving averge*) är en annan modifiering av Shewhart-diagrammet, som togs fram på 1950-talet (Roberts, 1959). Fördelen med EWMA är att kontrollgränserna, UCL och LCL, uppdateras rekursivt och därmed tillåter vissa avvikelser från en stationär signal, men samtidigt bibehåller möjligheten att detektera små ihållande processförändringar. I diagrammet visualiseras de senaste dataposterna genom att multiplicera den senaste mätningen med en s.k. glömskefaktor λ (typiskt används $0{,}05 \leq \lambda \leq 0{,}25$) och det glidande medelvärdet EWMA multipliceras med $1 - \lambda$, det vill säga $\mathrm{EWMA}_i = \lambda x_i + (1 - \lambda)\mathrm{EWMA}_{i-1}$. Genom att ge data en exponentiellt avtagande vikt, utifrån sitt avstånd från nuvarande mätning, så fås



Figur 6.32 Exempel på EWMA-diagram för syrehaltsmätningar (blå kryss) med kontrollgränserna UCL och LCL i rött och CL i svart. EWMA-modellens värden är i rött med cirklar med parametervärden: $L = 3$ och $\lambda = 0{,}25$ och 'subgroup sample size one' eller fönsterlängden ett.

den önskade effekten att de senaste värdena ger större påverkan på beräkningen av det glidande medelvärdet än de äldsta värdena. Om parametern λ väljs till ett litet värde så fungerar EWMA som ett CUSUM-diagram. Om λ istället är nära noll så fungerar EWMA mer som ett Shewhart-styrdiagram.

Kontrollgränserna för EWMA-metoden motsvarar det medelvärdet (μ_0) och kontrollgränserna; UCL och LCL beräknas med följande uttryck som diskuteras av Montgomery (2019):

$$\mathrm{UCL} = \mu_0 + L\sigma\sqrt{\frac{\lambda}{2-\lambda}\left[1-(1-\lambda)^{2i}\right]} \tag{6.4}$$

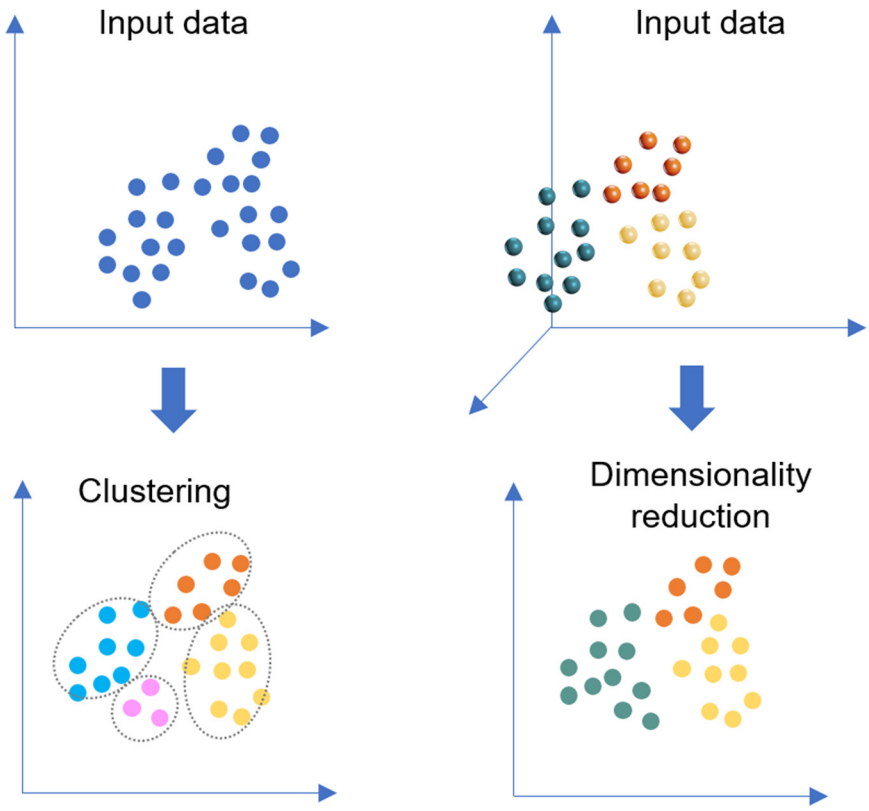
$$\mathrm{LCL} = \mu_0 - L\sigma\sqrt{\frac{\lambda}{2-\lambda}\left[1-(1-\lambda)^{2i}\right]} \tag{6.5}$$

Här är faktorn L bredden på kontrollgränserna, vars rekommenderade värden ligger runt 3. Figur 6.32 visar ett exempel på tillämpning av EWMA-styrdiagram på syrehaltsmätningar på ett reningsverk.

Den här typen av beräkningar används ofta som delkomponent i ett system för feldetektering eftersom metoden även kan användas för att lågpassfiltrera data och för att prediktera nästa värde för processer med icke-stationära medelvärden. Praktiska tillämpningar av dessa metoder inkluderar träning av diagrammodellen för att bestämma λ-parametern från ett träningsdataset, eller för att ställa in/konfigurera den via användarval. Valet av λ-parametern avgör i praktiken om diagrammet ska fungera som ett lågpassfilter eller som en feldetektor med viss förmåga att identifiera medelvärdesförändringar av olika magnituder.

6.7.2.3.2 Multivariata övervakningsmetoder

Multivariata övervakningsmetoder är en utvidgning av de univariata metoderna (Avsnitt 6.7.2.3.1) som istället för en variabel, nyttjar flera variabler samtidigt. Den viktigaste praktiska aspekten vid



Figur 6.33 Exempel på oövervakade multivariata metoder.

användning av multivariata metoder är möjligheten att dra nytta av de korrelationer som uppstår mellan de olika variablerna. Multivariata metoder används t.ex. för onlineövervakning av industriella processer (Kourti & MacGregor, 1996; Kourti *et al.*, 1996; Miletic *et al.*, 2008; Qin, 2012).

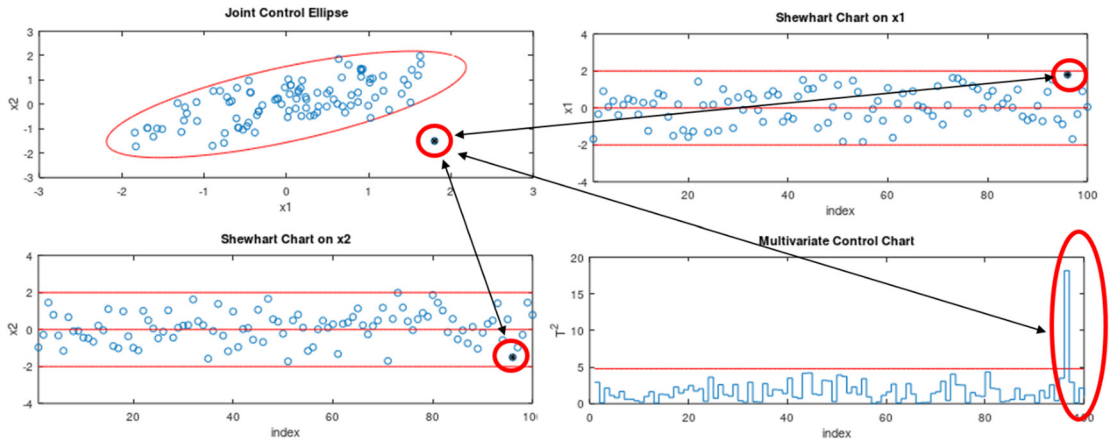
Avancerade metoder för *feature extraction* kan reducera antalet dimensioner av det ursprungliga multivariata rummet till ett lägre ordningens *feaure space* (t.ex. PCA), eller gruppera mätdata såsom illustreras i Figur 6.33.

6.7.2.3.2.1 Multivariata styrdiagram

Multivariata metoder kan användas för att generera styrdiagram, motsvarande de som beskrivits för univariata signaler. Som tidigare beskrevs så visualiseras tidserien för enskilda variabler tillsammans med kontrollgränser UCL och LCL för att detektera statistiskt signifikanta avvikelser (Oakland, 2003).

En begränsning med denna metod är att interaktioner mellan olika process- och produktionsvariabler försummas. Sådana interaktioner kan kvantifieras genom att beräkna korrelationer mellan variabler. När dessa korrelationer ignoreras så förloras information om den övervakade processen, vilket därmed begränsar övervakningssystemets förmåga att upptäcka avvikelser.

I praktiken så kan man förvänta sig att givardata inom olika reningssteg är korrelerade med varandra då de beskriver samma underliggande process. Omfattningen av dessa interaktioner avgör om univariata eller multivariata är mest framgångsrika för att detektera avvikelser.



Figur 6.34 Exempel på skillnaden mellan univariata och multivariata styrdiagram (*källa*: Miletic *et al., 2019b*).

De multivariata diagrammen kan konstrueras med hjälp av s.k. Hotellings T^2-värde, vilket är den multivariata motsvarigheten till den univariata välkända Student's *t-statistiken*. T^2-värdet (6.6) används för att mäta korrelationerna mellan processvariablerna i den $M \times M$ stora kovariansmatrisen $\mathbf{S}_x$ (se Mitra, 1998):

$$T^2(\mathrm{x}_n) = (\mathrm{x}_n - \overline{\mathrm{x}})^T \mathrm{S}_x^{-1} (\mathrm{x}_n - \overline{\mathrm{x}}) \tag{6.6}$$

För varje datapost $\mathbf{x}_n$ ger (6.6) det skalära värdet T^2 ($\mathbf{x}_n$), som mäter det kvadratiska avståndet mellan $\mathbf{x}_n$ och det förväntade värdet $\overline{\mathbf{x}}$. Detta avstånd viktas av den inverterade kovariansmatrisen och kvantifierar 'hur lik' $\mathbf{x}_n$ är de normala förhållandena $\bar{\boldsymbol{x}}$. Eftersom detta statistiska värde kan beräknas för varje observation av vektorn $\mathbf{x}_n$ kan T^2-värdena direkt visualiseras i en tidsserie, där även statistiska kontrollgränser anges utifrån empiriska percentilvärden. På så sätt kan T^2-diagrammet användas för att detektera avvikelser från den typiska normalvariationen i multivariat process.

Det finns flera reningsverkstillämpningar där T^2-värden använts för övervakning och optimering av SBR-processer (Villez *et al.*, 2010). Dessutom används T^2-värden ofta för oövervakade metoder, i synnerhet PCA, för processövervakning (se t.ex. Rosén, 2001).

Ett exempel på styrkan med multivariata styrdiagram ges i Figur 6.34 där först univariata Shewhart-diagram för två processvariabler (x_1 och x_2) visas. Inget av de Shewhart-diagrammen indikerar någon avvikelse. Men eftersom de två variablerna är starkt korrelerade så går det att detektera en avvikelse med det multivariata styrdiagrammet.

6.7.2.3.2.2 Principalkomponentanalys

Principalkomponentanalys (PCA) är den multivariata metod som haft störst genomslag inom reningsverkstillämpningar, se till exempel i Aguado *et al.* (2006), Haimi *et al.* (2013) och Newhart *et al.* (2019) och har nått en mognadsgrad som gör den praktiskt tillämplig (Corominas *et al.*, 2018). Mer specifikt så PCA använts för processövervakning, feldetektion och diagnostik av givare, visualisering av variationer i data, analys av avloppsvattnets sammansättning och även som en teknik och delkomponent vid förbehandling av data inför olika sorters modellering. PCA syftar till att reducera antalet dimensioner i det multivariata ursprungliga datasetet. Detta görs genom att projicera ursprungliga data på en mindre uppsättning (latenta) variabler. En PCA-modell byggs upp

från observationerna $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$, och den $N \times M$ stora datamatrisen, där var och en av de N raderna $\mathbf{X}_n = (x_{n,1}, x_{n,2}, \ldots, x_{n,m}, \ldots, x_{n,M})$ innehåller värden för en dataposten n med M processvariabler (Jackson, 1980).

Processvariablerna kan antingen vara givardata eller laboratoriedata.

PCA delar sedan upp datamatrisen $\mathbf{X}$ i två separata matriser $\mathbf{T}$ och $\mathbf{P}$, samt en residualmatris enligt

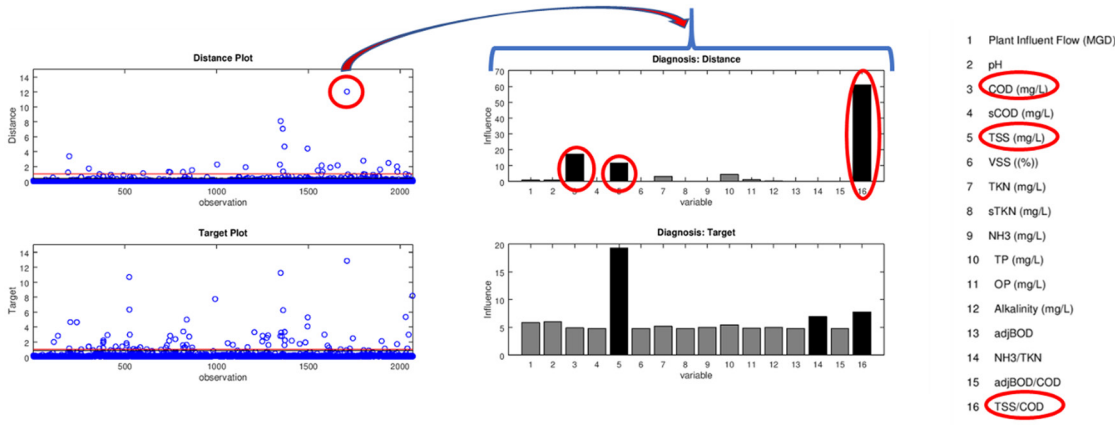
$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{6.7}$$

där $\mathbf{T}$ är en $N \times A$ matris som innehåller s.k. 'scores' och $\mathbf{P}$ är en $M \times A$ matris som innehåller s.k. 'loadings' där antalet principalkomponenter, A, väljs av användaren och $A \ll M$. Residualmatrisen $\mathbf{E}$ har storleken $N \times M$ och innehåller den variation som PCA-modellen inte kan förklara. Det finns flera metoder för att välja antalet principalkomponenter såsom korsvalidering och heuristiska tekniker (se Camacho & Ferrer, 2014; Jackson, 1993; Wold, 1978).

Kolumnerna i T-matrisen representerar oberoende variabler som ofta är relaterade till de kemiska, fysikaliska och biologiska processerna. Ur ett matematiskt perspektiv är dessa latenta variabler linjärkombinationer av de ursprungliga korrelerade processvariablerna. Dessa linjärkombinationer genererar skalära värden, s.k. *t-scores*, som lagras i raderna i T-matrisen. Varje kolumn i T-matrisen motsvarar en linjärkombination av variablerna. Vikterna för dessa kombinationer lagras istället i P-matrisen, som relaterar värdena i X-matrisen till T-matrisen. Denna matris projicerar variablerna från det ursprungliga 'X-rummet' till det mycket mindre 'P-rummet'. PCA skapar alltså en mappning av de ursprungliga processdata till de underliggande kemiska, fysikaliska och biologiska drivkrafterna.

Två statistiska mått används i PCA för att upptäcka processfel och för att mäta modellens lämplighet och datakvalitet. Vid sidan av Hotellings T^2-värde som har redan introducerats så används även det kvadrerade prediktionsfelet (*squared prediciton error – SPE*) som ett skalärt mått på PCA-modellens kvalitet. SPE beräknas med hjälp av matrismetoder för att ge ett enda värde som kan användas i ett styrdiagram. SPE kan i sin tur brytas ned för att undersöka vilka processvariabler som bidrar mest till en avvikelse i datasetet, vilket brukar visualiseras i en s.k. *contribution plot* (Westerhuis *et al.*, 2000). Dessa feldetektionstekniker är även användbara för automatiserad feldiagnostik där processdata analyseras vid flertalet tillfällen (Miletic *et al.*, 2004, 2019a).

Figur 6.35 illustrerar hur SPE- och T^2-måtten kan användas för avvikelsedetektion och hur contribution plots kan användas för efterföljande diagnostik. Exemplet baseras på verkliga driftdata



Figur 6.35 Exempel på detektion och diagnos av avvikelser med contribution plots (*källa*: inCTRL Solutions, 2018). Notera att '*Distance Plot*' avser SPE och '*Target Plot*' *avser* T^2.

från ett reningsverk i USA där COD, TSS och kvoten TSS per COD identifierades som de mest sannolika rotorsakerna till avvikelserna (inCTRL Solutions, 2018). Datakvaliteten i ett dataset kan alltså analyseras utifrån SPE- och T^2-måtten om de visualiseras i ett styrdiagram med kontrollgränser. Om dessa kontrollgränser överskrids indikerar det avvikande data.

Det är viktigt att notera att reningsprocesserna är dynamiska och icke-linjära processer av naturen. På grund av detta så har modifierade PCA-varianter föreslagits för att göra PCA mer adaptiv och för att även kunna hantera icke-linjära processvariationer, se t.ex. Haimi *et al.* (2016); Rosén (2001). Dessa modifieringar används då som plan B om standardversion av PCA inte fungerar tillfredsställande, se Avsnitt 6.7.2.3.3.3 på hur olinjäriteter kan hanteras.

PCA kan även tillämpas på data från batchprocesser såsom SBR-processer (Aguado *et al.*, 2007) med s.k. *Multiway-PCA*. Denna modifiering kräver dock att indata (X-matrisen) förbehandlas på ett särskilt sätt där bacthnumret och en extra dimension för tid eller reaktionsgrad inkluderas i datamatrisen. Dessutom behöver P-matrisen i (6.7) justeras med en enkel beräkning som kallas för *unfolding* (Nomikos & MacGregor, 1995).

6.7.2.3.2.3 Gruppering av data (clustering)

Klustringstekniker syftar till att gruppera data i informativa undergrupper eller kluster. Varje kluster består av data som liknar varandra men skiljer sig från data i andra kluster. Till skillnad från klassificeringstekniker (Avsnitt 6.7.2.2.1) som grupperar data i klasser baserade på annoterade data, är klustring en oövervakad metodik för att undersöka mönster i en datamängd utan annoteringar. Insikter om datakvalitet baseras således genom att definiera varje kluster och analysera om data ligger utanför eller inom konfidensområdet för respektive kluster. Det finns flera olika metoder för att identifiera kluster, som beskrivs i detalj i statistiska läroböcker, t.ex. Hastie *et al.* (2009).

På reningsverk används klustring främst för processövervakning och för att visualisera och diagnostisering av processens driftförhållanden. De mest populära klustringsteknikerna som tillämpats på reningsverk är *K-means-algorithm* och regelbaserade algoritmer (*fuzzy clustering*) på fuzzy-klustringsalgoritmer (Corominas *et al.*, 2018; Haimi *et al.*, 2013). Klustring kombineras dessutom ofta med en dimensionsreduktion med PCA (se t.ex. Aguado & Rosén, 2008; Maere *et al.*, 2012). Klustring av de scores som en PCA-modell genererar är ett verktyg för automatiserad feldetektion och processdiagnostik. Figur 6.36 ger ett exempel på hur klustring användes i en fallstudie på ett europeiskt reningsverk (inCTRL Solutions, 2020).

I exemplet så grupperas scores i t_1 och t_2-planet i två kluster, vart och ett som kan kopplas till ett driftfall, där ellipserna indikerar icke-parametriska konfidensintervall kring klustren.

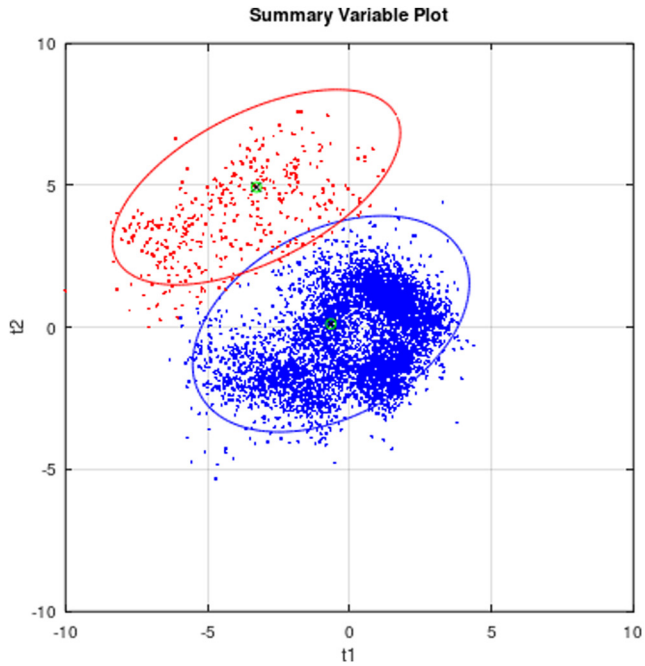
Förutom PCA så har klustringsmetoder kombinerats med oövervakade ANN-metoder som kallas för *self-organizing maps – SOM)* (Kohonen, 2001). SOM och klustring har visat sig vara användbar för processövervakning (García & González, 2004).

6.7.2.3.3 Praktiska utmaningar vid tillämpning

Dagens moderna automationssystem består av effektiva enheter som kan mäta, utföra beräkningar och lagra stora mängder processdata med mycket snabba samplingstider. Som en konsekvens av denna teknikutveckling med möjlighet att registreras snabba processförändringar så kommer även förändringar i korrelations- och autokorrelationensstrukturen att återspeglas i mätningarna, en egenskap som kallas icke-stationäritet.

I detta avsnitt beskrivs ett antal procedurer som kan användas för att modifiera standardtekniker till att kunna hantera autokorrelerade och icke-stationära processer. Några statistiska metoder kommer även beskrivas kan hantera outliers, data som inte är normalfördelade och en metod för att hantera olinjära data.

På samma sätt som tidigare förutsätts en $N \times M$ datamatris, där var och en av de N raderna $x_n = (x_{n,1}, x_{n,2}, \ldots, x_{n,m}, \ldots, x_{n,M})$ innehåller värdena för M processvariabler och n dataposter.



Figur 6.36 Exempel på användning av klustring för att gruppera processdata som förbehandlats med PCA (*källa*: inCTRL Solutions, 2020).

PCA syftar till att beskriva kovariansstrukturen i data med ett begränsat antal principalkomponenter. Dessa principalkomponenter definieras som linjärkombinationer av originalvariablerna och behöver vara parvis ortogonal. De underliggande antagandena för PCA faller ofta på grund av autokorrelation och icke-stationäritet, och avvikande värden kan påverka uppskattningen av modellparametrarna (dvs. medelvärdet och kovariansmatrisen för den antagna normalfördelningen). Nästa stycke beskriver hur både multivariata och liknanden univariata metoder påverkas av outliers.

6.7.2.3.3.1 Känslighet för avvikande värden

Driftdata innehåller vanligtvis dataposter som avviker från huvuddelen av datasetet. Dessa avvikande värden kallas även *outliers* och kan bero på mätfel alternativt orsakas av processtörningar. Oavsett orsak så påverkar outliers den statiska analysen genom att felaktigt snedvrida variationen i data. Detta kan observeras i förändrade medelvärden och eller variablernas kovarians, vilket därmed minskar tillförlitligheten i de statistiska metoder som nyttjar dessa statistiska mått. Vanligt oövervakade metoder såsom PCA- och *k-means-algorithm*, linjära regressionsmodeller och klassificeringsmetoder är därför känsliga för extremvärden och outliers.

Robust statistik är ett begrepp som används för de metoder som anpassats för att begränsa inverkan av outliers på modellen. Detta stycke beskriver några av dessa robusta metoder.

I det univariata fallet är det vanligaste robusta estimat av medelvärdet *medianen* respektive *medianen av alla absoluta avvikelser från medianen* som robust alternativ till standardavvikelsen. I det multivariata fallet baseras vanliga robusta skattningar på MCD-kriteriet (*Minimum Covariance Determinant*). I MCD identifieras den delmängd dataposter vars kovariansmatris har den minsta determinanten, och därmed minsta spridning, som sedan används för att beräkna den empiriska medelvärdesvektorn och kovariansmatrisen. Det är inte trivialt att beräkna MCD-estimatorer

eftersom sökningarna i delmängderna är kombinatoriska. Det finns även andra alternativ och robusta skattningsmetoder (Maronna *et al.*, 2019).

PCA baseras på klassiska skattningar av kovarians och medelvärde vilket gör PCA sårbar för extremvärden. Det finns ett antal varianter av robust PCA där de enklaste ersätter de klassiska skattningarna med robust motsvarigheter. Det finns även mer sofistikerade alternativ som baseras på *projection pursuit* (Hubert *et al.*, 2005) och *sparse matrices* (Candès *et al.*, 2011). Outliers skapar liknande problem hos klustringsmetoder såsom *k-means*. Detta beror på att *k-means* syftar till att gruppera data genom att minimera summan av det kvadrerade avstånden mellan dataposter och centrum för ett, av användaren valt, antal k kluster. Eftersom mittpunkten för de olika klustren uppskattas från data är denna teknik också känslig för extremvärden. Modifierade k-means-metoder syftar till att göra denna metod för klustring mer robust (Gallegos & Ritter, 2005). Corona *et al.* (2013) använde robust PCA för en softsensor på ett reningsverk, men bortsett från det exemplet finns få exempel på reningsverkstillämpningar med robusta klustringsmetoder.

6.7.2.3.3.2 Autokorrelation och icke-stationäritet

Dynamisk PCA (Ku *et al.*, 1995) är en modifierad version av PCA för att hantera autokorrelerade data. I metoden så adderas L-laggade dataposter på samma sätt som en autoregressiv process kan modelleras. Dynamisk PCA innebär i praktiken att man tillämpar PCA på den till den utökade datamatrisen $\widetilde{\mathbf{X}_L} = \left[\mathbf{X}_n \cdots \mathbf{X}_{n-l} \cdots \mathbf{X}_{n-L}\right]$ med $\mathbf{X}_{n-l}$med L tidsförskjutna värden, där L är ett användarval.

Dynamisk PCA har tillämpats i flertalet fall för feldetektion och processövervakning i både simulering och verkliga fallstudier (se Yoo *et al.*, 2004 och Villez *et al.*, 2005). Pågående forskning fokuserar på att anpassa de statistiska måtten T^2 och SPE till dynamisk PCA (Ferrer, 2014).

Eftersom antalet tidsförskjutningar L är konstant och inte varieras, är dynamisk PCA inte adaptiv och därför inte lämplig för data från icke-stationära processer. I icke-stationära tillämpningar behöver man därför modifiera dessa tekniker med mekanismer för att automatiskt justera parametrarna. Tekniker som rekursiv PCA (Li *et al.*, 2000) och *moving window PCA* (MWPCA, av Wang *et al.*, 2005) bygger på tanken att nya dataposter bör inkluderas i modellen allteftersom, medan gamla dataposter bör ges allt mindre relevans. I MWPCA drivs denna intuition till sin spets. Vid varje tidpunkt tas ett fast antal äldre dataposter bort från modellen medan ett lika stort antal läggs till, och modellens parametrar beräknas sedan på nytt utifrån ett glidande fönster. Fönsterstorleken och antalet borttagna dataposter kan vara fasta eller justeras adaptivt. Även här är tillämpningar inom felanalys vanliga och baseras på olika givardata. Exempelvis så använde Haimi *et al* (2016) MWPCA med en adaptiv fönsterlängd för feldetektion i en reningsverksapplikation, medan Awhangbo *et al* (2020) använde MWPCA med fasta fönsterlängd för övervakning av en rötkammare med utifrån IR-mätningar. Gernaey *et al* (2004) jämförde feldetekteringsförmågan hos rekursiv och traditionell PCA i simulerade och fullskaliga tillämpningar, och Rosen och Lennox (2001) studerade liknande problem med rekursiv och *multiscale* PCA.

6.7.2.3.3.3 Olinjära processer

Tekniker som *kernel PCA* (KPCA; Schölkopf *et al.*, 1998) ger en lösning på det faktum att PCA inte kan hantera olinjära data. KPCA diagonaliserar en positiv semidefinit kovariansfunktion som är begränsad till dataposterna i $\mathbf{X}$. Under antagandet att det finns en funktion ϕ som mappar dataposterna till ett godtyckligt högdimensionellt variabelrum (med nya egenskaper) genomför KPCA en PCA i det utökade variabelrummet. Explicit kunskap om ϕ är dock inte nödvändig eftersom alla beräkningar kan utföras direkt från den positiva semidefinita kovariansfunktionen $k(\mathbf{x}_n, \mathbf{x}_{n'j})$. Ett vanligt val av kovariansfunktion är 'Gaussiska bump/squared exponential kernel' $k\left(\mathbf{x}_n, \mathbf{x}_{n'}\right) = -exp\left(-\| \mathbf{x}_n - \mathbf{x}_{n'} \|^2 / \sigma^2\right)$, där standardformen för PCA fås om kovariansfunktionen väljs som $k\left(\mathbf{x}_n, \mathbf{x}_{n'}\right) = \mathbf{x}_n^T \mathbf{x}_{n'}$. Tillägget med en kovariansfunktion för robust PCA (kernel ROBPCA, Debruyne & Verdonck, 2010) används när

variablerna antas ha ett olinjärt samband. Kazor *et al.* (2016) jämför statiska, dynamiska, adaptiva och adaptiva-dynamiska versioner av PCA, KPCA och andra tekniker för 'manifold learning', och Klanderman *et al.* (2020) utvidgar detta till en plattform för feldiagnostik.

6.7.3 Riktlinjer för val av metod

Detta avsnitt ger allmän råd och ett ramverk för att välja och tillämpa av feldetektionsmetoder för att analysera datakvalitet, med utgångspunkt i de metoder som presenterades i Avsnitt 6.7.1–6.7.2.

6.7.3.1 Ramverk för att analysera datakvalitet med datadrivna metoder

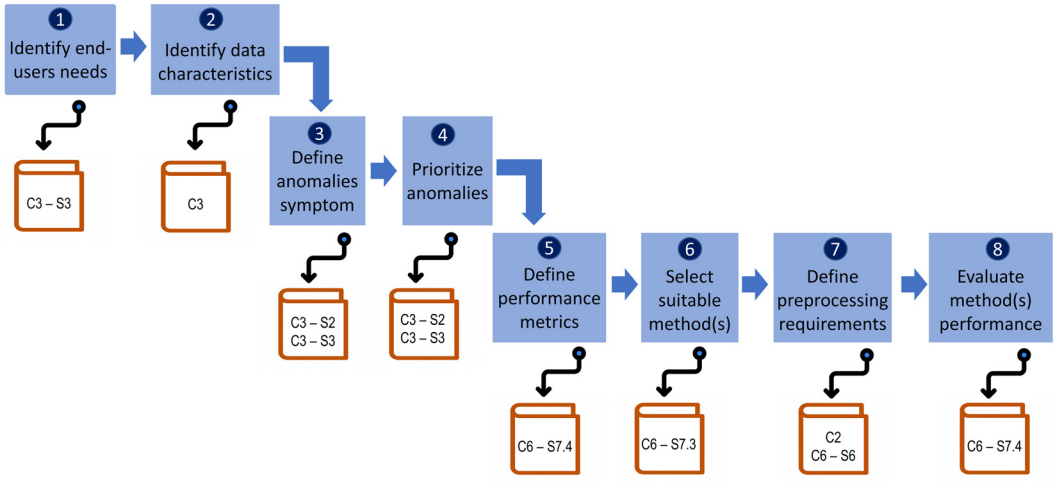
Ett ramverk för tillförlitlig och noggrann analys av datakvaliteten kräver flera olika steg. En god start är att alltid involvera domänexperter och slutanvändare av data, framför allt när ramverket byggs upp och när metoderna för att analysera datakvaliteten utvärderas (utifrån riktlinjer som beskrivs i Avsnitt 6.7.4). Val och tillämpning av metod bör inkludera stegen i Figur 6.37, där även korsreferenser ges till de kapitel (C) och stycken (S) som berör motsvarande steg:

(1) Identifiera slutanvändarnas behov både på kort sikt (t.ex. realtidsvisualisering och annotering) och på lång sikt (t.ex. uppskattning av koncentrationer och belastningar). Detta kan ge insikt i vilken typ av fel eller avvikelser som är viktigast att upptäcka, och därmed vilka metoder som är lämpligast att använda. Se Avsnitt 3.3 i Kapitel 3 för orsaker till försämrad datakvalitet eller anomalier som kan vara relevanta.

(2) Analysera vad som är karakteristiskt med data. Påverkas data av tidsberoende fenomen såsom säsongsvariationer, väderrelaterade händelser (t.ex. nederbörd), icke-stationära fenomen (t.ex. drift) eller korrelation mellan de uppmätta variablerna. Detta inkluderar även historiska data och metadata i den mån de finns tillgängliga, t.ex. kommentarer som beskriver med datakvaliteten (se Kapitel 3).

(3) Definiera symptomen för de avvikelser som listas i steg 1. Se Avsnitt 3.1 och 3.3 i Kapitel 3 för att koppla samman intressanta avvikelser med de symptom som kan upptäckas i data.

(4) Prioritera vilken typ av avvikelser som är viktigast att upptäcka. Prioritera dem utifrån hur viktigt det är att avvikelserna upptäcks och utifrån konsekvenserna av att avvikelserna inte upptäcks. Insikter från detta steg kommer användas för att utvärdera deldetektionsmetodens prestanda.



Figur 6.37 Ramverk för val av datadrivna metoder för att analysera datakvalitet.

(5) Definiera ett eller flera prestandamått som ska användas för att utvärdera feldetektionsmetodens prestanda (Avsnitt 6.7.4). Definiera hur utfallet av dessa prestandamått ska användas för att bedöma datakvaliteten. Om det är tillämpligt så kan man i detta steg även beskriva hur data ska delas upp för modellträning, -validering och test för att säkerställa att det blir en oberoende utvärdering.
(6) Välj den lämpligaste metoden. Både egenskaper hos data och användarnas behov är viktiga faktorer för att avgöra vilken datadriven metod som är lämpligast att använda för att analysera datakvaliteten. I praktiken kommer olika algoritmer att upptäcka vissa typer av anomalier bättre än andra. En effekt av detta är att det i många fall är en kombination av flera olika metoder som ger det bästa resultatet.
(7) Definiera kraven på dataförbehandling utifrån strukturen på tillgängliga data (se Kapitel 2) och de valda algoritmerna. Detta inkluderar val av data, centrering, skalning med mera (se Avsnitt 6.6). Det är viktigt att vara konsekvent med förbehandlingsstegen för att jämförelsen mellan de olika metoderna ska bli rättvisande. Det är även viktigt att se till att säkerställa att den tänkta uppdelningen av data för modellträning, validering och testning bibehålls.
(8) Kalibrera/träna, validera och testa de valda metoderna på ett representativt dataset.

Metadata enligt beskrivningen i Kapitel 3 har en viktig roll i processen för att bedöma datakvaliteten. Alla modifieringar av rådata (enligt beskrivningen i Avsnitt 3.2.4) behöver dokumenteras noggrant. De metadata som tas fram om datakvaliteten (enligt beskrivningen i Avsnitt 3.3), behöver förstås också dokumenteras.

6.7.3.2 Val av metod(er)

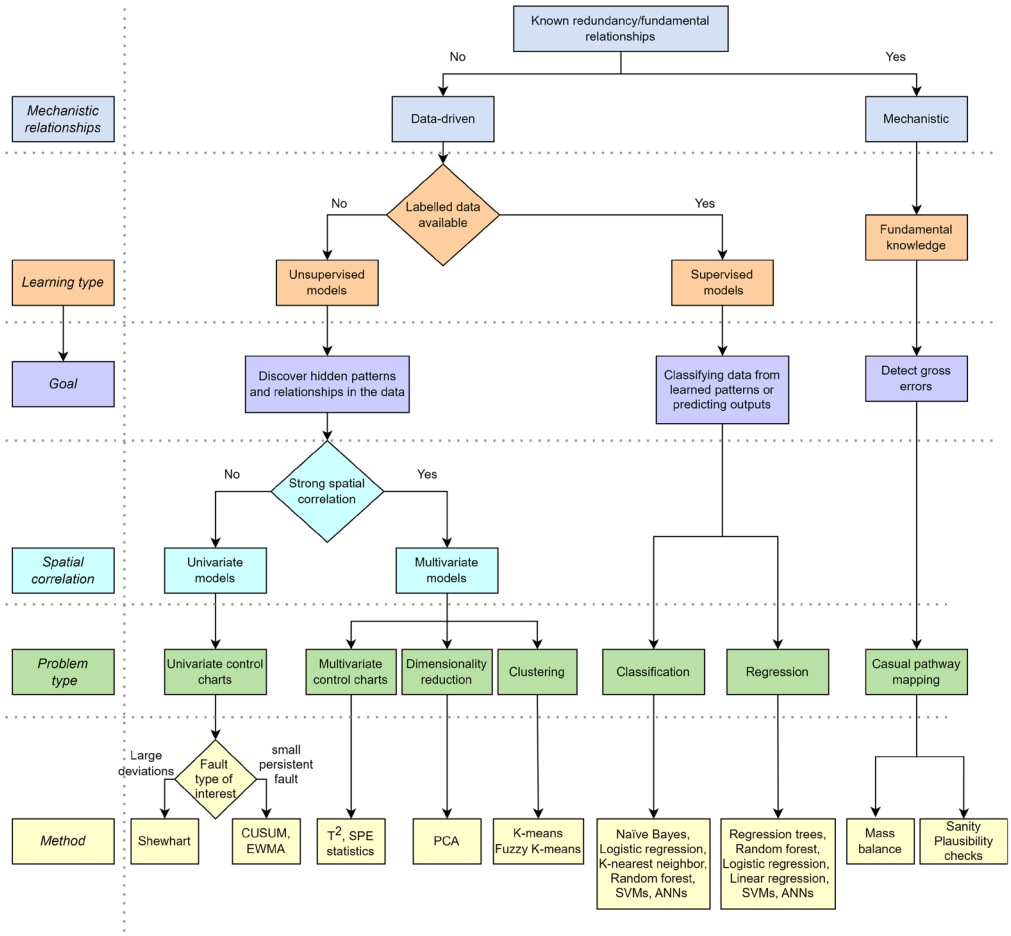
Idag väljs datadrivna analysmetoder till stor del *ad hoc* och utifrån *trial-and-error*, som förutom att vara en subjektiv strategi också kan vara tids- och kostnadskrävande. I idealfallet skulle beslutsprocessen vara mer systematisk för att göra ett bättre val av feldetektionsmetod.

Ett första vägskäl i Figur 6.38 är att förstå och bestämma om det finns mekanistisk kunskap om processen (känd redundans och grundläggande relationer) som kan användas. Sådana metoder beskrevs i Avsnitt 6.7.1. Om denna typ av kunskap eller information är okänd (kan vara för tidskrävande/resurskrävande att få fram data med försök) behöver informationen från historiska data istället utnyttjas genom att implementera datadrivna metoder som beskrevs i Avsnitt 6.7.2.

En bra rutin är att alltid genomföra rimlighetskontroller och använda massbalanser (när sådana finns tillgängliga) som en del av den grundläggande dataanalysen. Detta för att upptäcka tillfälliga mätfel i data innan den efterföljande analysen görs. De variabler och givare som ska analyseras behöver då ingå i massbalansen, som behöver vara komplett och en korrekt representation av det verkliga processchemat.

Det andra vägskälet vid val av datadriven metod handlar om främst om tillgång på data och därmed vilken typ av inlärning som är möjlig (övervakad – från annoterade dataset eller oövervakad – från omärkta dataset). Men valet påverkas även av önskat resultat från datakvalitetsvalideringen (*the goal*).

Tabell 6.7 ger en enkel översikt över krav på indata och utdata beroende på vald metod. Oavsett vilken feldetektionsmetod som används så kan utdata beskrivas med (a) en etikett som anger om en datapunkt klassas som avvikande eller ej, och (b) ett värde som anger graden av avvikelse (Markus & Uchida, 2016). Värdet kan även omvandlas till en etikett med hjälp av ett lämpligt tröskelvärde (Alferes *m.fl.*, 2013). I praktiken kommer det slutliga valet av metod för bedömning av datakvalitet att omfatta de punkter som diskuterats tidigare, inklusive de resurser som krävs för genomförande. Det kan även nämnas att oövervakade metoder ofta har föredragits eftersom det är komplext att ha ett komplett annoterat tränings och testdataset, vilket dessutom kräver mycket arbete och tekniska expertis.



Figur 6.38 Guide för att välja datadriven analysmetod av datakvalitet enligt beskrivningen i Avsnitten 6.7.1 och 6.7.2

6.7.3.3 Ytterligare praktiska anmärkningar

Oavsett vilken metod som väljs är det viktigt att ta hänsyn till att både mekanistiska och datadrivna metoder har sina egna styrkor och begränsningar (Tabell 6.8) och att de inte bör betraktas som motpoler. Aktuell forskning fokuserar faktiskt på att kompensera metodernas olika begränsningar och skapa synergier mellan dem (Bikmukhametov & Jaschke, 2020). Dessa s.k. hybridmetoder använder mekanistiska modeller för att: träna och utvärdera datadrivna metoder, minska antalet dimensioner, minska datamodellens komplexitet, minska beräkningstiden, komplettera data och minska den empiriska osäkerheten.

I praktiken är univariata styrdiagram (Avsnitt 6.7.2.3.1) de mest använda feldetektionsmetoderna av den enkla anledningen att de är enklare att implementera och använda. Alla typer av avvikelser kommer dock inte att upptäckas, särskilt inte när det gäller komplexa tidsserier. Tabell 6.9 belyser de viktigaste styrkorna och begränsningarna hos respektive metod. Ur ett praktiskt perspektiv rekommenderas att EWMA- eller CUSUM-diagram kombineras med Shewhart-diagram för att dra nytta av fördelarna med båda typerna av diagram, så att både små och stora förändringar i medelvärdet eller processens standardavvikelse (σ) kan upptäckas tillräckligt snabbt.

Tabell 6.7 Sammanfattning av krav på in- och utdata vid val av datadriven valideringsmetod för kvalitetsbedömning av data.

Metod				Indata	Nyckelparametrar	Utdata
Datadriven	Övervakad	Klassificering		Stora annoterade träningsdataset	Metodberoende, t.ex. antalet grannar (kNN)	Klassificering av datapunkter, detektion eller ej
		Regression			Metodberoende, t.ex. delningskriterium (reg. tree)	
	Oövervakade	Univariata	Styrdiagram	Stora oannotedae dataset	Styrgränser, medelvärde, standardavvikelse, fönsterlängd	Alarm vid överskriden kontrollgräns
		Multivariata	Styrdiagram		Kontrollgränser	Alarm vid överskriden kontrollgräns, contribution plots
			PCA		Antal principalkomponenter	
			Klustrings-tekniker		Metodberoende, t.ex. antalet tillåtna kluster (*k*-means)	Detektion av avvikande data som avviker från identifierade kluster
Mekanistisk	Univariata	Rimlighetskontroller		Driftdata, mindre dataset	Rimliga nyckeltal, förhållanden, intervall	Indikation på om data passerar rimlighetskontrollen
	Multivariata	Massbalansering		Driftdatat, mindre dataset	Variabler och stökiometriska samband (processmassbalanser)	Detektion av mätningar som avviker från förväntade massflöden

Tabell 6.8 Styrkor och begränsningar med datadrivna och mekanistiska metoder.

	Styrkor	Begränsningar
Datadriven	• Enklare konstruktion och uppställning. • Liten insats för att parametrisera modellen. • Används när det är svårt, dyrt eller omöjligt att beskriva systemet med fysikaliska ekvationer. • En empirisk metod kan ge tillförlitliga resultat när den baseras på en stor mängd testdata med tillräcklig variabilitet.	• Svårtolkad och otransparenta utdata. • Stora mängder historiska data behövs. • I de flesta fall är de endast tillämpliga på särskilda driftsförhållanden (liknande träningsdata). • I allmänhet ger den lägre noggrannhet och är ej att föredra för exakta prediktioner. Noggrannheten ökar dock med mängden data som tillhandahålls och • valet av en lämplig algoritm, och med annoterade in och utdata.
Mekanisk	• Högre noggrannhet. • Lägre antal datapunkter (observationer) krävs. • Extrapolering till olika driftsförhållanden är möjlig inom området den kalibrerats för. • Resultatet av metoden är transparent.	• Ofta är mekanistisk information inte tillgänglig eller kräver tidskrävande och komplexa beräkningar. • Alltför förenklade antaganden kan leda till felaktiga slutsatser. • Kräver experimentella data som kan vara svåra att få tag på. • Betydande investering i metod • utveckling, testning och tillämpning.

Tabell 6.9 Praktiska överväganden vid användning av styrdiagram.

	Styrkor	Begränsningar
Shewhart	• Enkelt och intuitivt. • Effektivt för att upptäcka stora förändringar i processparametrar. • Används ofta på grund av att de är relativt enkla att implementera och tolka.	• Det kan ta lång tid att upptäcka en liten bestående förändring.
CUSUM	• Effektivare när det gäller att upptäcka små förskjutningar i processmedelvärdet (2σ eller mindre). • Det är lätt att visuellt identifiera förändringarna i processens medelvärde. • Ger den snabbaste upptäckten av bestående avvikelser oavsett storlek på avvikelsen.	• Det är svårare att upprätta och underhålla CUSUM-diagram. • CUSUM-diagram är långsammare när det gäller att upptäcka stora processmedelvärdesförändringar. • Svårare att upptäcka och tolka mönster.
EWMA	• Tack vare den exponentiellt viktade faktorn är EWMA-diagrammet mycket känsligare för mindre förskjutningar i processens medelvärde (0,5 till 2σ). • Används också för att jämna ut effekten av brus i data. • Kan identifiera trender, vilket gör det möjligt att identifiera om en process förändrats positivt eller negativt. • Eftersom den är baserad på ett digitalt filter kan den användas för prognoser och reglering.	• EWMA-diagram är också långsammare när det gäller att upptäcka större förändringar. • Fördröjning. Om en uppåtgående förskjutning sker när EWMA befinner sig i den nedre delen av sitt intervall (LCL) måste den först komma tillbaka upp till mittlinjen innan den kan fortsätta upp till UCL. Detta fördröjer upptäckten av förskjutningen i medelvärdet. • Ingen explicit procedur för att hitta det lämpligaste värdet för glömskefaktorn λ.

Som framgår av Avsnitt 6.7.2 så krävs, oavsett metod, en korrekt uppdelning av data för att med en del av data träna och kalibrering av metoden, och sedan en validering metoden på ett annat dataset. Även om mekanistiska metoder kräver en djupare systemförståelse jämfört med empiriska metoder så krävs alltid processkunskap för att validera rimligheten i metodens utdata och därmed undvika missvisande resultat (se Avsnitt 6.7.4).

Genom att implementera flera metoder parallellt eller i serie så kan ett mer exakt och tillförlitligt resultat fås. Ett exempel är att använda enklare metoder (rimlighetskontroller och styrdiagram) för en grundläggande validering av datakvaliteten, följt av en mer avancerad analys. Univariata och multivariata metoder kan också kombineras för att upptäcka fel som påverkar enskilda givare och situationer som påverkar hela processen.

6.7.4 Prestandautvärdering av datadrivna feldetektionsmetoder

Detta avsnitt ger en enkel vägledning i att utvärdera prestandan för feldetektionsmetoder. Först beskrivs själva arbetsflödet för prestandautvärdering med råd om vart och ett av de olika stegen (Avsnitt 6.7.4.1). Därefter ges ett exempel som illustrerar hur fyra prestandan hos fyra olika feldetektionsmetoder kan jämföras utifrån verkliga driftsdata från ett reningsverk (Avsnitt 6.7.4.2). En bra prestandautvärdering och metodjämförelse är en förutsättning för att kunna välja rätt feldetektionsmetod. Avsnitten avslutas med en diskussion om fördelar samt typiska fallgropar som kan uppstå vid prestandautvärdering av datadrivna (Avsnitt 6.7.4.3).

6.7.4.1 Arbetsflöde

Det finns två syften med att bedöma prestandan hos en feldetektionsmetod. För det första ska bedömningen ge svar på om metodens prestanda uppfyller slutanvändarnas behov och praktiska krav. För det andra så kan man, utifrån en sådan bedömning, identifiera den metod som bäst kan identifiera avvikande mätvärden från normala data. De fem grundläggande stegen för en detaljerad prestandabedömning innefattar att:

(1) Definiera målet med feldetektionsproblemet och kvantifiera vilken prestanda som krävs för att datakvalitetsbedömning ska anses tillräckligt noggrann.
(2) Välja ett eller flera prestandamått som återspeglar målet i steg 1.
(3) Välja ut eller producera de data som ska användas för att utvärdera feldetektionsmetoderna.
(4) Jämför feldetektionsmetoderna utifrån de valda prestandamåtten i steg 2 och data i steg 3.
(5) Analysera resultaten utifrån syftet i steg 1) och det sammanhang de ska användas i praktiken.

6.7.4.1.1 Steg 1: Definiera och kvantifiera önskad prestanda

Det första steget är att definiera den lägsta acceptabla detektionsprestandan, som fortfarande uppfyller slutanvändarens behov. Att upptäcka subtila avvikelser, t.ex. en mindre sensordrift, med höga krav på att inga falsklarm ska förekomma kommer att vara svårare och sannolikt dyrare att utveckla, jämfört med metoder som ger grova indikationer och stöd, t.ex. för att prioritera sensorunderhåll. En tidig definition av omfattningen av anomalidetekteringen rekommenderas eftersom det kan avslöja omöjliga uppgifter tidigt, istället för efter slutförandet av en omfattande utvärderingsstudie.

6.7.4.1.2 Steg 2: Välj prestandamått

Prestandautvärderingsmått kan delas in i två kategorier: dynamiska och statiska (Kurtoglu *et al.*, 2008). Statiska mått kvantifierar prestandan hos en feldetektionsmetod utifrån data utan hänsyn till deras inbördes ordning i tiden. Dynamiska (och hybrida) prestandamått inkluderar däremot den kronologiska ordningen i data. Tabell 6.10 listar de viktigaste prestandamåtten (statiska och dynamiska/hybrid) för att analysera feldetektionsmetoder i reningsverkstillämpningar.

Tabell 6.10 Vanliga prestandamått för att bedöma feldetektionsmetoder i reningsverkstillämpningar.

Klassificering	Typ av prestandamått	Prestandamått	Referenser
Statiska	Prestandamått som baseras på confusion matrix	True positive rate/hit rate/sensitivity/recall False positive rate/false alarm rate False negative rate/miss rate True negative rate/specificity Negative predictive value Precision/Positive predictive value Accuracy F-score/F-measure	Fawcett (2006); Frank *et al.* (2019); Schneider *et al.* (2019)
	Probabilistiska*	Receiver Operating Characteristics (ROC) graphs Area Under the Curve (AUC)/Area under the ROC curve Partial area under curve Kappa Coefficient/Cohen's Kappa/Kappa statistics Matthews Correlation Coefficient (MCC)/Phi coefficient	Fawcett (2006); Frank *et al.* (2019); Samuelsson *et al.* (2019)
	Regressionsbaserade	Root Mean Square Error (RMSE) R2 score (coefficient of determination) Mean and Mean Absolute Errors (ME and MAE) Mean Square Error (MSE) Mean Square Derivative Error (MSDE) Percent Bias (PBIAS) Mean Absolute Error Relative (MEAR) Sum of residuals	Hauduc *et al.* (2015)
Dynamiska/hybrida	Dynamiska	Mean time to detection Detection stability factor	Kurtoglu *et al.* (2008); Fellini *et al.* (2019)
		Detection speed (ARL1)	Baklouti *et al.* (2021)
	Hybrida	Weighting: Anomaly Detection Evaluation Indexes Anomaly detection index (IFD)	Corominas *et al.* (2011); Alferes *et al.* (2017); Carlsson and Zambrano (2016)

I de fall en och samma metod har flera olika namn separeras de med '/'.
*Probabilistiska prestandamått som även baseras på mätetalen i confusion matrix.

6.7.4.1.2.1 Statiska mått

Som Tabell 6.10 visar så finns det ett stort antal statiska prestandamått i litteraturen. De mest populära måtten baseras på den s.k. *confusion matrix* som innehåller fyra mätetal:

- Sann positiv (*true positive – tp*) eller korrekt identifierade avvikelser
- Falsk positiv (*false positive – fp*) eller falsklarm
- Falsk negativ (*false negative – fn*) eller en oupptäckt avvikelse
- Sann negativ (*true negative – tn*) eller korrekt klassning av normala felfria data.

Utifrån dessa grundläggande mätetal av klassificeringarnas prestanda så kan prestandamåtten *accuracy*, *precision*, *recall* och *F1-score* beräknas. Accuracy är förhållandet mellan korrekt identifierade dataposter och det totala antalet dataposter:

$$\text{Accuracy} = (tp + tn)/(tp + tn + fp + fn)$$

Precision är den andel av de faktiska avvikelserna som upptäcks korrekt med metoden:

$$\text{Precision} = tp/\left(tp + fp\right)$$

Recall mäter metodens förmåga att identifiera hur stor andel av de faktiska avvikelserna som upptäcks:

$$\text{Recall} = tp/(tp + fn)$$

I allmänhet finns det ett omvänt förhållande mellan de två sista måtten. Av denna anledning så finns ett ytterligare prestandamått som tillämpas på s.k. *unbalanced classification problems* (*unbalanced* avser obalansen i träningsdata som vid feldetektion beror på att man naturligt har färre datapunkter för avvikelser än det normal driftfallet) där den s.k. *F1-score* beräknas enligt:

$$F1 = 2 \times \text{precision} \times \text{recall}/(\text{precision} + \text{recall})$$

Avsnitt 6.7.4.2. ger ett exempel på hur prestandamått från confusion matrix kan användas för att utvärdera olika feldetektionsmetoder.

Det finns även probabilistiska prestandamått som baseras på confusion matrix, vanligtvis genom att tillämpa ett specifikt tröskelvärde. Exempelvis så ger *area under curve – AUC* prestandan hos en binär klassificerare och sammanfattar metodens prestanda i ett enda mått som sträcker sig från 0 till 1. En optimal metod som kan separera avvikelser från normal data en AUC $\sim$1, medans en slumpmässigt val a vilka data som är avvikelser alltid får en en AUC på 0,5, se Frank *et al.* (2019) för detaljer. AUC kan visualiseras med en s.k. ROC-kurva ('receiver operating characteristic – ROC') (Bradley, 1997) genom att plotta andelen korrekta avvikelsedetektion som en funktion av andelen falsklarm. Den yta som fås om man beräknar integralen av ROC-kurva ger AUC-värdet (Russo *et al.*, 2021). Det är även möjligt att nyttja de prestandamått som användas för att analysera hur bra en regression är, se (Hauduc *et al.*, 2015) för detaljer.

6.7.4.1.2.2 Dynamiska och hybridmått

Dynamiska prestandamått kräver annoterade dynamiska data av t.ex. två binära signaler (t.ex. 0 betyder att data är normala och 1 att en avvikelse finns) eller en multivariat tidsserieanalys av två signaler.

Kurtoglu *et al.* (2008) definierade olika dynamiska mått för feldetektion såsom tid till detektion och *detection stability factor*. Dessa indikatorer används i system där en snabb motåtgärd krävs för at bibehåll en säker drift. Exempelvis så används tiden från det att en störning initieras till att den detekteras som prestandamått i flyg- och rymdtillämpningar (se t.ex. Vachtsevanos *et al.*, 2006). Därför är sådana dynamiska mått också intressanta i reningsverkstillämpningar där snabba motåtgärder är viktiga. Tidsberoende prestandamått är relevanta för processer där utdata från feldetektionen kan följas kontinuerligt och där tiden till upptäckt är kritisk.

Exempelvis så är en snabb detektion av avvikande syrehaltsvärden avgörande för att spara energi och för att kunna skifta till en reservdriftsstrategi som bibehåller reningsfunktionen, läs mer om feltolerant styrning i (Abbaspour *et al.*, 2020). Ett exempel på ett mer avancerat dynamiskt prestandamått beskrivs av Corominas *et al.* (2011), som föreslår ett tidsbaserat hybridindex för att utvärdera feldetektionsmetoder. Detta index tar hänsyn till flera olika egenskaper såsom falsklarm, missade avvikelser och om metoden felaktigt indikerar att en avvikelse återgått till normaltillståndet

(kod prestandamåtten ges i samma artikel). Mer specifikt så straffar indexet falsklarm med en konstant hastighet och straffar missade avvikelser med en exponentiellt avtagande funktion, som ger en maximal straffpoäng över tid. De prestandamått som beskrivs av Corominas *et al.* (2011) inkluderar: G (straffpoäng) som erhålls med den utvärderade metoden, G_{max} (maximala straffpoängen) när helt felaktig detektering antas, J övergripande tillförlitlighet, J_{FAC} metodens tillförlitlighet mot missade detektioner, och J_{FAL} metodens tillförlitlighet mot falsklarm.

6.7.4.1.3 Steg 3: välja ut data för prestandautvärdering

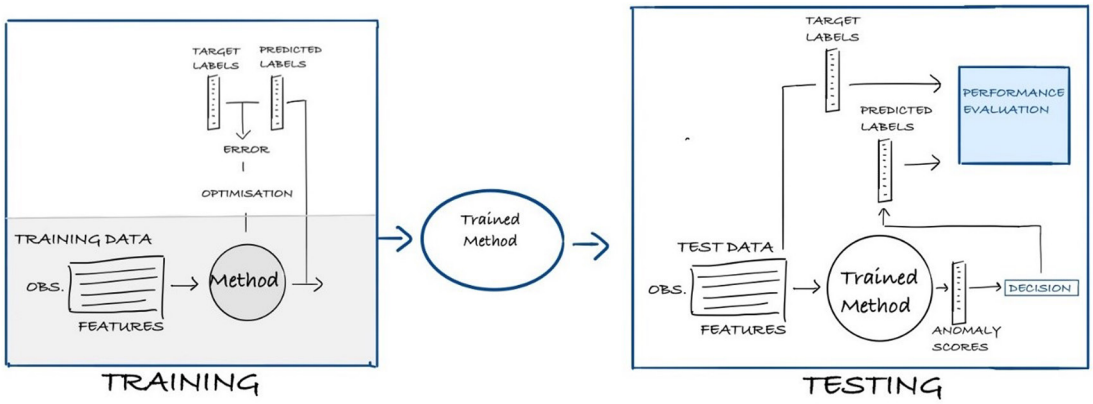
Att välja ett bra dataset för prestandautvärdering (benämns här som utvärderingsdataset) har liknande utmaningar som att välja ett bra dataset för att träna metoden (träningsdataset). En skillnad är att utvärderingsdatasetet behöver innehålla annoterade data från båda klasserna, det vill säga annoteringar av både normala och avvikande data. Oövervakade metoder kan däremot tränas med oannoterade data.

I idealfallet så används *ground truth* och det kända tillståndet för annotering. Tyvärr är det ofta svårt att ta fram i praktiken. Som alternativ kan subjektiva expertannoteringar användas för att manuellt skilja mellan normala och avvikande datasegment i ett dataset. Det bör noteras att avvikelser naturligt är sällsynta, vilket också återseglas i data. Detta gäller både verkliga data och utvärderingsdatasetet. Rent teoretiskt är det önskvärt att ha en balanserad andel avvikande och normala data i både tränings- och utvärderingsdata. Men detta är svårt att uppnå i praktiken av flera skäl. Anta som exempel att ett obalanserat scenario för feldetektion där 1 % av observationerna utgörs av avvikelser. För balansera utvärderingsdatasetet behöver 98 % av normaldata exkluderas från datasetet, vilket därmed riskerar att slänga bort informativa datapunkter. För att förtydliga, ett vanligt antagande är att normaldata kommer från en och samma population, vilket kanske inte alltid är fallet. Istället kan det finnas data i normalklassen där feldetektionsmetoden inte fungerar bra. Att ta bort sådana data vid utvärderingen begränsar analysen och möjligheten att förstå hur feldetektionsmetoden modellen presterar i ett allmänt scenario och om det behövs fler åtgärder, till exempel omkalibrering med olika hyperparametrar. Att bestämma vilka data i normalklassen som ska tas bort skapar dessutom ytterligare en hyperparameter som skulle behöva testas och bestämmas. En sådan teknik (benämns med *undersampling*), rekommenderas inte för utvärderingsdatasetet. Istället bör prestandamått väljas som tar hänsyn till obalansen i endera klassen, t.ex. F1-score, precision-recallkurvor etc. Då behövs ingen undersampling, utan istället en noggrann tolkning av hur obalansen påverkar utfallet av prestandamåttet.

Behovet av annoterade data beror också på komplexiteten i de valda prestandamåtten. Till exempel kräver dynamiska prestandamått kunskap om när avvikelsen initierades, det vill säga övergången från normaltillstånd till avvikelse i tiden. De statiska mätvärdena kan däremot utvärderas oberoende av kronologisk ordning. Ur annoteringssynpunkt är det mycket att notera tidpunkten när ett fel har åtgärdats, snarare än när felet började. Därför är det lättare att få fram utvärderingsdata för statiska mått där det räcker med ett fel följt av en åtgärd, till skillnad från de dynamiska måtten där den motsatta ordningsföljden krävs. Särskilda experiment med avsiktliga genererade störningar och avvikelser kan behövas för att utvärdera dynamiska mått, såvida inte avvikelsen är ett frekvent och tydligt problem (och i så fall bör termen avvikelse kanske revideras).

Oavsett om statiska eller dynamiska prestandamått används så bör datasetet vara representativt för den valda tillämpningen och de förväntade förhållandena. Med representativitet avses här process- och inflödesvariationer, brusnivåer och det driftsätt som överensstämmer med de förhållanden där feldetektionsmetoden är avsedd att användas. Representativitet avser också hur väl avvikelsen uttrycks, vilket kan liknas vid ett signal/brusförhållande. Att t.ex. utvärdera prestandan för en feldetektionsmetod på data som innehåller en 300 % stor avvikelse ger begränsad insikt i hur bra metoden är om målet är att detektera en bias på 10 %.

Slutligen är en generell regel (och förmodligen den viktigaste) att *alltid* använda separata dataset för träning och utvärdering av feldetektionsmetoden för att säkerställa oberoende prestandaresultat.



Figur 6.39 Arbetsflöde för att träna och utvärdera datadrivna feldetektionsmetoder. Vänster: den övervakade metoden använder ett annoterat träningsdataset. Om annoteringar delvis eller helt saknas används oövervakade metoder eller 'one-class calssification methods' vid träning vilket indikeras med grå ruta. Höger: När feldetektionsmetoden har tränats analyseras dess prestanda på testdata med prestandamått. Notera att testdata alltid behöver vara annoterade oavsett metod.

6.7.4.1.4 Steg 4: jämföra feldetektionsmetoderna

Figur 6.39 beskriver ett klassiskt arbetsflöde för att utvärdera prestandan av feldetektionsmetoder. I den vänstra figuren tränas metoden utifrån annoterade eller oannoterade data (se Avsnitt 6.7.2). Observera att det dataset som används för prestandautvärdering behöver vara annoterat, oavsett om feldetektionsmetoden oövervakad eller övervakad.

Den högra figuren illustrerar hur testdata används för att utvärdera feldetektionsmetoderna. Prestandamåtten predikterar tillståndet för de olika testdatapunkterna, vilket sedan jämförs med det sanna (annoterade) tillståndet. Som tidigare nämnts så kan testdata antingen annoteras subjektivt av en domänexpert eller utifrån kännedom om det sanna tillståndet.

6.7.4.1.5 Steg 5: tolka och sätt in resultaten i sitt sammanhang

Det avslutande steget innebär att tolka utfallet av de olika prestandamåtten utifrån den tänkta tillämpningen. Detta innebär att man reflekterar kring det ursprungliga syftet (steg 1), identifierar avvikelser mellan verkligheten och de använda dataseten (steg 3) samt jämför resultaten för de studerade metoderna med en enkel referensmetod, förslagsvis utifrån en av följande två alternativ.

Som referensmetod kan en enkel metod väljas som kan utvecklas och implementeras utan särskilda statistikkunskaper. Denna metod fungerar som referens för hur mycket bättre en avancerade metoder är, jämfört med de feldetektionsmetoder som analyseras. En enkel metod har flera fördelar. Den är lätt att lära ut, implementera och använda. Mer avancerade metoder bör prestera bättre än en enkel referensmetod för att motivera dess högre implementeringskostnad.

Ett annat alternativ för referensmetod är att jämföra med en slumpmässig detekteringsmetod. Alla feldetektionsmetoder bör prestera bättre än slumpmässiga gissningar.

Genom att jämföra feldetektionsmetoderna prestanda med referensmetoder blir det enklare att tolka de studerade metodernas prestanda i absoluta tal.

6.7.4.2 Ett praktisk exempel på prestandautvärdering

I detta avsnitt illustreras hur prestandamåtten kan tillämpas i praktiken genom att replikera en del av de resultaten i Samuelsson *et al.* (2018). Först beskrivs de data och därefter confusion matrix ned relaterade mått. Därefter beskrivs de hybridmått som används och kombinerar statiska och dynamiska

mätvärden. Två feldetektionsmetoder, A och B, jämförs med två referensmetoder: den pragmatiska ingenjörsstrategin och den slumpmässiga krona-eller-klave-detektorn.

Exemplet baseras på data från nedsmutsade och rena syregivare på ett större reningsverk (Bromma, Sverige). Datasetet innehåller syregivarnas avdrift på grund av successiv igensättning under 7 månader.

Steg 1: Definiera syfte med feldetektionen

Målet är att upptäcka felaktiga syrehaltsmätningar på grund av igensättning. En bias på 0,5 mg/l definierades som den lägsta detektionsgräns för att feldetektionsmetoden skulle vara användbar för prediktivt givarunderhåll. Det är önskvärt att uppnå detektionsmålet med en maximal falsklarmsfrekvens på 10 %. Att uppnå en låg andel falsklarm anses prioriterat jämfört med att uppnå en hög detektionsfrekvens (dvs vi gör hellre underhåll lite sent än i onödan).

Steg 2: Välj prestandamått

Det viktigaste prestandamåttet är att beräkna confusion matrix (Figur 6.40), som är lätt att tolka och används vidare för att visualisera ROC-kurvan, accuracy, precision, recall och F1-score. Hybridmåtten i Corominas *et al.* (2011) samt mean-time-to-detection inkluderades för att täcka in tidsaspekterna (dynamiska aspekter), som ett komplement till de statiska måtten.

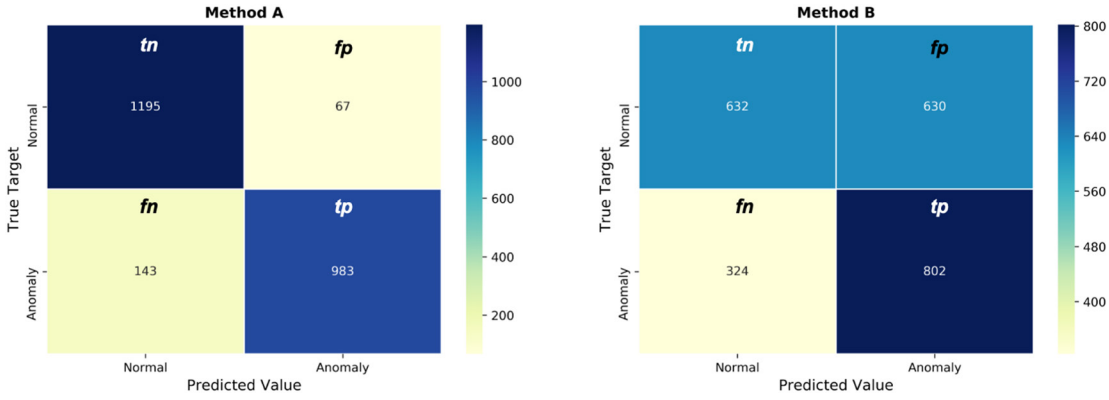
Steg 3: Välj dataset

Data i Samuelsson *et al.* (2018, 2019) togs fram i syfte att utvärdera metoder för feldetektering. Datasetet innehåller därför annoteringar för normaldata och avvikande observationer, samt fem perioder för träningsdata. Detaljer om data finns i originalpublikationen.

Steg 4: Jämför feldetektionsmetoderna utifrån prestandamått

Statiska prestandamått: I detta exempel och för metod A så visar confusion matrix att av 1262 normala datapunkter och 1126 avvikande så klassificerar metoden de flesta korrekt (1195 respektive 983 i Figur 6.40).

Metod B, å andra sidan, klassificerar ungefär hälften av de normal datapunkter korrekt (632) och 802 av de avvikande datapunkterna. Dessa resultat kan extrapoleras ytterligare enligt accuracy, precision, recall och F1-score vilka redovisas i Tabell 6.11 för båda metoderna.



Figur 6.40 Confusion matrix för metoderna A (vänster) och B (höger). Observera att höga värden i rutorna på diagonalen indikerar en metod med hög precision och recall.

Tabell 6.11 PFAC,0, PFAC,sat och tFAC-värdena för missad detektion var 20, 50 respektive 3 prover (45 min) och kswitch valdes till 1.

	Metod A	Metod B	Den pragmatiska metoden	Slumpen
Falsklarm	**0,053**	0,50	0,13	0,10
Precision	**0,94**	0,56	0,87	0,46
Recall	0,87	0,71	**0,99**	0,094
$F1$-score	0,90	0,63	**0,91**	0,16
Accuracy	**0,91**	0,60	**0,91**	0,52
Mean time to detection (dygn)	1,60	0,94	0,42	18,71
G (straffpoäng) för metoderna	9,1e3	2,9e + 4	6,1e3	5,2e4
$G_{\max}$ (maximalt möjliga antal straffpoäng)	8,0e4	8,0e4	8,0e4	8,0e4
J (bästa värde är 1) tillförlitlighet	**0,88**	0,64	0,92	0,35
	0,88	0,71	0,97	0,09
J_{FAL} (bästa värde är 1) tillförlitlighet mot falsklarm	**0,95**	0,50	0,86	0,90

PFAL,0 för falsklarm sattes till 20. Fet text anger bästa metod för de olika prestandamåtten.

Det kan noteras att precision, recall och F1-score ger liknande värden för normala och avvikande data. Faktum är att dessa mått vanligtvis endast används vid obalanserade datamängder (i det här fallet är i stället normala data och avvikelser balanserade). Det är dock alltid viktigt att normalisera data per klass. Även om metod B uppvisar en total accuracy på 0,60, vilket kan verka godtagbart, så är dess recall för normala data inte bättre än en slumpmässig gissning på 0,50.

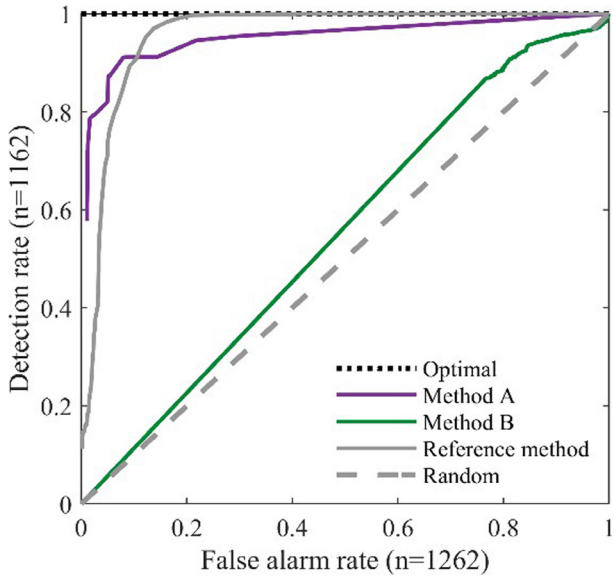
Statistiken från confusion matrix visualiseras också med en graf för *receiver operating characteristics* (ROC) (Kay, 1998) i Figur 6.41.

Många maskininlärningsstudier inkluderar istället motsvarande precision-recall-kurva, som är att föredra framför ROC om obalanserad dataset med få avvikande data används för utvärderingen Ozenne *et al* (2015).

Hybrida och dynamiska prestandamått:

I detta exempel används också det dynamiska prestandamått i Corominas *et al.* (2011). Flera parametrar måste ställas in för metoden där följande parametervärden användes: PFAC = 0, PFAC,sat och tFAC-värdena för missade avvikelser var 20, 50 respektive 3 datapunkter (45 min) och kswitch valdes till 1. PFAL,0 för falsklarm sattes till 20. Utvärderingen är upplagd så att en perfekt detektering ger 0 straffpoäng (G) och ett helt felaktigt klassificering får det maximala straffet $G_{\max}$ ($8{,}0 \times 10^4$ i detta exempel). Följaktligen skulle en perfekt detektion ge 100 % tillförlitlighet (J) och en helt felaktig klassificering ge 0 % tillförlitlighet (J). Om feldetektionssystemet resulterar i falsklarm, fördröjda detektioner, fördröjda med tröghet eller intermittenta detektioner, ökar straffpoängen jämfört med fallet med perfekt detektion. Generellt sett var metod A mer tillförlitlig (88 %) än metod B (64 %). Ändå är metod A inte mycket bättre än den pragmatiska ingenjörsstrategin (92 %). Den slumpmässiga detekteringen resulterar i det överlägset sämsta resultatet (35 %).

När man endast utvärderar systemets tillförlitlighet mot missade detektioner (J_{FAC}) är metod B mindre tillförlitlig (71 % tillförlitlig) jämfört med metod A (88 % tillförlitlig) eftersom den har sämre förmåga att identifiera avvikande händelser. Den pragmatiska metoden är dock mer tillförlitlig jämfört med metod A. Utifrån perspektivet falsklarm får metod B fler straffpoäng jämfört med metod A (G_{FAL} av X respektive Y poäng), vilket ger en tillförlitlighet mot falsklarm på 50 % respektive 95 % (J_{FAL}). Metod A är mycket mer tillförlitlig mot falsklarm. Den slumpmässiga metoden är tillförlitlig till 90 % och den pragmatiska metoden till 86 %.



Figur 6.41 Receiver operating characteristics (ROC) för metod A (lila heldragen linje) och metod B (grön heldragen linje). Den optimala prestandan visas som referens (svart streckad linje) med en konstant 100 % detekteringsgrad och noll falsklarm. Den pragmatiska referensmetoden (grå heldragen linje) visar indikerar att metod B inter tillför något mervärde jämfört med referensmetoden. Den lägsta prestandan indikeras med en diagonal grå streckad linje, som ger resultatet för en detektor som singlar slant som detektionsmekanism (lika stor sannolikhet att utlösa ett larm som att inte utlösa ett larm).

Överlag presterar metod A bättre än metod B, men inte uppenbart bättre än den pragmatiska ingenjörsstrategin. Även om metod B har sämre prestanda än metod A så är den fortfarande bättre än den slumpmässiga metoden.

Steg 5: Tolka resultaten utifrån syftet

Av Figur 6.40 framgår att både metod A och B presterar bättre än den slumpmässiga detektorn, vilket är ett grundläggande krav för att de ska vara praktiskt användbara. Metod A, som presterar bättre än B, är dock inte bättre än den pragmatiska referensmetoden. Detta tyder på att den mer avancerade metoden ger ett mindre mervärde och om implementeringen är svår skulle referensmetoden vara ett bättre val. Ett annat sätt att tolka resultatet är att data är lätta att analysera, utan behov av avancerade metoder för att detektera avvikelserna.

Prestandamåtten i Tabell 6.11 stöder resultaten i Figur 6.41. Dessutom uppnås det nödvändiga detektionsmålet på 10 % falsklarm endast med metod A, som därför lämpar sig för implementering och finjustering. Metodernas prestanda utifrån hybridmåtten liknar de som erhålls med de statiska måtten. Hybridmåtten innehåller parametrar som kan justeras för att ge större vikt åt den dynamiska eller den statiska komponenten.

6.7.4.3 Nyttor och fallgropar vid prestandautvärdering av datadrivna metoder

I praktiken är det inte alltid så enkelt och tydligt att utvärdera feldetektionsmetoder som i exemplet ovan. Detta avsnitt beskriver fördelar och fallgropar som kan hjälpa slutanvändarna att tillämpa det beskrivna arbetsflödet i praktiken.

Godtyckliga namn på prestandamåtten: Många statiska prestandamått har vaga namn såsom accuracy och precision, vilket till synes kan tyda på en bra prestanda. Men för att kunna tolka resultaten är det viktigt att man förstår hur det aktuella prestandamåttet beräknas.

Obalanserade dataset: Balanserade dataset, det vill säga lika stora mängder avvikande och normal datapunkter, är önskvärda för de flesta mätvärden. Ett dataset med huvudsakligen normal data kan lätt ge nära 100 % accuracy om man använder en metod som aldrig ger ett larm.

Representativa data: Man bör alltid komma ihåg att metodens prestanda i en verklig tillämpning endast kommer att likna resultaten av prestandautvärderingen när verkligheten är representerad i prestandautvärderingsdata. Att använda ett dataset med uppenbara avvikelser kan ge bra resultat mätt med ett visst mått, men kan misslyckas i praktiken om avvikelsen är mindre uttalad.

Använd endast sofistikerade och flexibla metoder: Flexibla metoder som *deep neural networks* med många parametrar är i allmänhet benägna att överanpassa träningsdata och är mycket mer komplexa att underhålla när de används. Jämför alltid komplexa metoder med enkla referensmetoder för att bedöma hur sofistikerad metod som är nödvändig.

Bra och tillräckligt bra prestanda: Vanligtvis är förväntningarna på tillförlitligheten i metodens förutsägelser högre än vad som kan uppnås med hjälp av den tillgängliga noggranheten i data. Det är därför viktigt att tidigt bestämma vad som är tillräckligt bra för den önskade tillämpningen för inte slösa resurser på tillämpningar där datatillgängligheten eller datakvaliteten är bristfällig.

Man kan sammanfatta avsnittet med följande:

(1) Definiera tydligt vilken detektionsprestanda som behövs i ett tidigt skede av utvärderingen.
(2) Välj dataset som är tillämpliga med tanke på de data som finns tillgängliga för den specifika tillämpningen och som är representativa för det verkliga problemet. I annat fall måste data- och metadatainsamlingen revideras för att till exempel inkludera kommentarer om avvikelser och kända fel.
(3) Genomför prestandautvärderingen utifrån det övergripande syftet och dess sammanhang. Säkerställ att de data som används för utvärdering också återspeglar verkligheten.

6.8 SLUTSATSER

Det ställs allt högre krav på att datadrivna tekniker för bedömning av datakvalitet ska implementeras på reningsverk, som beslutsstöd och för att optimera driften. För att uppnå detta har ett arbetsflöde för databehandling definierats med huvudstegen inledande datagranskning, förbehandling av data och tekniker för att analysera datakvaliteten. Ett stort antal datadrivna tekniker har kortfattat presenterats, från grundläggande metoder till avancerade, inklusive både mekanistiska och datadrivna metoder. Kriterier har definierats för korrekt val av teknik:

- De databehandlingssteg som ingår i ett dataflöde är specifika för det aktuella fallet och syftet.
- Valet av teknik för förbehandling av data beror på egenskaperna i rådata och den valideringsmetod som ska tillämpas.
- Om redundans/grundläggande samband är kända bör mekanistiska tekniker tillämpas.
- Om redundans/grundläggande samband till största delen är okända bör empiriska datadrivna metoder tillämpas.
- Rimlighetskontroller och massbalanser bör vara en del av dataflödet då det är möjligt.
- Valet av en datadriven metod beror på tillgång på data och om de är annoterade eller omärkta data, samt syftet med datavalideringen.
- I skrivande stund undersöks hybridmetoder (kombination av mekanistiska och datadrivna) för att kompensera för begränsningarna i de båda metoderna.
- Bedömningen av prestandan hos feldetektionsmetoder beror i hög grad på de valda prestandamåtten (dynamiska och statiska), som måste väljas med hänsyn till de data som finns tillgängliga för den specifika tillämpningen.

Referenser

Abbaspour A., Mokhtari S., Sargolzaei A. and Yen K. K. (2020). A survey on active fault-tolerant control systems. *Electronics*, **9**(9), 1513, https://doi.org/10.3390/electronics9091513

Aguado D. and Rosén C. (2008). Multivariate statistical monitoring of continuous wastewater treatment plants. *Engineering Applications of Artificial Intelligence*, **21**(7), 1080–1091, https://doi.org/10.1016/j.engappai.2007.08.004

Aguado D., Ferrer A., Ferrer J. and Seco A. (2006). Comparison of different predictive models for nutrient estimation in a sequencing batch reactor for wastewater treatment. *Chemometrics and Intelligent Laboratory Systems*, **84**, 75–81, https://doi.org/10.1016/j.chemolab.2006.03.009

Aguado D., Ferrer A., Ferrer J. and Seco A. (2007). Multivariate SPC of a sequencing batch reactor for wastewater treatment. *Chemometrics and Intelligent Laboratory Systems*, **85**(1), 82–93, https://doi.org/10.1016/j.chemolab.2006.05.003

Aguado D., Noriega-Hevia G., Ferrer J., Seco A. and Serralta J. (2022). PLS-based soft-sensor to predict ammonium concentration evolution in hollow fibre membrane contactors for nitrogen recovery. *Journal of Water Process Engineering*, **47**, 102735, https://doi.org/10.1016/j.jwpe.2022.102735

Alferes J., Tik S., Copp J. and Vanrolleghem Peter A. (2013). Advanced monitoring of water systems using in situ measurement stations: data validation and fault detection. *Water Science & Technology*, **68**(5), 1022–1030, https://doi.org/10.2166/wst.2013.302

Alferes J. and Vanrolleghem P. A. (2016). Efficient automated quality assessment: dealing with faulty on-line water quality sensors. *AI Communications*, **29**(6), 701–709, https://doi.org/10.3233/AIC-160713

Alferes J., Copp J., Weijers S. R., Cussonneau G., Fay G., Dembele A. and Vanrolleghem P. A. (2017). Validating data quality for water quality monitoring: Objective comparison of different data quality assessment approaches. ICA 2017.

Allison P. D. (2001). Missing Data 2001. Sage: Thousand Oaks.

Arteaga F. and Ferrer A. (2002). Dealing with missing data in MSPC: several methods, different interpretations, some examples. *Journal of Chemometrics*, **16**, 408–418, https://doi.org/10.1002/cem.750

Arteaga F. and Ferrer A. (2005). Framework for regression-based missing data imputation methods in on-line MSPC. *Journal of Chemometrics*, **19**(8), 439–447, https://doi.org/10.1002/cem.946

Arteaga F., Folch-Fortuny A. and Ferrer A. (2020). Missing data. In: Comprehensive Chemometrics, Elsevier, pp. 615–639. https://doi.org/10.1016/b978-0-12-409547-2.14718-3

Awhangbo L., Bendoula R., Roger J. and Béline F. (2020). Fault detection with moving window PCA using NIRS spectra for the monitoring of anaerobic digestion process. *Water Science and Technology*, **81**(2), 367–382, https://doi.org/10.2166/wst.2020.117

Baklouti I., Mansouri M., Ben Hamida A., Nounou H. and Nounou M. (2018). Monitoring of wastewater treatment plants using improved univariate statistical technique. *Process Safety and Environmental Protection*, **116**, 287–300, https://doi.org/10.1016/j.psep.2018.02.006

Baklouti I., Mansouri M., Hamida A. B., Nounou H. N. and Nounou M. (2021). Enhanced operation of wastewater treatment plant using state estimation-based fault detection strategies. *International Journal of Control*, **94**(2), 300–311, https://doi.org/10.1080/00207179.2019.1590735

Balmér P. and Hellström D. (2012). Performance indicators for wastewater treatment plants. *Water Science and Technology*, **65**(7), 1304–1310, https://doi.org/10.2166/wst.2012.014

Basseville M. (1993). Detection of abrupt changes theory and application. Prentice-Hall, Englewood Cliffs, New Jersey.

Basseville M. (2001). On fault detectability and isolability. *European Journal of Control*, **7**(6), 625–637, https://doi.org/10.3166/ejc.7.625-637

Bickel P. J. and Doksum K. A. (1981). An analysis of transformations revisited. *Journal of the American Statistical Association*, **76**(374), 296–311, https://doi.org/10.1080/01621459.1981.10477649

Bigot J. (2020). Statistical data analysis in the Wasserstein space. *ESAIM: Proceedings and Surveys*, **68**, 1–19, https://doi.org/10.1051/proc/202068001

Bikmukhametov T. and Jaschke J. (2020). Combining machine learning and process engineering physics towards enhanced accuracy and explainability of data-driven models. *Computers and Chemical Engineering* **2020**. https://doi.org/10.1016/j.compchemeng.2020.106834

Bishop C. M. (2006). Pattern Recognition and Machine Learning. Springer, United States of America.

Boger Z. (1992). Application of neural networks to water and wastewater treatment plant operation. *ISA Transactions*, **31**(1), 25–33, https://doi.org/10.1016/0019-0578(92)90007-6

Box G. E. P., Jenkins G. M., Reinsel G. C. and Ljung G. M. (2015). Time series analysis: Forecasting and control. 5th edn, John Wiley & Sons, Hoboken, New Jersey.

Bradley A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, **30**(7), 1145–1159, https://doi.org/10.1016/S0031-3203(96)00142-2

Buck S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society Series B*, **22**, 302–306.

Camacho J. and Ferrer A. (2014). Cross-validation in PCA models with the element-wise *k*-fold (EKF) algorithm: practical aspects. *Chemometrics and Intelligent Laboratory Systems*, **131**, 37–50, https://doi.org/10.1016/j.chemolab.2013.12.003

Candès E., Li X., Ma Y. and Wright J. (2011). Robust principal component analysis? *Journal of the ACM*, **58**(3), 1–37, https://doi.org/10.1145/1970392.1970395

Carlsson B. and Zambrano J. (2016). Fault detection and isolation of sensors in aeration control systems. *Water Science and Technology*, **73**(3), 648–653, https://doi.org/10.2166/wst.2015.529

Carvajal G., Roser D. J., Sisson S. A., Keegan A. and Khan S. J. (2015). Modelling pathogen log10 reduction values achieved by activated sludge treatment using naïve and semi naïve Bayes network models. *Water Research*, **85**, 304–315, https://doi.org/10.1016/j.watres.2015.08.035

Corominas L., Villez K., Aguado D., Rieger L., Rosén C. and Vanrolleguem P. (2011). Performance evaluation of fault detection methods for wastewater treatment processes. *Biotechnology and Bioengineering*, **108**(2), 333–344, https://doi.org/10.1002/bit.22953

Corominas L., Garrido-Baserba M., Villez K., Olsson G., Cortés U. and Poch M. (2018). Transforming data into knowledge for improved wastewater treatment operation: a critical review of techniques. *Environmental Modelling & Software*, **106**, 89–103, https://doi.org/10.1016/j.envsoft.2017.11.023

Corona F., Mulas M., Haimi H., Sundell L., Heinonen M. and Vahala R. (2013). Monitoring nitrate concentrations in the denitrifying post-filtration unit of a municipal wastewater treatment plant. *Journal of Process Control*, **23**, 158–170, https://doi.org/10.1016/j.jprocont.2012.09.011

Crowe C. M. (1996). Data reconciliation – progress and challenges. *Journal of Process Control*, **6**, 89–98, https://doi.org/10.1016/0959-1524(96)00012-1

Debruyne M. and Verdonck T. (2010). Robust kernel principal component analysis and classification. *Advances in Data Analysis and Classification*, **4**(2), 151–167, https://doi.org/10.1007/s11634-010-0068-1

Devlin J., Ming-Wei C., Kenton L. and Kristina T. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT (1).

Dutta A. and Sarkar S. (2015). Sequencing batch reactor for wastewater treatment: recent advances. *Current Pollution Reports*, **1**, 177–190, https://doi.org/10.1007/s40726-015-0016-y

Eriksson L., Byrne T., Johansson E., Trygg J. and Wikström C. (2013). Multi and Megavariate Data Analysis: Principles and Applications. ISBN 978-91-973730-5-0. Umetrics Academy.

Fan X. W., Du S. X. and Wu T. J. (2004). Rough support vector machine and its application to wastewater treatment processes. *Control and Decision*, **19**, 573–576.

Fawcett T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, **27**(8), 861–874, https://doi.org/10.1016/j.patrec.2005.10.010

Fellini S., Vesipa R., Boano F. and Ridolfi L. (2019). Fault detection in level and flow rate sensors for safe and performant remote-control in a water supply system. *Journal of Hydroinformatics*, **22**(1), 132–147, https://doi.org/10.2166/hydro.2019.002

Ferrer A. (2014). Latent structures-based multivariate statistical process control: a paradigm shift. *Quality Engineering*, **26**(1), 72–91. Huber, P. J. and Wiley, J. Robust statistics. Wiley, New York, https://doi.org/10.1080/08982112.2013.846093

Ferrer-Riquelme A. J. (2009). Statistical control of measures and processes. In: Comprehensive Chemometrics, S. Brown, R. Tauler and R. Walczak (eds.), Elsevier, Oxford, Vol. **1**, pp. 97–126.

Folch-Fortuny A., Arteaga F. and Ferrer A. (2015). PCA model building with missing data: new proposals and a comparative study. *Chemometrics and Intelligent Laboratory Systems*, **146**, 77–88, https://doi.org/10.1016/j.chemolab.2015.05.006

Forouzanmehr F., Le Q. H., Solon K., Maisonnave V., Daniel O., Buffiere P., Gillot S. and Volcke E. I. P. (2021). Plant-wide investigation of sulfur flows in a water resource recovery facility (WRRF). *Science of the Total Environment*, **801**, 149530, https://doi.org/10.1016/j.scitotenv.2021.149530

Frank S., Lin G., Jin X., Singla R., Farthing A., Zhang L., and Granderson J. (2019). Metrics and Methods to Assess Building Fault Detection and Diagnosis Tools. https://www.nrel.gov/docs/fy19osti/72801.pdf

Friedman J., Hastie T. and Tibshirani R. (2001). The Elements of Statistical Learning. Springer Series in Statistics, New York, Vol. **1**, No. 10.

Frogner C., Zhang C., Mobahi H., Araya M. and Poggio T. A. (2015). Learning with a Wasserstein loss. In: Advances in Neural Information Processing Systems, pp. 2053–2061.

Gallegos M. T. and Ritter G. (2005). A robust method for cluster analysis. *The Annals of Statistics*, **33**(1), 347, https://doi.org/10.1214/009053604000000940

García H. L. and González I. M. (2004). Self-organizing map and clustering for wastewater treatment monitoring. *Engineering Applications of Artificial Intelligence*, **17**(3), 215–225, https://doi.org/10.1016/j.engappai.2004.03.004

Gernaey K. V., Wade M. and Jørgensen S. B. (2004). Application of recursive principal component analysis to a continuous biological phosphorus removal process. *IFAC Proceedings Volumes*, **37**(3), 363–368, https://doi.org/10.1016/S1474-6670(17)32608-3

Goldberg Y. and Hirst G. (2017). Neural Network Methods in Natural Language Processing. Morgan & Claypool Publishers. ISBN-13: 978-1627052986.

Goodfellow I., Bengio Y. and Courville A. (2016). Deep Learning (Adaptive Computation and Machine Learning Series). ISBN-13: 9780262035613.

Granata F., Papirio S., Esposito G., Gargano R. and de Marinis G. (2017). Machine learning algorithms for the forecasting of wastewater quality indicators. *Water*, **9**(2), 105, https://doi.org/10.3390/w9020105

Grung B. and Manne R. (1998). Missing values in principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, **42**, 125–139, https://doi.org/10.1016/S0169-7439(98)00031-8

Gustafsson F. (2000). Adaptive Filtering and Change Detection. John Wiley & Sons Ltd, West Sussex, England.

Haimi H., Mulas M., Corona F. and Vahala R. (2013). Data-derived soft-sensors for biological wastewater treatment plants: an overview. *Environmental Modelling & Software*, **47**, 88–107, https://doi.org/10.1016/j.envsoft.2013.05.009

Haimi H., Corona F., Mulas M., Sundell L., Heinonen M. and Vahala R. (2015). Shall we use hardware sensor measurements or soft-sensor estimates? Case study in a full-scale WWTP. *Environmental Modelling and Software*, **72**, 215–229, https://doi.org/10.1016/j.envsoft.2015.07.013

Haimi H., Mulas M., Corona F., Marsili-Libelli S., Lindell P., Heinonen M. and Vahala R. (2016). Adaptive data-derived anomaly detection in the activated sludge process of a large-scale wastewater treatment plant. *Engineering Applications of Artificial Intelligence*, **52**, 65–80, https://doi.org/10.1016/j.engappai.2016.02.003

Haitovsky Y. (1968). Missing data in regression analysis. *Journal of the Royal Statistical Society Series B*, **30**, 67–82.

Hanbay D., Turkoglu I. and Demir Y. (2008). Prediction of wastewater treatment plant performance based on wavelet packet decomposition and neural networks. *Expert Systems with Applications*, **34**(2), 1038–1043, https://doi.org/10.1016/j.eswa.2006.10.030

Harrou F., Dairi A., Sun Y. and Senouci M. (2018). Wastewater treatment plant monitoring via a deep learning approach. IEEE International Conference on Industrial Technology (ICIT), 1544–1548.

Hastie T., Tibshirani R. and Friedman J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd edn, Springer Series in Statistics, Springer.

Hauduc H., Neumann M. B., Muschalla D., Gamerith V., Gillot S. and Vanrolleghem P. A. (2015). Efficiency criteria for environmental model quality assessment: a review and its application to wastewater treatment. *Environmental Modelling & Software*, **68**, 196–204, https://doi.org/10.1016/j.envsoft.2015.02.004

Henze M., Harremoes P., Jansen J. L. C. and Arvin E. (2002). Wastewater Treatment, Biological and Chemical Processes. Springer-Verlag Berlin, Heidelberg, New York.

Higham N. J. (2002). Computing the nearest correlating matrix – a problem from finance. *IMA Journal of Numerical Analysis*, **22**(3), 329–343, https://doi.org/10.1093/imanum/22.3.329

Hubert M., Rousseeuw P. and Verboven S. (2005). A fast method for robust principal components with applications to chemometrics. *Engineering Applications of Artificial Intelligence*, **60**(1–2), 101–111.

inCTRL Solutions (2018). Internal report.

inCTRL Solutions (2020). Internal report.

ISO 5725-1 (1994). Accuracy (trueness and precision) of measurement methods and results — Part 1: General principles and definitions.

IWA Task Group on Good Modelling Practices (2012). Guidelines for Using Activated Sludge Models (STR 22). IWA Publishing. eISBN: 9781780401164.

Jackson D. A. (1993). Stopping rules in principal components analysis: A comparison of heuristical and statistical approaches. *Ecology*, **74**(8), 2204–2214, https://doi.org/10.2307/1939574

Jackson J. E. (1980). Principal components and factor analysis: part I – principal components. *Journal of Quality Technology*, **12**(4), 201–213, https://doi.org/10.1080/00224065.1980.11980967

Jung D. (2015). Diagnosability Performance Analysis of Models and Fault Detectors. Linköping University Electronic Press, Linköping.

Kay S. M. (1998). Fundamentals of Statistics Signal Processing: Detection Theory. Prentice-Hall PTR, Upper Saddle River, NJ, USA.

Kazor K., Holloway R., Cath T. and Hering A. (2016). Comparison of linear and nonlinear dimension reduction techniques for automated process monitoring of a decentralized wastewater treatment facility. *Stochastic Environmental Research and Risk Assessment*, **30**(5), 1527–1544, https://doi.org/10.1007/s00477-016-1246-2

Kim J.-O. and Curry J. (1977). Treatment of missing data in multivariate analysis. *Sociological Methods & Research*, **6**, 215–240, https://doi.org/10.1177/004912417700600206

Kim M., Kim Y., Kim H., Piao W. and Kim C. (2016). Evaluation of the *k*-nearest neighbor method for forecasting the influent characteristics of wastewater treatment plant. *Frontiers of Environmental Science & Engineering*, **10**(2), 299–310, https://doi.org/10.1007/s11783-015-0825-7

Klanderman M. C., Newhart K. B., Cath T. Y. and Hering A. S. (2020). Case studies in real-time fault isolation in a decentralized wastewater treatment facility. *Journal of Water Process Engineering*, **38**, 101556, https://doi.org/10.1016/j.jwpe.2020.101556

Kohonen T. (2001). Self-Organizing Maps. Volume 30 of Springer Series in Information Sciences. Springer.

Kourti T. and MacGregor J. F. (1996). Multivariate SPC methods for process and product monitoring. *Journal of Quality Technology*, **28**(4), 409–428, https://doi.org/10.1080/00224065.1996.11979699

Kourti T., Lee J. and MacGregor J. F. (1996). Experiences with industrial applications of projection methods for multivariate statistical process control. *Computers & Chemical Engineering*, **20**, S745–S750, https://doi.org/10.1016/0098-1354(96)00132-9

Ku W., Storer R. and Georgakis C. (1995). Disturbance detection and isolation by dynamic principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, **30**(1), 179–196, https://doi.org/10.1016/0169-7439(95)00076-3

Kurtoglu T., Mengshoel O. J. and Poll S. (2008). International Conference on Prognostics and Health Management.

Le Q. H. (2019). Mass-balance-based Experimental Design and Data Reconciliation for Wastewater Treatment Processes. PhD thesis, Ghent University. *Chapter 4: Data evaluation for full-scale wastewater treatment plant: linear vs bilinear mass balances*. ISBN 9789463572453.

Le Q. H., Verheijen Peter J. T., Loosdrecht Mark C. M. van and Volcke Eveline I. P. (2018). Experimental design for evaluating WWTP data by linear mass balances. *Water Research*, **142**, 415–425, https://doi.org/10.1016/j.watres.2018.05.026

Le Q. H., Verheijen Peter J. T., van Loosdrecht Mark C. M. and Volcke Eveline I. P. (2022). Application of data reconciliation to a dynamically operated wastewater treatment process with off-gas measurements. *Environmental Science: Water Research and Technology*, **8**(10), https://doi.org/10.1039/D2EW00006G

Li W., Yue H., Valle-Cervantes S. and Qin S. (2000). Recursive PCA for adaptive process monitoring. *Journal of Process Control*, **10**(5), 471–486, https://doi.org/10.1016/S0959-1524(00)00022-6

Little R. J. A. and Rubin D. B. (2002). Statistical Analysis with Missing Data. Wiley, New York.

Lloyd J. R., Duvenaud D., Grosse R., Tenenbaum J. B. and Ghahramani Z. (2014). Automatic construction and natural-language description of non parametric regression models. 28th Association for the Advancement of Artificial Intelligence (AAAI), pp. 1242–1250.

Longo S., d'Antoni B. M., Bongards M., Chaparro A., Cronrath A., Fatone F., Lema J. M., Mauricio-Iglesias M., Soares A. and Hospido A. (2016). Monitoring and diagnosis of energy consumption in wastewater treatment plants. A state of the art and proposals for improvement. *Applied Energy*, **179**, 1251–1268, https://doi.org/10.1016/j.apenergy.2016.07.043

Loosdrecht M. C. M. V., Nielsen P. H., Lopez-Vazquez, C. M. and Brdjanovic D. (2016). Experimental Methods in Wastewater Treatment.

Lopes V. V. and Menezes J. C. (2005). Inferential sensor design in the presence of missing data: a case study. *Chemometrics and Intelligent Laboratory Systems*, **78**, 1–10, https://doi.org/10.1016/j.chemolab.2004.11.004

Madron F., Veverka V. and Vanecek V. (1977). Statistical-analysis of material balance of a chemical reactor. *AIChE Journal*, **23**, 482–486, https://doi.org/10.1002/aic.690230412

Maere T., Villez K., Marsili-Libelli S., Naessens W. and Nopens I. (2012). Membrane bioreactor fouling behaviour assessment through principal component analysis and fuzzy clustering. *Water Research*, **46**(18), 6132–6142, https://doi.org/10.1016/j.watres.2012.08.027

Mallat S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**(7), 674–693, https://doi.org/10.1109/34.192463

Mamandipoor B., Majd M., Sheikhalishahi S., Modena C. and Osmani V. (2020). Monitoring and detecting faults in wastewater treatment plants using deep learning. *Environmental Monitoring and Assessment*, **192**(2), 148, https://doi.org/10.1007/s10661-020-8064-1

Mark H. and Workman J. (2007). Chemometrics in Spectroscopy. Elsevier. ISBN: 978-0-12-374024-3. https://doi.org/10.1016/B978-0-12-374024-3.X5000-4

Markus Goldstein M. and Uchida S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PLoS ONE*, **11**(4), e0152173.

Maronna R., Martin D. and Yohai V. (2019). Robust Statistics: Theory and Methods (with R). Wiley, New York.

Marsili-Libelli S. (2006). Control of SBR switching by fuzzy pattern recognition. *Water Research*, **40**(5), 1095–1107, https://doi.org/10.1016/j.watres.2006.01.011

Metcalf and Eddy (2004). Wastewater Engineering, Treatment and Reuse. McGraw-Hill, New York.

Miletic I., Quinn S., Dudzic M., Vaculik V. and Champagne M. (2004). An industrial perspective on implementing on-line applications of multivariate statistics. *Journal of Process Control*, **14**(8), 821–836, https://doi.org/10.1016/j.jprocont.2004.02.001

Miletic I., Boudreau F., Dudzic M., Kotuza G., Ronholm L., Vaculik V. and Zhang Y. (2008). Experiences in applying data-driven modelling technology to steelmaking processes. *The Canadian Journal of Chemical Engineering*, **86**, 937–946, https://doi.org/10.1002/cjce.20090

Miletic I., Rieger L., Rosenthal A. and Schraa O. (2019a). Get More Out of Your Data – Advanced Methods for Automated Data Integration, Cleaning and Augmentation. WEFTEC 2019 Proceedings, 3266–3271.

Miletic I., Rieger L., Schraa O. and Wade M. (2019b). Measure What You Cannot Measure: Robust Multivariate Soft-Sensors for Influent PO4-P Prediction. WEFTEC 2019 Proceedings, 3272–3277.

Mitra A. (1998). Fundamentals of Quality Control and Improvement. 2nd edn, Prentice Hall, Upper Saddle River, NJ, USA.

Mitra A. (2008). Fundamentals of Quality Control and Improvement. 3rd edn, John Wiley and Sons, Hoboken, NJ, USA.

Montgomery D. G. (2019). Introduction to Statistical Quality Control. 8th edn, Wiley, New York, 768pp.

Montgomery D. C., Jennings C. L. and Kulahci, M. (2015). Introduction to Time Series Analysis and Forecasting. John Wiley & Sons. ISBN 978-1118745113.

Mulder C. de, Flameling T., Weijers S., Amerlinck Y. and Nopens I. (2018). An open software package for data reconciliation and gap filling in preparation of water and resource recovery facility modeling. *Environmental Modelling & Software*, **107**, 186–198, https://doi.org/10.1016/j.envsoft.2018.05.015

Muteki K., MacGregor J. F. and Ueda T. (2005). Estimation of missing data using latent variable methods with auxiliary information. *Chemometrics and Intelligent Laboratory Systems*, **78**, 41–50, https://doi.org/10.1016/j.chemolab.2004.12.004

Narasimhan S. and Jordache C. (2000). Data reconciliation and gross error detection: an intelligent use of process data, Annals of the New York Academy of Sciences. Gulf Publishing Company, Houston, Texas, US. https://doi.org/10.1111/j.1749-6632.1972.tb54778.x

Nelson P. R. C. (2002). Treatment of Missing Measurements in PCA and PLS Models. PhD dissertation, Department of Chemical Engineering, McMaster University, Hamilton, Ontario, Canada.

Newhart K. B., Holloway R. W., Hering A. S. and Cath T. Y. (2019). Data-driven performance analyses of wastewater treatment plants: a review. *Water Research*, **157**, 498–513, https://doi.org/10.1016/j.watres.2019.03.030

Ng A. Y. and Jordan M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. Advances in neural information processing systems, pp. 841–848.

Nomikos P. and MacGregor J. F. (1995). Multivariate SPC charts for monitoring batch processes. *Technometrics*, **37**(1), 41–59, https://doi.org/10.1080/00401706.1995.10485888

Nyberg M. and Frisk E. (2020). Model Based Diagnosis of Technical Processes. Linköping University.

Oakland J. S. (2003). Statistical Process Control. 5th edn, Butterworth-Heinemann, Bodmin, Cornwall, UK.

Ohmura K., Thürlimann C. M., Kipf M., Carbajal J. P. and Villez K. (2019). Characterizing long-term wear and tear of ion-selective pH sensors. *Water Science and Technology*, **80**(3), 541–550, https://doi.org/10.2166/wst.2019.301

Orlova D. Y., Zimmerman N., Meehan S., Meehan C., Waters J., Ghosn E. E. B., Filatenkov A., Kolyagin G. A., Gernez Y., Tsuda S., Moore W., Moss R. B., Herzenberg L. A. and Walther G. (2016). Earth mover's distance (EMD): A true metric for comparing biomarker expression levels in cell populations. PLoS ONE, https://doi.org/10.1371/journal.pone.0151859

Ozenne B., Subtil F. and Maucort-Boulch D. (2015). The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *Journal of Clinical Epidemiology*, **68**, 855–859, https://doi.org/10.1016/j.jclinepi.2015.02.010

Ozyurt D. B. and Pike R. W. (2004). Theory and practice of simultaneous data reconciliation and gross error detection for chemical processes. *Computers & Chemical Engineering*, **28**, 381–402, https://doi.org/10.1016/j.compchemeng.2003.07.001

Peters M. E., Neumann M., Iyyer M., Gardner M., Clark C., Lee K. and Zettlemoyer L. (2018). Deep contextualized word representations. Proceedings of NAACL-HLT 2018, pp. 2227–2237.

Pimentel M. A. F., Clifton D. A., Clifton L. and Tarassenko L. (2014). A review of novelty detection. *Signal Processing*, **99**, 215–249, https://doi.org/10.1016/j.sigpro.2013.12.026

Qin S. J. (2012). Survey on data-driven industrial process monitoring and diagnosis. *Annual Reviews in Control*, **36**(2), 220–234, https://doi.org/10.1016/j.arcontrol.2012.09.004

Rajaraman A. and Ullman J. D. (2011). 'Data Mining' (PDF). Mining of Massive Datasets, pp. 1–17.

Ramsay L. O. and Silverman B. W. (2005). Functional Data Analysis. 2nd edn, Springer Series in Statistics, https://www.springer.com/gp/book/9780387400808.

Rasmussen C. E. and Williams C. K. I. (2005). Gaussian Process for Machine Learning (Adaptive Computation and Machine Learning series). The MIT Press, Cambridge, Massachusetts, London, England.

Rieger L., Takacs I., Villez K., Siegrist H., Lessard P., Vanrolleghem P. A. and Comeau Y. (2010). Data reconciliation for wastewater treatment plant simulation studies-planning for high-quality data and typical sources of errors. *Water Environment Research*, **82**, 426–433, https://doi.org/10.2175/106143009X12529484815511

Roberts S. W. (1959). Control chart tests based on geometric moving averages. *Technometrics*, **1**(3), 239–250, https://doi.org/10.1080/00401706.1959.10489860

Rosén C. (2001). A Chemometric Approach to Process Monitoring and Control – with Applications to Wastewater Treatment Operation. PhD thesis, Lund University, Department of Industrial Electrical Engineering and Automation. Lund, Sweden.

Rosen C. and Lennox J. (2001). Multivariate and multiscale monitoring of wastewater treatment operation. *Water Research*, **35**(14), 3402–3410, https://doi.org/10.1016/S0043-1354(01)00069-0

Ruano M. V., Ribes J., Seco A. and Ferrer J. (2012). An advanced control strategy for biological nutrient removal in continuous systems based on pH and ORP sensors. *Chemical Engineering Journal*, **183**, 212–221, https://doi.org/10.1016/j.cej.2011.12.064

Rubin D. B. (1987). Multiple Imputation for Nonresponse in Surveys. Wiley & Sons, New York.

Rubner Y., Tomasi C. and Guibas L. J. (2000). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, **40**(2), 99–121, https://doi.org/10.1023/A:1026543900054

Russo S., Besmer M. D., Blumensaat F., Bouffard D., Disch A., Hammes F., Hess A., Moritz L., Blake M., Camille M., Eberhard M., Viet T. and Villez K. (2021). The value of human data annotation for machine learning based anomaly detection in environmental systems. *Water Research*, **206**, 117695, https://doi.org/10.1016/j.watres.2021.117695

Russo (2022). in preparation

Salman B. and Salem O. (2012). Modeling failure of wastewater collection lines using various section-level regression models. *Journal of Infrastructure Systems*, **18**(2), 146–154, https://doi.org/10.1061/(ASCE)IS.1943-555X.0000075

Samuelsson O., Björk A., Zambrano J. and Carlsson B. (2017). Gaussian process regression for monitoring and fault detection of wastewater treatment processes. *Water Science and Technology*, **75**(12), 2952–2963, https://doi.org/10.2166/wst.2017.162

Samuelsson O., Björk A., Zambrano J. and Carlsson B. (2018). Fault signatures and bias progression in dissolved oxygen sensors. *Water Science and Technology*, **78**(5–6), 1034–1044, https://doi.org/10.2166/wst.2018.350

Samuelsson O., Zambrano J., Björk A. and Carlsson B. (2019). Automated active fault detection in fouled dissolved oxygen sensors. *Water Research*, 166, https://doi.org/10.1016/j.watres.2019.115029

Sánchez-Fernández A., Baldán F. J., Sainz-Palmero G. I., Benítez J. M. and Fuente M. J. (2018). Fault detection based on time series modelling and multivariate statistical process control. *Chemometrics and Intelligent Laboratory Systems*, **182**, 57–69, https://doi.org/10.1016/j.chemolab.2018.08.003

Schafer J. L. (1997). Analysis of Incomplete Multivariate Data. 1st edn, Chapman and Hall/CRC, Boca Raton.

Schölkopf B., Smola A. and Müller K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, **10**(5), 1299–1319, https://doi.org/10.1162/089976698300017467

Schneider M. Y., Carbajal J. P., Furrer V., Sterkele B., Maurer M. and Villez K. (2019). Beyond signal quality: the value of unmaintained pH, dissolved oxygen, and oxidation–reduction potential sensors for remote performance monitoring of on-site sequencing batch reactors. *Water Research*, **161**, 639–651, https://doi.org/10.1016/j.watres.2019.06.007

Sousa A. C., Lucio M. M., Bezerra Neto, O. F., Marcone G. P., Pereira A. F., Dantas E. O., Fragoso W. D., Araujo M. C. and Galvão R. K. (2007). A method for determination of COD in a domestic wastewater treatment plant by using near-infrared reflectance spectrometry of seston. *Analytica Chimica Acta*, **588**(2), 231–6, https://doi.org/10.1016/j.aca.2007.02.022

Tanner M. A. and Wong W. H. (1987). The calculation of posterior distribution by data augmentation (with discussion). *Journal of the American Statistical Association*, **82**, 528–540, https://doi.org/10.1080/01621459.1987.10478458

Thomann M., Rieger L., Frommhold S., Siegrist H. and Gujer W. (2002). An efficient monitoring concept with control charts for on-line sensors. *Water Science and Technology*, **46**(4–5), 107–116, https://doi.org/10.2166/wst.2002.0563

Thürlimann C. M., Dürrenmatt D. J. and Villez K. (2015). Evaluation of qualitative trend analysis as a tool for automation. *Computer Aided Chemical Engineering*, **37**, 2531–2536, https://doi.org/10.1016/B978-0-444-63576-1.50116-3

Thürlimann C. M., Dürrenmatt D. J. and Villez K. (2018). Soft-sensing with qualitative trend analysis for wastewater treatment plant control. *Control Engineering Practice*, **70**, 121–133, https://doi.org/10.1016/j.conengprac.2017.09.015

Tian Z. X., Jiang J. P., Guo L. and Wang P. (2012). Anomaly detection of municipal wastewater treatment plant operation using support vector machine. International Conference on Automatic Control and Artificial Intelligence (ACAI 2012), pp. –521.

Vachtsevanos G., Lewis F., Roemer M., Hess A. and Wu B. (2006). Fault Diagnosis and Prognosis Performance Metrics. In: Intelligent Fault Diagnosis and Prognosis for Engineering Systems, John Wiley & Sons, Inc., pp. 355–399. ISBN: 978-0-470-11784-2. https://doi.org/10.1002/9780470117842.ch7

van der Heijden R. T. J. M., Heijnen J. J., Hellinga C., Romein B. and Luyben K. C. A. M. (1994). Linear constraint relations in biochemical reaction systems. 1. Classification of the calculability and the balanceability of conversion rates. *Biotechnology and Bioengineering*, **43**, 3–10, https://doi.org/10.1002/bit.260430103

Villez K., Rosen C., Stjin V.-H., Yoo C. K. and Vanrolleghem, P. (2005). On-line dynamic monitoring of the SHARON process for sustainable nitrogen removal from wastewater. *Computer-Aided Process Engineering*, **20**: 1297–1302, https://doi.org/10.1016/S1570-7946(05)80058-6

Villez K., Rosén C., D'hooge E. and Vanrolleghem P. A. (2010). Online phase length optimization for a sequencing batch reactor by means of the Hotelling's T^2 statistic. *Industrial & Engineering Chemistry Research*, **49**(1), 180–188, https://doi.org/10.1021/ie801907n

Villez K., Vanrolleghem P. A. and Corominas L. (2016). Optimal flow sensor placement on wastewater treatment plants. *Water Research*, **101**, 75–83, https://doi.org/10.1016/j.watres.2016.05.068

Vitorino D., Coelho S. T., Santos P., Sheets S., Jurkovac B. and Amado C. (2014). A random forest algorithm applied to condition-based wastewater deterioration modeling and forecasting. *Procedia Engineering*, **89**, 401–410, https://doi.org/10.1016/j.proeng.2014.11.205

Von Sperling M., Verbyla E. M. and Oliveira S. M. A. C. (2020). Assessment of Treatment Plant Performance and Water Quality Data. IWA Publishing.

Wade M. J., Sánchez A. and Katebi M. R. (2005). On real-time control and process monitoring of wastewater treatment plants: real-time process monitoring. *Transactions of the Institute of Measurement and Control*, **27**(3), 173–193, https://doi.org/10.1191/0142331205tm140oa

Walczak B. and Massart D. L. (2001). Dealing with missing data. Part I. *Chemometrics and Intelligent Laboratory Systems*, **58**, 15–27, https://doi.org/10.1016/S0169-7439(01)00131-9

Walther G., Zimmerman N., Moore W., Parks D., Meehan S., Belitskaya I., Pan J. and Herzenberg L. (2009). Automatic clustering of flow cytometry data with density-based merging. Advances in Bioinformatics. https://doi.org/10.1155/2009/686759

Wang X., Kruger U. and Irwin G. (2005). Process monitoring approach using fast moving window PCA. *Industrial & Engineering Chemistry Research*, **44**(15), 5691–5702, https://doi.org/10.1021/ie048873f

Wang X., Fan Y., Huang Y., Ling J., Klimowicz A., Pagano G. and Li B. (2021). Solving sensor reading drifting using denoising data processing algorithm (DDPA) for long-term continuous and accurate monitoring of ammonium in wastewater. *ACS ES&T Water* **1**(3), 530–541, https://doi.org/10.1021/acsestwater.0c00077

Westerhuis J. A., Gurden S. P. and Smilde A. K. (2000). Generalized contribution plots in multivariate statistical process monitoring. *Chemometrics and Intelligent Laboratory Systems*, **51**(1), 95–114, https://doi.org/10.1016/S0169-7439(00)00062-9

Wold S. (1978). Cross validatory estimation of the number of components in factor and principal component analysis. *Technometrics*, **20**(4), 397–405, https://doi.org/10.1080/00401706.1978.10489693

Wold S., Albano C., Dunn W. J., Esbensen K., Hellberg S., Johansson E. and Sjöström M. (1983). Pattern recognition: finding and using regularities in multivariate data. In: Food Research and Data Analysis, H. Martens and H. Russwurm Jr. (eds.), Applied Science Pub, London, Vol. **3**, pp. 183–185.

Wold S., Sjöström M. and Eriksson L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, **58**(2), 109–130, https://doi.org/10.1016/S0169-7439(01)00155-1

Yoo C., Lee J.-M., Lee I.-B. and Vanrolleghem P. (2004). Dynamic monitoring system for full-scale wastewater treatment plants. *Water Science and Technology*, **50**(11), 163–171, https://doi.org/10.2166/wst.2004.0684

Zambrano J., Samuelsson O. and Carlsson B. (2019). Machine learning techniques for monitoring the sludge profile in a secondary settler tank. *Applied Water Science*, **9**(6), 146, https://doi.org/10.1007/s13201-019-1018-5

Zhang D., Hølland E. S., Lindholm G. and Ratnaweera H. (2018). Hydraulic modeling and deep learning based flow forecasting for optimizing inter catchment wastewater transfer. *Journal of Hydrology*, **567**, 792–802, https://doi.org/10.1016/j.jhydrol.2017.11.029

Zhou P., Li Z., Snowling S., Baetz B., Na D. and Boyd G. (2019). A random forest model for inflow prediction at wastewater treatment plants. *Stochastic Environmental Research and Risk Assessment*, **33**, 1781–1792, https://doi.org/10.1007/s00477-019-01732-9

doi: 10.2166/9781789065107_0239

Kapitel 7
Slutsatser och framtidsutblick

D. Aguado[1], J. Alferes[2], A. Menniti[3], Q. Plana[4,5], M. V. Ruano[6], O. Samuelsson[7], K. Villez[8] och E. Zegers[9]

[1]*CALAGUA – Unidad Mixta UV-UPV, Institut Universitari d'Investigació d'Enginyeria de l'Aigua i Medi Ambient – IIAMA, Universitat Politècnica de València, València, Spanien*
[2]*VITO – WaterClimateHub, Ostende, Belgien*
[3]*Clean Water Services, Tigard, OR, USA*
[4]*modelEAU, Université Laval, Québec, QC, Canada*
[5]*Eurecat – Kataloniens teknikcentrum, Barcelona, Spanien*
[6]*CALAGUA – Unidad Mixta UV-UPV, Departament d'Enginyeria Química, Universitat de València, Burjassot, Spanien*
[7]*IVL Svenska Miljöinstitutet, Stockholm, Sverige*
[8]*Oak Ridge National Laboratory, Oak Ridge, TN, USA*
[9]*Staden Toronto, Toronto, ON, Kanada*

I det här kapitlet beskriver vi nuläget för insamling och hantering av metadata och reflekterar över vad IWA task group MetaCO har bidragit med. Vi ger också praktiska råd på hur denna rapport kan användas i den dagliga driften för datahanteringen med en kort guide som beskriver hur bästa möjliga praxis för metadata kan initieras. Kapitlet, och även denna rapport, avslutas med en kort framtidsutblick, med tankar om framtida möjligheter.

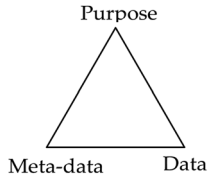
7.1 IDENTIFIERADE UTMANINGAR MED METADATA

Den här rapporten har gett praktiska och tekniska råd för insamling och användning av metadata, det vill säga *hur* metadata kan lagras och nyttiggöras. Mindre fokus har lagts på *varför* metadata bör vara ett strategiskt ämne som en del av en *digital roadmap*. Tydliga svar krävs på båda frågorna för att driva på utvecklingen och tillämpningen inom VA-branschen. I kommande avsnitt beskriver vi några viktiga utmaningar som är relaterade till dessa frågor, som har identifierats under arbetet med denna rapport.

7.1.1 Varför? – hur man motiverar metadata

Metadata har sällan ett egenvärde i sig själva. Deras värde ligger snarare i att göra data användbara för ett visst syfte. Det räcker inte med att metadata beskriver givardata, utan de insamlade metadata behöver också stödja ett specifikt *syfte* inom organisationen. Med väl hanterade givardata och ett tydligt definierat syfte kan metadata tillföra värde till en organisation (Figur 7.1). Om däremot någon

© 2025 IWAP. Detta är ett Open Access-bokkapitel som distribueras enligt villkoren i Creative Commons Attribution License (CC BY-NC-ND 4.0) som tillåter kopiering och vidaredistribution för icke-kommersiella ändamål utan derivat, förutsatt att arbetet citeras korrekt (https://creativecommons.org/licenses/by-nc-nd/4.0/). Kapitlet är från boken *Bästa möjliga data från givare i avloppsreningsprocesser*, Kris Villez, Daniel Aguado, Janelcy Alferes, Queralt Plana, Maria Victoria Ruano, Oscar Samuelsson (redaktörer).



Figur 7.1 Värdetriangeln för metadata som visar de komponenter som behövs för att skapa användbara insikter från data.

en av de tre komponenterna i Figur 7.1 saknas, kommer detta att begränsa värdet av tillgängliga (givar)data och i många fall inte ge användbara svar.

Att känna till syftet med givardata är avgörande för att bestämma vilka metadata som ska lagras. Detta kan vara enkelt för en specifik frågeställning, men blir genast utmanande när man överväger framtida datatillämpningarna (s.k. *unknown-unkowns*). Kapitel 3 ger dock en bra start till en grundstruktur. I de flesta fall rekommenderar vi att datahanteringsstrategier (*data management plans*) inledningsvis fokuserar på väldefinierade frågor (s.k. *known-unknowns*). Detta leder typiskt till ett fokus på metadata som beskriver givarunderhåll och givarnas datakvalitet.

7.1.1.1 Delning av data

En utmaning som är direkt kopplad till datadrivna modeller (ibland kallat AI) är det ökande behovet av att kunna dela data – även utanför den egna organisationen (*outside-the-fence*). Detta kommer sannolikt leda till att även personer utan domänkunskap inom vattenrening kommer bidra med dataanalys och modellering, där de resulterande black-box-modellerna (som beskrivs i Kapitel 6) används som beslutsunderlag.

En förutsättning är då strukturerade metadata för att förklara grundläggande antaganden om givardata som är underförstådda av en erfaren processingenjör. Ett exempel kan vara förväntad avdrift och opålitliga data från givare som inte underhållits. Interaktion med externa nya aktörer som vanligtvis inte samarbetar med reningsverk kommer också att skapa nya möjligheter och generera nya datakällor. På samma sätt som personer utan domänexpertis behöver vägledning för att tolka vattendata, behöver VA-organisationerna ett strukturerat arbetssätt för att analysera inkommande data i förhållande till dess metadata. Att ha en systematisk metadatastrategi kan även underlätta för extern dataanvändning.

7.1.1.2 Främja en metadatabejakande organisation

Metadata används redan för att underlätta datadelning inom VA-organisationen, där delning ofta innebär att data återanvänds för ett annat syfte. Grundläggande metadata såsom enheter och beskrivande datakvalitet kan minska risken för att felaktiga beslut tas av personal som har bristande insikter i hur data har genererats. Ett vanligt exempel är det korta organisatoriska minnet av givar- eller processtörningar. Utan en strukturerad och framtidssäkrad metadatainsamling kommer sådana kritiska anteckningar att försvinna i takt med att personal slutar och kommer i praktiken bara vara tillförlitliga och tillgängliga under en kort tid (veckor).

7.1.1.3 Kostnaden för metadata

Det kan vara frestande att lagra metadata i loggböcker eller kalkylark eftersom det går snabbt och enkelt utan några extra mjukvarukostnader. Den här typen av manuell registrering är dock tidskrävande och leder lätt till fel vid inmatning och export/delning. Vi rekommenderar istället en automatiserad metadatainsamling. Detta kräver dock investeringar i både skräddarsydd hårdvara, mjukvara och utbildning. Sådana investeringar betalar sig sällan på kort sikt, utan förbättrar snarare

datavärdet över tid. Att motivera sådana kostnader med en snabb investeringskalkyl är naturligtvis svårt, särskilt om syftet är vagt.

När man försöker implementera datahanteringsplaner eller digitala strategier är en av de största flaskhalsarna att kommunicera värdet av data till ledningen, i konkurrens med andra operativa prioriteringar. Att kunna översätta mervärdet av metadata och i allmänhet av databaserade applikationer i ekonomiska termer är fortfarande en stor utmaning.

Ett alternativt synsätt är att metadata, i viss utsträckning, är grundläggande och att det kommer att leda till förseningar och ökade kostnader för projekt om man inte lagrar dem eller har tillgång till dem vid strategiska beslut. Utifrån detta perspektiv bör kostnader för metadata inkluderas som en given delpost vid investeringsbeslut i givare eller digitala verktyg.

En fördel med metadata är att det sannolikt minskar kostnaden för att utveckla datadrivna modeller. En s.k. *supervised* modell (data med annoterade metadata, se Kapitel 6) kräver generellt mindre data och är billigare att kalibrera, i motsats till s.k. oövervakade (*unsupervised*) modeller som baseras på oannoterade data. I en maskininlärningstillämpning börjar man vanligtvis med att utforska stora datamängder för att identifiera givarsignaler som ger korrekta förutsägelser eller prognoser. Metadata kan stödja dataanalysen genom att automatiskt sålla ut relevanta signaler och tidsperioder för specifika prediktions- eller prognosändamål.

7.1.1.4 Minskad osäkerhet och ökat förtroende för ett datadrivet beslutsfattande

Avslutningsvis så är ett starkt motiv för att använda metadata att minska osäkerheten i beslut som baseras på data. Ett exempel är vid ombyggnation av ett reningsverk där den historiska slamproduktionen användas för att dimensionera en utökad nitrifikationsvolym. Om metadata för validering av suspgivaren indikerar en hög noggrannhet räcker mindre säkerhetsmarginaler för att täcka upp för mätosäkerheten. I en motsatt situation saknas historiska valideringsdata för suspgivaren, och därför genomförs en dedikerad mätkampanj för att validera alla suspgivare. Utvärderingen indikerar en 30-procentig offset i flera givare, vilket i sin tur indikerar att den historiska slamhalten sannolikt har överskattats. En stor säkerhetsfaktor behövs då för att täcka upp för den bristande tilltron till historiska data. Detta kan undvikas genom systematisk insamling och hantering av metadata.

7.1.2 Utmaningar med att implementera metadatastrategier

7.1.2.1 Förankra värdet med metadata

Det är inte alltid lika tydligt som i exemplet med suspgivaren. Faktum är att den svåraste utmaningen förmodligen är hur man ska kunna förankra värdet och behovet av metadata, eftersom de vanligtvis är anläggningsspecifika och har ett vagt syfte (Figur 7.1). Det är i allmänhet svårt att prissätta data om inte kostnaden för att *inte mäta* är hög. Ett tydligt syfte med en förväntad nytta och avkastning på investeringen kan dock bidra till att motivera investeringar i metadata.

7.1.2.2 Främja en metadatakultur

En lika viktig utmaning är att främja en metadatakultur på reningsverket. Detta innefattar både sociala och organisatoriska aspekter samt användarperspektivet kopplat till mjukvaror och IT-system. Inom MetaCO drevs en fokusgrupp, s.k. *utility group*, med inriktning på praktisk tillämpning av resultaten inom MetaCO. Under diskussionerna inom denna utility group så var det tydligt att det behövs nya roller i reningsverkets organisation. Ett bra exempel är en *data steward* som ansvarar för att säkerställa en bra datastruktur och datakvalitet, d.v.s. en god datahanteringspraxis (*data governance*). Det första steget mot en datadriven organisation är att inse, och ta konsekvenserna av, att datadrivna applikationer kräver underhåll, precis vilken fysisk pump som helst.

En annan social aspekt är fikonspråket för data som kan vara förvirrande med varierande förklaringar beroende på sammanhanget. I den här rapporten har vi försökt att klargöra den allmänt vedertagna terminologi som behövs för att kunna diskutera frågor som rör metadata (se Kapitel 2 och ordlistan med definitioner i slutet av rapporten).

7.1.2.3 Tekniska utmaningar för att implementera god metadatapraxis

Tekniska utmaningar utgör fortfarande ett hinder för en god praxis för insamling och strukturering av metadata. Det saknas kommersiella system för att underhålla och integrera reningsverkets metadata, vilket kräver skräddarsydda systemintegrationer som ställer ännu högre krav på andra domänspecialister. De mest uppenbara tekniska utmaningarna inkluderar:

- Automatisera insamlingen av metadata: Detta kan vara så enkelt som att lagra tidpunkten för rengöring av en givare eller hålla reda på den faktiska mätplatsen där en givare är monterad.
- Lagring av metadata. Att integrera och dela data mellan olika system är nyckeln till att skapa en användbar metadatabas. Som betonas i Kapitel 4 så måste både lagring och export av metadata vara smidigt, vilket det inte alltid är idag.
- Brist på standardisering: Det stora utbudet av kommersiella lösningar för datalagring och datahantering (inklusive kommunikationsprotokoll, databaser etc.) kan vara överväldigande, vilket i sin tur gör det svårt att välja en lämplig applikation.

7.2 BAKGRUNDEN TILL DENNA RAPPORT OCH HANTERING AV OBSERVERADE METADATAUTMANINGAR

Drivkraften till MetaCO kom från den gnagande känslan som ibland uppstår när man använder någon annans data. I många maskininlärnings- och modelleringsprojekt lägger man vanligtvis betydande andel av sin tid på att samla in metadata. Detta är ofta nödvändigt för att få fram data som är relevanta, är tillräckligt omfattande och av tillräcklig kvalitet. Ofta kräver detta diskussion med flera personer med kompetens inom informationsteknik (IT), givarhantering och personer från driften. Även om datainsamlingen görs noggrant kan de metadata som produceras vara osäkra, vilket kräver subjektiva beslut av dataanalytiker och modellerare i efterföljande steg. Det är den utmaningen som drev vårt initiativ.

Initialt så identifierade vi fyra utmaningar som vi hoppades kunna hantera i denna rapport (*Scientific and Technical Report – STR*): (a) okunskap om vad metadata är eller kan vara; (b) brist på förståelse för hur metadata bäst organiseras, (c) avsaknad av brett accepterade valideringsmetoder för metadata som beskriver en givares tillstånd, och (d) avsaknad av en översikt av datadrivna metoder som kan analysera givarsignaler och hur de kan användas för att producera metadata. Dessa utmaningar behandlades i Kapitel 3–6 vilket diskuteras i följande avsnitt med målsättningen att ge en subjektiv bedömning av om hur vart och ett av dessa kapitel har uppnått sitt mål.

7.2.1 Förstå vad metadata är eller kan vara

Efter Kapitel 2, som beskriver rapportens övergripande omfattning, gör Kapitel 3 en analys av frågan *Vad är metadata*?. Vi valde en övergripande strategi genom att definiera metadata för givarsignaler i ordets bredaste bemärkelse. Det visade sig vara enkelt att genom interna diskussioner samla in en omfattande lista över de typer av metadata som uppfyller vår definition. En insikt var att metadata för givare ibland begränsas till en beskrivning av hårdvaran och var, när och hur den används. Denna typ av metadata kan härledas från en kombination av processflödesschema, rör- och instrumentdiagram (*Process and Instrumentation Diagram – P&ID*). I motsats till denna snäva definition omfattar vår definition av metadata även datakvalitet, kontextuell information, underhåll och driftanteckningar. Detta gör att det är fritt att avgränsa sitt urval av metadata utifrån identifierade prioriteringar eller specifika projektbehov.

7.2.2 Förståelse för hur metadata bäst organiseras

I Kapitel 4 besvarade vi frågan *Hur hanteras metadata*?. Kapitlet ger en introduktion till både äldre och nyare system för datalagring och datahantering. Centralt i detta kapitel är idén att en god

datahanteringspraxis är nyckeln till att användbara givardata får en lång livslängd. Inledningsvis visade det sig vara svårt att samla in information, specifikt för reningsverk, eftersom utformningen och driften av databassystem ofta implementeras av tredjepartsleverantörer och IT-företag. Dessutom tenderar den här typen av teknik att vara generisk och inte specialiserad för just reningsverk. Denna generella karaktär innebär att behovet av att hantera metadata kräver noggranna förklaringar och diskussioner vid databasens utformning och implementering. Därför innehåller detta kapitel en översikt över databastekniker, rekommendationer för implementering och användning samt några exempel på databasdesign. Detta bör göra det möjligt för processexperter och anläggningsägare att initiera en konstruktiv diskussion med databasingenjörer och andra IT/OT-experter.

7.2.3 En generell praxis för metadata från givare

Givare nämns ofta som den viktigaste komponenten vid uppbyggnad, kalibrering och användning av dynamiska processmodeller och avancerade system för processövervakning och styrning. Trots detta allmänt erkända behov har vi noterat att valideringsrutiner i praktiken oftast baseras på metoder som inte är väldokumenterade. Därför var vår ursprungliga idé till Kapitel 5 att samla in standardiserade valideringsrutiner från olika reningsverk och sedan sammanställa dessa till en enhetliga rutin, en för varje givartyp. Valideringsrutiner för givare används vanligtvis för att jämföra mätvärdet från en givare med en referensmätning eller ett medium med ett känt värde för den uppmätta variabeln (t.ex. en standardlösning eller s.k. spikat vattenprov). Vi fann dock att utförandet ofta anses vara enkelt men att de jämförande resultaten sällan registreras för vidare analys eller för inspektion av trender. Dessutom blev det tydligt att valideringsrutiner för givare sällan finns dokumenterade. Detta står i tydlig kontrast till de analyser som sker på laboratorier där man följer skriftliga ruiner och standarder som även revideras regelbundet. Valideringsrutiner för givare skulle vara till stor nytta för att få kvalitetsbedömningen av givardata till en nivå som är jämförbar med laboratorier, som även skulle bidra till att dokumentera givarens prestanda. Vi tror dock att detta kommer förbli en utmaning under överskådlig tid eftersom det kan vara svårt att etablera systematiska valideringsrutiner för givare.

Även om det fortfarande saknas konsensus bland de reningsverken om vilka rutiner som bör användas, och hur de ska utformas, så ger Kapitel 5 förslag på flera möjliga valideringsrutiner för givare. Dessa rutiner baseras på befintliga rutiner hos reningsverk och har modifierats med det specifika syftet att noggrant samla in metadata som beskriver givarsignalens noggrannhet, bias, precision etc. Kapitlet ger en grund för givarvalidering med mallar och principiell utformning för de vanligast givarna på ett reningsverk. Vi uppmuntrar alla att prova dessa och bidra till förbättrade valideringsmetoder för givare.

7.2.4 En översikt över datadrivna metoder för validering av givardata

Datadrivna metoder för givarvalidering brukar delas in i två grupper. Den ena gruppen baseras på fysikalisk kunskap om processen (mekanistiska modeller) och fokuserar på att använda redundanta samband mellan givare för att upptäcka avvikelser och data med låg tillförlitlighet. Flödesbalanser och massbalanser används brett för datavalidering. Den andra gruppen baseras på en stor uppsättning datakrävande metoder såsom statistiska processövervakning, *data mining*, mönsterigenkänning (*pattern recognition*) och maskininlärning. Dessa två olika grupper beskrivs sällan samtidigt eftersom de har ursprung i olika vetenskapliga domäner. Av denna anledning gav Kapitel 6 en gemensam översikt över de vanligaste datavalideringsteknikerna på ett och samma ställe. Som förväntat var intresset stort bland dataanalytiker för att bidra, vilket resulterade i ett långt och omfattande kapitel. De beskrivna teknikerna har tillämpats minst en gång med data från en verklig process och reningsverk. Utifrån hittills publicerade litteratur har vi gett vägledning om urval och implementering av dessa tekniker för datavalidering. Kapitlet är därför en bra start och utgångspunkt för att implementera datavalidering i praktiken.

7.3 ANVÄNDARPERSPEKTIVET: ATT KOMMA IGÅNG MED DATA

Att generera och hantera metadata ingår i det större begreppet datahantering, vilket är en generell utmaning för många reningsverk. Av denna anledning leder diskussioner om metadata på reningsverk ofta till större diskussioner om givarhantering, system och processer för datahantering och reningsverkets dataarkitektur. Diskussioner i *utility groups* belyste tre teman för utmaningar och framgångsfaktorer med metadatahantering:

- Skapa en kultur som stöder och värdesätter datakvalitet.
- Skapa system för att hantera datakvalitet.
- Dataarkitektur och integration.

7.3.1 Skapa en kultur för datahantering – data governance

Ett bra ramverk med strukturer för datahantering är en viktig del i arbetet med att säkerställa att reningsverkets data är ändamålsenliga. Datahantering är en process där man tar fram policyer för hur data ska samlas in, lagras och bearbetas. Typiska grundelement inkluderar:

- Etablera ansvarsområden och ansvariga för olika delmängder av data.
- Kommunicera förväntad datakvalitet till dataanvändare (se t.ex. Kapitel 3),
- Dokumentera datakällor, beräkningar eller andra beräkningar och göra denna information tillgänglig för datakonsumenter (se t.ex. Kapitel 4).

När policyn för datahantering har börjat tillämpas kan typiska roller omfatta:

- ***Data steward***: Person som har ansvar för valda datakällor och ansvarar för att utreda dataproblem och underhålla datakällorna (i normalfallet finns en data steward som ansvarar för alla dataset).
- ***Data validator***: Person som kontrollerar att data är korrekta, giltiga, uppdaterade och kompletta (i allmänhet endast tillämpligt för utvalda dataset som kräver förhöjd kvalitet).
- ***Report owner***: Person som ansvarar för att ta fram en rapport som baseras på data, göra ändringar i den, svara på frågor om rapporten och hålla rapporten uppdaterad.

Reningsverk med välutvecklade policyer för datahantering som även omsatts i befintliga verksamhetsrutiner för att genomföra dessa policyer har vanligtvis en stark datakultur där värdet av data kommuniceras och där cheferna stöder de insatser som krävs. Denna kulturella komponent i en organisation är en kritisk framgångsfaktor för att fördelarna med de tekniska idéerna i den här rapporten ska kunna förverkligas fullt ut.

7.3.2 System för hantering av datakvalitet

Många reningsverk värdesätter högkvalitativa data och inser vikten av metadata. Men även reningsverk med stor erfarenhet av framgångsrikt arbete med givare kämpar med att hantera givarna och säkerställa kvaliteten på de data de producerar. Även om onlinegivare kan vara ett kraftfullt verktyg för driften, så kräver generellt en framgångsrik givarinstallation att samtliga villkor är uppfyllda:

- Ett tydligt och validerat behov av att en onlinegivare verkligen är nödvändig (realtidsdata eller nära realtidsdata är viktiga för tillämpningen; den insats som krävs för att underhålla ett online-instrument motiveras av värdet av att ha information i realtid).
- Ett väl genomtänkt val av givartyp (referenser och/eller jämförande fältförsök) och en representativ placering av givaren i processen.
- Lämpliga underhållsprocesser som är systematiserade med tydlig definition av roller och ansvarsområden för olika arbetsgrupper.
- Rutinmässig validering av givaren (t.ex. jämförelse av givarens mätvärde mot en referens, t.ex. en laboratoriemätning, se Kapitel 5).

- Möjlighet att enkelt dela givardata och valideringsdata inom hela organisationen.
- Policyer för dataförvaltning som vägleder dessa insatser.

7.3.3 Dataarkitektur och integrationer

Data som genereras på reningsverk finns i allmänhet i flera olika system. Några av de typiska system som är mest relevanta för hantering av givardata och metadata inkluderar:

- Historikdatabaser som lagrar data som genereras av SCADA-systemet, inklusive givardata.
- Underhållssystem lagrar data om kalibrering av utrustning och underhåll som utförs av underhållspersonal.
- Laboratoriedatasystem (*Laboratory Information Management System – LIMS*) som lagrar resultaten från laboratorieanalyser.
- Datahanteringssystem är en benämning som kan användas för att beskriva många helt olika typer av system. Men vissa reningsverk har ett system för att aggregera data som tar in information från andra system för att möjliggöra en mer intelligent drift. Den här typen av system kan omfatta:
  - Funktionalitet för databassökningar
  - Lagring av data som samlas in manuellt av operatörer (t.ex. avläsningar från handhållna givare)
  - Rapporter
  - Realtidsvisning av data i paneler (*dashboards*)

Att aggregera data från olika källor så att de kan analyseras på ett holistiskt och effektivt sätt är ett viktigt steg mot att förbättra användningen av metadata för att identifiera och diagnostisera problem och öka medvetenheten om datakvalitet. Oförmågan att analysera och visualisera data från en rad olika källor är en vanlig utmaning när det gäller att skapa och hantera metadata på ett vettigt sätt. Det finns en rad olika metoder för att integrera datakällor (se nedan, Avsnitt 7.4). Oavsett arkitektur är en framgångsrik användning av metadata, och effektiv hantering av givardata i allmänhet, beroende av integration mellan dessa olika datakällor.

Exempel på användbara integrationer inkluderar:

- LIMS och SCADA:
  - Validering av givardata (se Kapitel 5), särskilt för att jämföra mätvärden från givare med analysresultat från laboratoriet. Digitalisering och automatisering av denna valideringsprocess kan underlätta lagring och kommunikation av åtgärder och resultat från givarvalideringen till flera roller i organisationen.
- SCADA- och datahanteringssystem:
  - Återigen relevant för validering av givardata (t.ex. jämförelse av mätvärden från en givare med ett referensvärde, enligt Kapitel 5, eller automatisk detektering av avvikelser, enligt Kapitel 6). Genom att digitalisera och automatisera denna valideringsprocess minimeras användarnas arbetsinsats och informationen sprids till fler roller i organisationen.
- Underhållssystem och datahanteringssystem:
  - Registrering av genomförda underhållsaktiviteter (vad och när).
  - Registrering av kalibrering (typ av kalibrering, datum och tid, kalibreringsparametrar).
- Information i SCADA/OT-systemet som är av intresse för driftledningssystemet:
  - Givarens drifttimmar
  - Diagnostisk information från givaren om aktuell status.

Det är inte bara viktigt att systemen integreras, utan varje integration måste också hanteras över tid och framgången beror på hur mycket arbete som krävs för att göra detta. Exempel på frågor som är relevanta för ett långsiktigt ägandeskap är bland annat:

- Hur lätt är det att underhålla integrationen?
- Vad händer när fysiska komponenter tillkommer eller ersätts?

Historiska data om givare som tagits bort kan fortfarande visa sig vara användbara i framtiden.

7.3.4 Digital mognadsgrad

Alla reningsverk är på väg mot mer och/eller bättre användning av *digitala* verktyg (benämns ibland som den digitala transformationen). Hur långt ett enskilt VA-bolag har kommit på sin resa beskrivs ibland som dess *digitala mognadsgrad*. Digital mognad kan beskrivas som ett VA-bolags förmåga att på ett effektivt sätt utnyttja teknik och digitala verktyg. Även om en omfattande diskussion om digital transformation på reningsverk är utanför den här rapportens fokusområde, så berör vi konceptet här eftersom digital mognad ligger till grund för, och stöder, implementeringen av de koncept som utforskas i avsnitten ovan, till exempel datahantering och integration av olika IT-system. I slutändan skulle en högre digital mognadsgrad stödja en mer effektiv användning av metadata.

Under de senaste decennierna har mängden data som samlas in på reningsverken ökat dramatiskt. SCADA/DCS-system introducerades i mitten på 1900-talet och antalet dataposter som samlas in fortsätter att öka. Förutom SCADA-data samlar VA-bolagen även in laboratoriedata, vattenförbrukningsdata, kunddata, ingenjörs- och konstruktionsdata samt anläggningsdata om dess prestanda och underhåll. Denna explosion av data har dock medfört utmaningar på alla nivåer. På den mest grundläggande nivån är underhållet av givare en utmaning som bara har vuxit med antalet installerade givare. Ett steg uppåt i kedjan av dataanalys och visualisering har bristen på integration mellan datasystemen begränsat möjligheterna att nyttiggöra data. Med all den data som genereras har många reningsverk rapporterat att de saknar enkel åtkomst till sin data och att de inte litar på dess kvalitet.

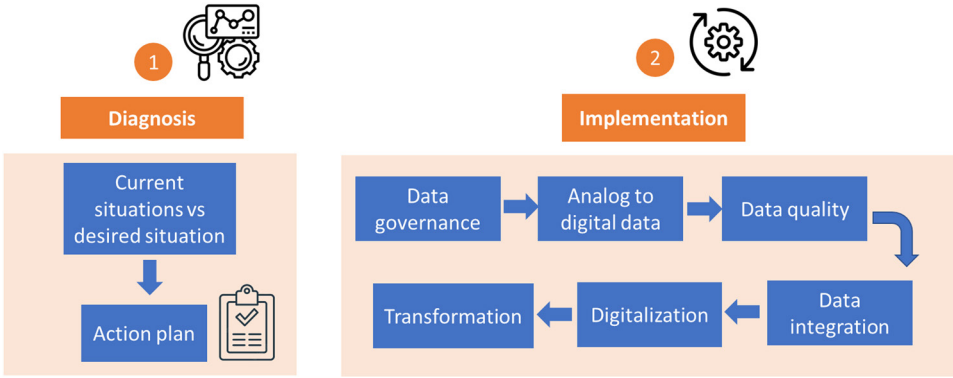
En ny studie från Water Research Foundation (2022) utforskar begreppet digital mognad och betonar att det inte bara handlar om teknik, utan även om underliggande beslutsprocesser och mänskliga faktorer. Nedan listas några faktorer som är relevanta för ett VA-bolags digitala mognad:

- Tillräcklig personalstyrka för att stödja digitala applikationer.
- Bemanningskompetens för att stödja digitala applikationer (t.ex. affärsanalytiker, dataanalytiker, företagsarkitekter, cybersäkerhetsspecialister).
- Definierat ramverk för styrning av företagets tekniska applikationer.
- Förekomst och effektiv implementering av policyer för data- och dokumenthantering.
- Policyer för IT-upphandling.
- Intern utbildning och supportresurser för digitala applikationer.
- Kopplingar mellan reningsverkets IT och den administrativa IT-gruppen.
- Kopplingar mellan operativ teknik (OT) och IT-gruppen.
- Kopplingar mellan olika yrkesroller (operatörer, ingenjörer, högre chefer).

Som nämnts ovan är alla VA-bolag i ett eller annat skede av den digitala transformationen. Enskilda VA-bolag kan ha kommit olika långt på denna resa som dessutom kan se mycket olika ut. Oavsett var ett VA-bolag befinner sig i den digitala transformationsprocessen är det värdefullt att påbörja arbetet med att skapa system för att hantera datakvalitet och tillhörande metadata. I takt med att man får erfarenhet av detta arbete kan dessa processer automatiseras eller på annat sätt omarbetas för att bli mer effektiva. Genom att skaffa sig erfarenhet får man en bredare kunskapsbas som kan vägleda framtida beslut på den digitala resan.

7.4 EN PLAN FÖR DATATRANSFORMATION

De behov som VA-bolagen nämnde ovan användes för att ta fram en plan för datatransformation med fokus på metadata. Datatransformation på reningsverken är en del av den större bilden av digital



Figur 7.2 Riktlinjer med de viktigaste stegen mot en omvandling av data och metadata.

transformation, och en central del av den digitala resan som VA-bolag, och i större utsträckning hela vattensektorn, måste genomföra för att möta de nuvarande utmaningarna: åldrande infrastruktur, minskade vattentillgång, befolkningstillväxt, klimatrelaterad störningar och striktare utsläppskrav.

I praktiken är diskussioner om att skapa och hantera metadata en del av en större diskussion om datahantering, givare och processtyrning, informationssystem, dataflöden och dataarkitektur. Denna digitala resa kan vara svår att genomföra för reningsverken. Konkurrerande prioriteringar gör det svårt att veta vilken riktning som är rätt att gå. Flera inledande frågor kan ställas:

- Hur initieras datatransformationen?
- Vilka investeringar eller uppgraderingar från nuvarande situation behövs?
- Vilken prioriteringen är rätt för de olika steg som ska tas?
- Hur stor budget krävs för data/digital transformation?
- Hur påverkar detta den allmänna verksamhetsplanen?
- Hur kan data och metadata inkluderas utan att de viktigaste verksamhetsmålen påverkas negativt? (Uppfylla standarder, minska kostnader, minimera miljöpåverkan).

Det sistnämnda borde ses som en möjlighet, i stället för ett hot, men i många fall är det inte så, eller åtminstone inte tillräckligt tydligt på operativ nivå och ledningsnivå. Det är som tidigare nämnts svårt att mäta avkastningen på investeringen av datahantering, även om det skulle kunna motivera investeringar för att minska kostnaderna och förbättra prestandan på lång sikt.

Som en riktlinje bör all data- eller digital transformation hos VA-bolagen följa den tvåstegsprocedur som beskrivs nedan och illustreras i Figur 7.2.

7.4.1 Steg 1: Diagnos

Det första steget är att kritiskt utvärdera den aktuella situationen och de specifika behoven utifrån VA-bolagets anläggningar. Detta kommer att göra det möjligt att definiera *nuläget* och definiera det *önskade läget*. Följande aspekter bör beaktas:

- Hur stor är den befintliga infrastrukturen, vilken kapacitet har den och hur gammal är den?
- Vilken nivån av styrning och automatiseringen används givarna till?
- Vilken är den generella digitala mognadsgraden?
- Hur stor är den interna kompetensen? Behövs ytterligare tekniska kompetenser eller specifika utbildningar?
- Vilka regler behöver vi förhålla oss till på olika lokal, national, EU, och global nivå?
- Vad är den långsiktiga målsättningen med reningsverket/anläggningen?

Svaren på dessa frågor bör resultera i en anpassad handlingsplan för att uppnå de önskade målen, med hänsyn till begränsningar och lokala förhållanden.

7.4.2 Steg 2: Genomförande

När nuläget har identifierats och mål har satts upp så genomförs den digitala resan i sex steg.

7.4.2.1 Automatisera manuell insamling av data och metadata

Detta innebär att analog information omvandlas till digital form och gäller all data och metadata som beskrivs i Kapitel 2 och 3. Ett bra exempel är de data som samlas in av operatörer (normalt med penna och papper) när de utför givarunderhåll. Om denna information finns tillgänglig som digitaliserade data kan den göras mer lättillgänglig, vilket underlättar dataanalys och påskyndar beslutsfattandet, vilket i slutändan kommer att leda till att risker och kostnader kan undvikas.

Som exempel kan nämnas att traditionella sätt att hantera givardata kan kräva en betydande och kontinuerlig rutininsats av instrumenttekniker. Ju mer rutinarbete som krävs, desto mindre sannolikt är det att ett system för datakvalitet kan upprätthållas över tid. Det kan dock vara svårt att veta vilka delar som kan eller bör automatiseras för att spara tid om man inte har erfarenhet av att implementera ett mer manuellt system till att börja med. Erfarenheterna kan användas för att planera förändringar av policyer för datahantering, system för hantering av datakvalitet, dataarkitektur och dataintegration. Att bygga upp dessa system på ett sätt som fungerar för varje enskilt VA-bolag kräver tid och tålamod, men det är värt besväret att få data som man kan lita på och ta beslut utifrån.

7.4.2.2 Utforma ett ramverk för datahantering

Att utforma ett ramverk för datahantering innebär att man fastställer interna policyer för hur data ska samlas in, lagras och bearbetas och att man säkerställer att användardata är ändamålsenliga. Grundläggande delar av ett sådant ramverk förklarades i Avsnitt 7.3.

7.4.2.3 Inrätta ett system för hantering av datakvalitet

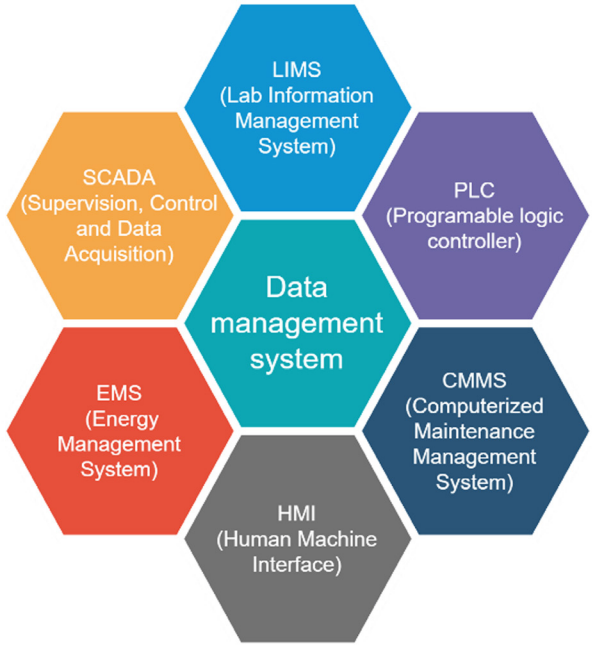
Oberoende av instrumenteringsnivå är givardata endast användbara om datakvaliteten är tillräckligt bra för den avsedda tillämpningen. Många reningsverk värdesätter högkvalitativa data och inser vikten av metadata, men faktum är att många reningsverk också kämpar med att säkerställa kvaliteten på de data som produceras. Det är viktigt att fastställa metoder för att bedöma och säkerställa datakvaliteten såsom:

- Standardiserade valideringsrutiner enligt beskrivningen i Kapitel 5, med lämpliga underhållsrutiner för att validera givarens tillstånd, med en tydlig definition av roller och ansvarsområden för olika arbetsgrupper. Korrekta valideringsrutiner och korrekt tillämpning av dessa kan kräva särskild utbildning.
- Rutiner för bedömning av datakvalitet enligt beskrivningen i Kapitel 6. Flera metoder finns tillgängliga, allt ifrån grundläggande beräkningar till mer komplicerad metoder. Implementering av sådana metoder kan också kräva särskild utbildning, beroende på den valda algoritmen.
- Samla in all data och metadata som producerats i de två punkterna ovan från givarvalideringen och datakvalitetsbedömningen, på ett enkelt sätt så att det kan delas inom hela organisationen (se Kapitel 3 och 4).

7.4.2.4 Dataintegration och implementering av ett datahanteringssystem

Att integrera olika datakällor är ett viktigt steg mot att förbättra användningen av data och metadata (se Kapitel 4). Följande aspekter bör inkluderas:

- Ett centraliserat datahanteringssystem gör det möjligt att analysera data på ett holistiskt och effektivt sätt för att följa upp den dagliga driften, diagnostisera störningar i tid och öka medvetenheten om datakvaliteten.



Figur 7.3 Integrering av datakällor i ett datahanteringssystem på ett reningsverk.

- Förutom givardata måste metadata om givarens validering och datakvalitet integreras i ett centralt datasystem. Exempelvis kan integrerade data från laboratoriemätningar (LIMS), underhållsåtgärder (CMMS) och givarsignaler (SCADA/PLC) underlätta givarvalideringen (t.ex. jämförelse med laboratoriemätningar) och optimera kalibreringsrutinerna.
- Dataintegration är också grunden för datavisualisering och dataanalys, även om det är en av de största utmaningarna i praktiken som är en flaskhals för att nyttiggöra metadata.
- Ur implementeringsperspektiv så finns det en rad olika metoder för att integrera datakällor (Kapitel 4), t.ex. datalager, datasjöar eller dataintegrationer (dvs. data förvaras i distribuerade system men kan hämtas och kopieras till andra system vid behov).
- Oavsett vilken integrationsarkitektur som väljs så är det centralt att integrera de mest kritiska datakällorna för att nyttiggöra metadata, vilket illustreras i Figur 7.3.

7.4.2.5 Smarta beslutsstöd

När ett datahanteringssystem har införts kan reningsverkets data och metadata användas till alla möjliga olika digitala lösningar (t.ex. molnberäkningar, IoT, AI etc.), vilket kan generera mervärde för befintliga data. Hur framgångsrikt detta blir beror på instrumentnivån och den digitala mognaden vid varje anläggning, överkomliga investeringar och ambitionen att använda digitala verktyg i den dagliga verksamheten.

7.4.2.6 Fortsatt digitalisering

Den digitala transformationen är en kontinuerlig resa där processer, teknik och affärsmodeller ständigt utvecklas mot ett *digitalt reningsverk*. Ett tydligt exempel är utvecklingen av digitala tvillingar (digital representation av en fysisk enhet). För att detta ska lyckas måste man skapa en data- och metadatakultur där värdet av data kommuniceras och där cheferna stöder arbetet. Denna datakultur

inom en organisation är en kritisk framgångsfaktor för att fördelarna med de tekniska idéer som delas i denna rapport ska kunna förverkligas fullt ut.

7.5 FRAMTIDSUTBLICK

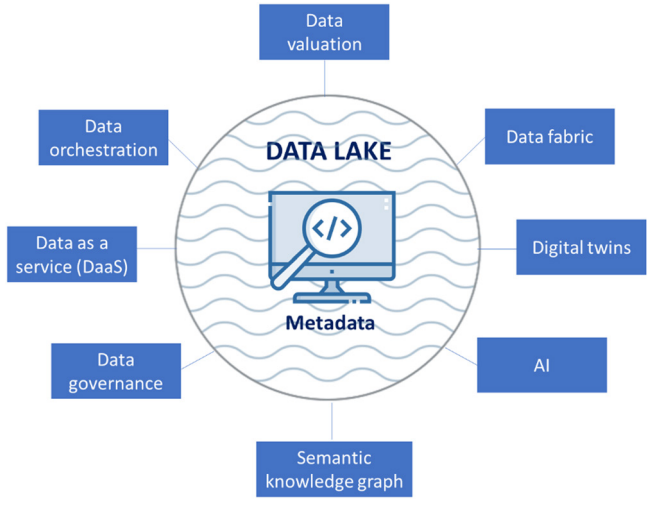
Med tanke på det explosionsartade intresset för digital teknik och den ökande produktionen av data blir generering och användning av metadata än mer viktigt för en framgångsrik digital utveckling av vattenbranschen.

I de föregående avsnitten har vi diskuterat de utmaningar som identifierats i samband med införandet av data- och metadatapraxis, och vi har också redogjort för de viktigaste åtgärder som kan underlätta införandet på reningsverk. Men detta är bara början. Potentialen med metadata är stor och det är uppenbart att metadata i framtiden kommer att underlätta att hitta, identifiera, välja, erhålla och utforska information på olika nivåer.

När man funderar på metadata och dess roll i framtiden för metadata kan man ställa sig frågorna:

- Vilken typ av ny metadata kommer att behövas?
- Hur kommer metadata att delas mellan olika system?
- Kommer metadata att standardiseras?
- Hur kommer metadata att hanteras?
- Hur kan generering och insamling av metadata automatiseras?
- Vilka förändringar i paradigmskiftet för datahantering kommer att krävas?
- Hur kommer metadata att värderas?
- Vem i organisationen kommer att behöva komma åt och hantera metadata?

Framtida insatser bör fokuseras på att skapa korrekta, kompletta, konsekventa, länkade, användarvänliga och högkvalitativa metadata som underlättar interoperabilitet mellan olika grupper och enheter på reningsverket, olika reningsverk och VA-bolag samt andra intressenter. Det är fortfarande en lång väg dit, men det finns tydligt identifierade områden som inom en snar framtid både kommer behöva metadata och generera metadata inom kort, vilket illustreras i Figur 7.4:



Figur 7.4 Framtida områden för VA-branschen där metadata kommer att spela en stor roll.

- Strukturering av data (*data orchestrion*): Att på ett automatiserat sätt sammanföra data från olika källor och lagringsplatser och kombinera dem på ett rationellt sätt kommer endast att vara möjligt med hjälp av en metadatadriven strukturering, så att data kan göras tillgängliga för framtida analys på ett organiserat och konsekvent sätt och undvika stuprörshanering av metadata.
- Värdering av data (*data valuation*): Att värdera data och metadata som en tillgång för verksamheten kommer att vara avgörande för att genomföra den digitala transformationen och bli en datadriven organisation. Hur man nyttiggör data på bästa sätt kan bedömas genom att värdera data utifrån det värde de skapar.
- Dataväv (*data fabric*): I takt med utvecklingen av s.k. *data fabrics* (som i korthet är en sammanlänkning av olika datakällor, system och sätt att hantera data på) så kommer metadata att förbättra förutsättningarna för dataanalys, datahantering, datasökbarheten och datakvaliteten (både översikt och förbättring av noggrannhet och tillförlitligheten hos data).
- Data fabrics kommer att bygga på en metadatadriven arkitektur för att leverera integrerade och förfinade data till olika dataanvändare.
- Digitala tvillingar: Ökningen av digitala tvillingar som digitala motsvarigheter till fysiska enheter och processer (anläggningar, givare, processer, komponenter) stöds av de data och metadata som genereras och associeras med den specifika enheten. Metadata som en del av digitala tvillingar inkluderar beskrivande information (t.ex. vem som skapade modellen, hur den ska hanteras) och teknisk information (t.ex. krav på datatillgänglighet och format, vilka parametrar som används, hur modellen kan användas). Sådana digitala versioner kommer inom kort att användas i allt större utsträckning för att testa driftstrategier, och för att förbättra reningsprocesserna.
- Datahantering (*data governance*): Vi befinner oss mitt i ett paradigmskifte där behov och dataanvändning drivs nedifrån och upp och inte uppifrån och ned. Denna användardrivna datahantering kommer att kräva en helt ny typ av plattform för metadatahantering där datatillgångar hanteras på ett mer holistiskt sätt.
- Semantiska kunskapsmodeller (*semantic knowledge graphs*): Kunskapsmodeller (en samling av beskrivningar av funktioner och begrepp för ett visst sammanhang eller en specifik kunskapsdomän) gör det enkelt att integrera heterogena data från olika källor genom att använda semantiska metadata för att beskriva dessa data. Den resulterande kunskapsgrafen ger ett ramverk för dataintegration, gemensam tolkning, analys och delning.
- AI: Det är tydligt att AI och maskininlärning (ML) kommer att förändra VA-sektorns framtid. Bra exempel är generering av insikter och prognoser för viktiga parametrar. I likhet med digitala tvillingar genererar varje steg i utvecklingen och implementeringen av en ML-baserad lösning metadata (vilka dataset som används, experimentkörningar, parametervärde vid modellträning, modellversion, modellparametrar, prestandamått etc.). För att sådana lösningar ska kunna användas på ett framgångsrikt sätt krävs att det alltid finns ett uppdaterat, komplett, tillförlitligt, rikt och välhanterat metadataskikt.
- Data som tjänst (*data-as-a-service*): En ny affärsmodell där dataåtkomst, dataintegration, databehandling och/eller dataanalys tillhandahålls på begäran via en molnanslutning. Data/applikationer levereras till slutanvändarna via nätverket i stället för att de körs lokalt. Metadata som tjänst skulle också göra det möjligt för användare att söka efter användbara dataset (rådata, bearbetade data etc.).

VA-branschen befinner sig för närvarande i en brytpunkt när det gäller metadatahantering. Metadata kommer att driva hela datamiljön, vilket möjliggör en betydande och hållbar omvandling på lång sikt. God metadatapraxis kommer att vara avgörande på den här resan och omvandla framtida data (och nuvarande *data grave yards* och datastuprör) till värdefull information för att bättre utnyttja de tillgängliga tillgångarna på reningsverket. Detta kommer att förenklas med ett strukturerat och

integrerat datahanteringssystem, som omfattar insamling och organisering av metadata genom automatisk loggning och datadrivna analysmetoder.

REFERENSE

Water Research Foundation (2022). Definition, Framework, and Maturity Assessment for Intelligent Water Systems: A Review of Existing Frameworks and Maturity Models.

Ordlista

Annotering av givardata	*Definition D3.32:* **Annotering av givardata** innebär att addera ytterligare information till dataposter eller tidsstämplar för att underlätta framtida användning och tolkning av givardata.
Automatiserad annotering	*Definition D3.43:* En **automatiserad annotering** är en annotering som har tilldelats genom dataanalys eller algoritmiska verktyg utan någon mänsklig inblandning.
Avvikande data	*Definition D3.28:* En avvikelse är en del av en datamängd som inte stämmer överens med de andra delarna av datamängden.
Bias	*Definition D3.20:* **Bias** är den förväntade skillnaden mellan en uppsättning datapunkter och det accepterade referensvärdet (ISO, 2006).
Chock	*Definition D3.26:* En stegvis förändring eller **chock** är en drastisk och tillfällig förändring av givarens bias från en tidsstämpel till nästa.
Dataflöde	*Definition D4.9*: Ett **dataflöde** (*data pipeline*) är en sekvens av processer och verktyg som används för att överföra och omvandla data mellan ett eller flera system och ett målarkiv. Processen bidrar till att strukturera data för att underlätta kommande analyser.
Datahantering	*Definition D4.2*: **Datahantering** (*data management*) består av metoder, arkitektur och verktyg för att få tillgång till och kunna hantera leverans av data på ett konsekvent sätt med olika datatyper i en organisation. Syftet med datahantering är att uppfylla organisationens alla datakrav för tillämpningar, applikationer och processer.
Datahanteringsstrategi	*Definition D4.4*: En tillämpad **datahanteringsstrategi** (*data governance*) innebär att organisationen har en verksamhetsstyrning för att hantera data (planering, övervakning och tillämpning).
Datahanteringsteknik	*Definition D4.1*: **Datahanteringsteknik** (*data management technology*) avser en kombination av hårdvara och mjukvara som används för systematisk insamling, beskrivning, utforskning, modifiering, lagring och hämtning av givarsignaler och deras metadata.
Datakomprimering	*Definition D4.6*: **Datakomprimering** är processen för att koda information med hjälp av färre bitar än den ursprungliga representationen. Detta minskar den volym som krävs på den fysiska lagringsenheten.
Datakvalitetsstämpel	*Definition D2.4:* En **datakvalitetsstämpel** är ett enskilt tal eller en text som produceras av ett instrument och som anger en nivå av förtroende för tidsstämpeln och/eller datapunkten.

Datapost	*Definition D2.2:* En **datapost** (*data record*) är ett par eller en triplett av (i) en tidsstämpel (obligatorisk), (ii) en datapunkt (obligatorisk) och (iii) en datakvalitetsstämpel (*qualifier*) (valfri).
Dataproveniens	*Definition D4.10*: **Dataproveniens** (*data lineage*) avser metadata som beskriver de processer som påverkat den aktuella datamängden och därigenom ger en historisk bild av data och deras ursprung. Detta inkluderar ursprung, modifiering och lagring. Detta leder till en större transparens som även underlättar felsökning vid dataanalysen.
Datapunkt	*Definition D2.3:* En **datapunkt** består av ett eller flera värden som uppfyller följande tre krav: (1) Alla värden inom datapunkten har producerats av samma instrument. (2) Alla värden är kopplade till samma tidpunkt eller, i förekommande fall, samma tidpunkt för insamling av vattenprover. (3) Alla värden beskriver samma typ av variabel och produceras med samma mätprincip.
Datorstödd annotering	*Definition D3.42:* En **datorstödd annotering** är en annotering som tilldelas efter en inspektion av en mänsklig expert av dataposter som valts ut för inspektion genom dataanalys eller algoritmiska verktyg.
Direkt annotering	*Definition D3.39:* En **direkt annotering** är en annotering som endast baseras på inspektion eller analys av de dataposter som annoteras.
Direkt mätning	*Definition D5.11:* En **direkt mätning** är när givaren direkt mäter den komponent som är av intresse.
Drift	*Definition D3.27:* **Drift** är en långsam förändring av bias, varians eller responstid. Drift är ett symtom som utvecklas gradvis och kontinuerligt över tid.

Etikett	*Definition D3.33:* En **etikett** eller flagga (*flag*) är en strukturerad och förutbestämd typ av annotering med följande egenskaper:
	• Den används upprepade gånger för att kommentera liknande händelser eller företeelser
	• Den innehåller minst två komponenter:
	  ◦ En unik identifierare (alfanumerisk sträng) som beskriver typen av händelse eller företeelse.
	  ◦ En beskrivning av en typ av åtgärd, kausal händelse eller symtom.
	• Dessutom kan etiketten innehålla följande:
	  ◦ Beskrivning av det specifika instrumentet.
	  ◦ Tidpunkt eller tidsfönster som etiketten avser.
	  ◦ Namn på den person skapat annoteringen.
	  ◦ Tidpunkt då etiketten skapades.
	  ◦ Syftet med annoteringen.
	Programvara och version som använts för anteckningar.

Expertannotering	*Definition D3.41:* En **expertannotering** är en anteckning som tilldelas efter att en mänsklig expert har granskat dataposter utan hjälp av dataanalys eller algoritmiska verktyg.
Fel	*Definition D3.17:* Ett **fel** är en avvikelse mellan ett uppmätt eller beräknat värde och det sanna, specificerade eller teoretiskt korrekta värdet.
Förändring av enheter	*Definition D3.15:* En **ändring av enheter** är en avsiktlig ändring av enheter från den måttenhet som tillhandahålls av instrumentet till en enhet som anses lämplig för långtidsförvaring och användning.
Förlorade data	*Definition D2.6:* En datapost, tidsstämpel eller datapunkt anses vara **förlorad** (*missing data*) om ett instrument installerats för att samla in data, men misslyckats med att producera förväntade mätvärden.

Förlustfri komprimering	*Definition D4.7*: **Förlustfri komprimering** (*lossless compression*) är datakomprimering som säkerställer att originaldata kan rekonstrueras perfekt från de komprimerade data, det vill säga lagring utan informationsförlust. Komprimering uppnås genom att eliminera redundanta delar i data.
Frånvarande data	*Definition D2.5:* En datapost, tidsstämpel eller datapunkt anses vara **frånvarande** (*absent data*) om tillgång till dessa data är värdefull vid tidpunkten för dataanalysen, men inget instrument har installerats för att samla dessa data.
Givarplacering	*Definition D3.4:* **Givarens placering** är den fysiska positionen av givaren (givarens mätelement).
Givarunderhåll	*Definition D5.3:* **Givarunderhåll** innefattar alla de åtgärder som krävs för att bibehålla givaren i full funktion under sin livslängd.
Givarvalidering	*Definition D5.2:* **Givarvalidering** (*Quality assurance, QA*) inkluderar alla åtgärder som utförs för att verifiera och dokumentera att en givare har full funktion och producerar data med önskad noggrannhet.
Hysteres	*Definition D5.7:* **Hysteres** är ett fenomen där givarens mätvärde beror på riktningen för förändringen av mätvärdet.
Indirekt annotering	*Definition D3.40:* En **indirekt annotering** är en annotering som baseras på information som finns utanför den datamängd som annoteras.
Information från tillverkaren	*Definition D3.12:* **Tillverkarinformation** består av instrumentets märke eller fabrikat, modellnummer och serienummer.
Instrumentets historik	*Definition D3.14:* **Instrumentets historik** består av all dokumentation som beskriver viktiga händelser i den signalproducerande enhetens historia.
Kalibrering	*Definition D5.10:* **Kalibrering** innebär att justera en givare för att uppnå önskad noggrannhet genom att matcha givarens utdata med ett eller flera accepterade referensvärden.
Kollektiv annotering	*Definition D3.36:* En **kollektiv annotering** som används för flera dataposter på en gång. De annoterade dataposterna kan ha samma eller olika tidsstämplar och kan tillhöra endast en signal eller flera signaler. Kollektiva annoteringar används vanligen för händelser som börjar och slutar vid olika tidpunkter.
Kommentar	*Definition D3.34:* En **kommentar** (*comment*) är ett stycke fritext som används för annotering. En enda kommentar kan kopplas till ett valfritt antal dataposter, ett valfritt antal tidsstämplar eller ett valfritt antal tidsspann.
Komprimering med förlust	*Definition D4.8*: **Datakomprimering med förlust** (*lossy compression*) är datakomprimering som innebär att viss information går förlorad. Därför kan man bara konstruera en approximation av originaldata baserat från komprimerade data. Även om information går förlorad i processen kan höga komprimeringsnivåer ofta uppnås med små informationsförluster.
Laboratoriedata	*Definition D2.7:* **Laboratoriedata** (även: laboratoriemätningar) består av värden som (a) registreras genom en sekvens av steg som omfattar minst ett manuellt steg för beredning av vattenprover och (b) beskriver observerade värden för egenskaper och storheter i ett fysiskt system av intresse.
Linjäritet	*Definition D5.6:* En givares **linjäritet** (*linearity*) beskriver i vilken utsträckning en förändring i den sanna halten/variabeln (indata) återspeglas som en linjär förändring i själva mätningen (utdata) över hela dess mätområde.
Lutning/känslighet	*Definition D5.5:* En givares **lutning/känslighet** (*slope/sensitivity*) är den förändring av den uppmätta variabeln (utdata) som fås för en standardiserad förändring av de sanna variabeln/halten (indata).
Mätenhet	*Definition D3.5:* **Mätenheten** är det valda standardmåttet med vilket storleken på givaravläsningen uttrycks.

Mätområde	*Definition D3.6:* **Mätområdet** är det intervall som definieras av de minimi- och maximivärden som kan produceras av en givare som är korrekt konfigurerad och används på rätt sätt.
Mätprincip	*Definition D3.9:* En givares **mätprincip** är en beskrivning av det grundläggande fysikalisk-kemiska förhållandet mellan den intressanta variabeln och den råsignal som givaren producerar.
Metadata	*Definition D2.8:* **Metadata** definieras som den information som används för att beskriva signaldata eller laboratoriedata.
Modifiering av signal med informationsförlust	*Definition D3.16:* En **modifiering av en signal som innebär förlust av information** är en procedur som omvandlar originalsignalen till en derivatsignal så att originalsignalen inte kan rekonstrueras perfekt från derivatsignalen, dvs. viss information går förlorad i processen.
Noggrannhet	*Definition D3.18:* **Noggrannheten** beskriver hur god överensstämmelsen är mellan ett mätning och det accepterade referensvärdet (ISO, 2006).
Numerisk upplösning	*Definition D3.3:* Den **numeriska upplösningen** är den minsta ökning av den uppmätta variabeln som kan registreras digitalt (dvs. analog-till-digital-omvandling). Inom signalbehandlingsområdet benämns detta som kvantisering.
Offset	*Definition D5.4:* En givares **offset** (*offset*) är en konstant avvikelse från det accepterade referensvärdet.
Orsaksannotering	*Definition D3.38:* En **orsaksannotering** är en annotering som innehåller information om grundorsaken till det observerade symptomet.
Ostrukturerade data	*Definition D2.10:* Data sägs vara **ostrukturerade** så snart ett av dessa villkor är uppfyllt: (1) Antalet dimensioner för datapunkterna kan variera. (2) Antalet värden längs en eller flera dimensioner kan variera. (3) Ett värdes position längs en eller flera dimensioner kan ändras utan att det får konsekvenser för den fortsatta analysen. (4) Datapunktsvärden kan tillhöra flera olika taltyper, t.ex. binärt, heltal eller decimaltal.
Policy för datalagring	*Definition D4.5*: En **policy för datalagring** beskriver vilka data som ska lagras eller arkiveras, var det ska ske och hur länge de ska lagras innan de ska rensas bort.
Precision	*Definition D3.21:* **Precision** beskriver hur nära enskilda datapunkter ligger varandra.
Primärdata	*Definition D2.12:* **Primärdata** (*primary data*) är data som består av den mest exakta, omfattande och uppdaterade versionen av en given uppsättning data. Denna datamängd används som den *sanna datakällan*, det vill säga en datamängd som varje dataanvändare accepterar som den bästa tillgängliga informationskällan.
Produktspecifikation	*Definition D3.11:* **Produktspecifikationer** och guider består av dokumentation som beskriver installation, användning och underhåll av ett instrument.
Programvara, versionshantering	*Definition D4.3:* **Versionshantering** av mjukvara (*software versioning*) avser processen att dokumentera stegvisa förändringar i programvara och tilldela unika versionsnamn eller unika versionsnummer för att referera till unika, historiska tillstånd av programvaran. Målen med versionshantering inkluderar (a) effektivt samarbete mellan programmerare, (b) hantera kompatibilitet med annan programvara, och (c) garantera reproducerbarhet av resultat som erhållits med programvaran.
Punktannotering	*Definition D3.35:* En **punktannotering** är en annotering som enbart nyttjas till en enstaka datapost.

Rådata	*Definition D2.11:* **Rådata** är data som produceras av en signalkälla, t.ex. en givare.
Repeterbarhet	*Definition D3.23:* **Repeterbarhet** är precisionen när mätproceduren (vanligtvis givarvalideringen) upprepas under exakt identiska förhållanden. Det vill säga att oberoende testresultat erhålls med samma metod i identiska tester i samma laboratorium, av samma operatör, med samma utrustning, inom ett kort tidsintervall.
Reproducerbarhet	*Definition D3.24:* **Reproducerbarhet** är den precision som erhålls från testresultat där samma metod används på identiska testobjekt med olika laboratorier, operatörer och utrustning (ISO, 2006).
Responstid	*Definition D3.25:* **Responstiden** (*response time*) är den tid det tar för ett instrument att visa ett nytt värde när det utsätts för en stegvis förändring. Den kan också definieras som den tid som krävs för att uppnå en procentandel av en stegvis förändring av förändringen av det sanna värdet (typiskt 90 % eller 95 %).
Riktighet	*Definition D3.19:* **Riktighet** är överensstämmelsen mellan det medelvärde som erhålls från en stor uppsättning datapunkter och ett accepterat referensvärde (ISO, 2006).
Sekundärdata	*Definition D2.13:* **Sekundärdata** är data som har genererats genom urval och modifiering av primärdata för ett specifikt sammanhang eller syfte.
Sensor	*Definition D5.1:* En **givare** är en enhet som kommunicerar värdet på en processvariabel (definition utifrån Asociación Española de Normalización, 2004).
Signaldata	*Definition D2.1:* **Signaldata** (även: signaler) består av värden som produceras av en givare för kvaliteter och kvantiteter i ett fysiskt system av intresse. Signaldata består av serier av dataposter.
Signalöverföring	*Definition D3.8:* **Signalöverföring**ssättet är det sätt på vilket en signal som produceras av en givare överförs till den fysiska lagringsenheten (t.ex. PLC, SCADA).
Signalverifiering	*Definition D5.9:* **Signalverifiering** (*end-to-end-testing*) avser ett systematiskt förfarande för att verifiera att givarsignalen förblir oförändrad mellan givaren och platsen för själva datalagring, t.ex. en historikdatabas.
Slumpmässigt fel	*Definition D3.31:* Ett **slumpmässigt fel** (*random error*) orsakas av oförutsägbara men ihållande variationer som påverkar alla mätningar av samma givare på samma eller liknande sätt.
Spårbara data	*Definition D2.14:* **Data är spårbara** om det finns system som möjliggör övervakning av hela processen för insamling, lagring och bearbetning av data för en korrekt och effektiv användning.
Strukturerad data	*Definition D2.9:* En signal sägs innehålla **strukturerade data** när varje datapunkt i varje datapost uppfyller följande krav: (1) Antalet dimensioner är konstant. (2) Antalet värden i alla dimensioner är konstant. (3) Ett värdes position längs varje dimension är meningsfull och dess betydelse förändras inte över tiden. (4) Alla datapunktsvärden är begränsade till att tillhöra en enda taltyp, t.ex. binär (*boolean*), heltal (*integer*) eller decimaltal (*float*).
Syfte med mätningen	*Definition D3.10:* **Syftet med mätningen** beskriver hur signalen är tänkt att användas.
Symtomannotering	*Definition D3.37:* En **symtomannotering** är en annotering som beskriver observerade symtom men som inte ger någon kausal orsaksförklaring.

Systematiska fel	*Definition D3.29:* Ett **systematiskt fel** är beständigt i tiden och kommer sannolikt att ge upphov till ett mätfel som är konstant eller förändras långsamt över tid.
Tidsstämpelns ursprung	*Definition D3.1:* **Tidsstämpelns ursprung** är den tekniska enhet som producerar tidsstämpeln i en datapost.
Tidsupplösning	*Definition D3.2:* **Tidsupplösningen** är tidsintervallet mellan efterföljande dataposter som produceras av en givare.
Tillfälligt mätfel	*Definition D3.30:* Ett **tillfälligt mätfel** (*gross error*) orsakas av ett fenomen som inträffar sällan och under korta perioder och som ofta bara påverkar en enda datapunkt.
Unik identifierare	*Definition D3.13:* En **unik identifierare** (*unique identifer*) är ett namn eller nummer som är unikt för ett visst instrument eller en viss komponent i ett instrument, vilket möjliggör en detaljerad beskrivning av signalkällan.
Validering, givarvalidering	*Definition D5.8:* **Givarvalidering** eller givarverifiering avser ett systematiskt förfarande för att (a) samla in och registrera mätvärden under kontrollerade förhållanden för att kvantifiera datakvaliteten genom kvalitetsmått som noggrannhet, bias, precision, offset, lutning och/eller linjäritet och (b) för att utvärdera om kvalitetsmåtten uppfyller förutbestämda specifikationer för den specifika tillämpningen. Den enklaste formen av validering är att jämföra givarens mätvärde med en referensmätning.
Variabelområde	*Definition D3.7*: **Variabelområdet** är det intervall som definieras av det minimi- och maximivärde som är acceptabelt eller typiskt observerat under normal drift, antingen i allmänhet eller i ett visst sammanhang (t.ex. plats och/eller tid).
Varians	*Definition D3.22:* Mätfelets **varians** är det förväntade värdet för den kvadratiska avvikelsen mellan en enskild mätning och det förväntade värdet (t.ex. medelvärdet) för mätningen.
Verifiering	Se: validering

Scientific and Technical Report Series No. 31a

Bästa möjliga data från givare i avloppsreningsprocesser

Redigerad av Kris Villez, Daniel Aguado, Janelcy Alferes, Queralt Plana, Maria Victoria Ruano och Oscar Samuelsson

Utvecklingen med instrumentering på avloppsreningsverk har revolutionerat möjligheterna för processövervakning och reglering. Effektiv dataöverföring och kraftigt reducerade kostnader för datalagring har inneburit att reningsverk i dag förfogar över stora mängder data – eran med Big data har nått VA-branschen. Samtidigt har de senaste årens utveckling inom AI och maskininlärning gjort det möjligt att utvinna värdefull information ur dessa stora datamängder.

Men trots digitaliseringens möjligheter så krävs både förmågan att hantera stora datamängder och tillräcklig datakvalitet för att fullt ut nyttiggöra AI och datadriven automatisering på ett effektivt och framgångsrikt sätt. Dokumentering och lagring av metadata, exempelvis karakterisering av datakvaliteten, är ofta en bristvara. Det leder till att stora delar av insamlade data riskerar att gå till spillo eftersom de med tiden blir svårare och svårare att tolka och därmed snabbt blir utdaterade. Den krassa verkligheten är att datadrivna beslutsstöd med AI och maskininlärning aldrig kommer att realiseras om reningsverkets organisation saknar kunskap och förmåga att hantera de fyra V som Big data innebär: datamängd (*Volume*), hastighet (*Velocity*), variation (*Variability*) samt noggrannhet och tillförlitlighet (*Veracity*).

Rapporten *Bästa Möjliga Data från Givare i Avloppsreningsprocesser* är en svensk översättning av den engelska rapporten *Metadata Collection and Organization in Wastewater Treatment and Wastewater Resource Recovery Systems*. Rapporten är en guide till att hantera dagens datatekniska utmaningar och syftar till att beskriva vad metadata innebär och hur de rent praktiskt kan användas för att nyttiggöra reningsverksdata. Rapporten ger råd om hur metadata bör samlas in, tolkas, struktureras och lagras utifrån befintliga standarder och de krav som datadrivna algoritmer ställer på data. Den är en bra utgångspunkt för läsare som vill förbättra sina datahanteringsrutiner och förstå hur de kan utveckla metoder för datadrivna beslutsstöd. Den riktar sig till en bred målgrupp såsom instrumenttekniker, driftpersonal, driftchefer och allmänt dataintresserade personer på reningsverk.

iwaponline.com

@IWAPublishing

ISBN: 9781789065107 (eBook)

ISBN: 9781789065114 (ePub)