

METODER FÖR DATADRIVEN DAMMÖVERVAKNING

RAPPORT 2022:855



METODER FÖR DATADRIVEN DAMMÖVERVAKNING

ANTON JACOBSON, HÅKAN FRIDÉN OCH ANDERS BJÖRK

ISBN 978-91-7673-855-9 | © Energiforsk maj 2022

Energiforsk AB | Telefon: 08-677 25 30 | E-post: kontakt@energiforsk.se | www.energiforsk.se

Förord

Detta arbete behandlar utveckling och implementering av datadrivna metoder för att fånga upp avvikande värden från givare monterade på eller intill dammar samt nya sätt att visualisera och fånga avvikelser från normalt tillstånd för en kraftverksdamm.

Vi presenterar en kort introduktion till dammsäkerhetsområdet, syfte för arbetet, beskrivning av data, framtagning av metoder för detektion av givar- och kommunikationsfel och dess resultat, samt förbehandling av data för modellering. Vi tillämpar multivariat dataanalys för att både fånga upp avvikande tillstånd hos enskilda givare och för dammtillståndet.

Rapporten innehåller beskrivning av använda teknologier för implementation, och en beskrivning om hur denna implementation gjorts, både off-line och on-line. Rapporten avslutas med slutsatser.

Rapporten har skrivits av:

Anton Jacobson, Håkan Fridén och Anders Björk på IVL Svenska Miljöinstitutet
Projektledare har varit:

Anders Björk på IVL Svenska Miljöinstitutet

Projektet har haft en referensgrupp bestående av:

Uno Kuoljok (Uniper), Kerim Genel Waldenström (Vattenfall Vattenkraft AB),
Sezar Moustafa Näsvall (Fortum), Romanas Wolfsborg (Vattenfall Vattenkraft AB)
Gustav Enmark (Fortum) och Andreas Fridh (Statkraft)

Projektet "Optimerade och anpassade datadrivna metoder för dammsäkerhet" har finansierat av Energiforsks program för Dammsäkerhet och Stiftelsen Institutet för Vatten- och Luftvårdsforskning.

I denna publikation presenteras resultat och slutsatser från projektet, den publiceras både i Energiforsks rapportserie och i IVL:s rapportserie B. För innehållet ansvarar IVL. Rapporten har granskats och godkänts i enlighet med IVL:s ledningssystem.

Sammanfattning

Dammövervakning är ett stort och viktigt område för svenska dammägare. Det är en utmaning att kontinuerligt övervaka en dammkonstruktion med dess givare och snabbt kunna reagera på förändringar som visar på en avvikelse i dammen. Denna rapport behandlar detektion och identifiering av givar- och kommunikationsfel vid dammövervakning samt tillämpning av multivariata metoder för att visualisera och fånga avvikande tillstånd för en damm.

Rapporten är resultatet av ett forskningsprojekt som drivits av IVL Svenska Miljöinstitutet AB under 2021–2022. Projektet finansierades av Energiforsk och Stiftelsen Institutet för Vatten- och Luftvårdsforskning (SIVL) och hade som mål att implementera övervakning av multivariata metoder för detektion av avvikelser i givarsignaler och dammtillstånd. Detta mål uppnåddes, genom flertalet steg rörande databearbetning, signalövervakning, maskininlärning och metodutveckling.

Multivariata modeller utvecklades baserat på etablerade metoder inom maskininlärning. Modellerna syftar till övervakning av givarsignaler samt det övergripande dammtillståndet. Flera olika typer av multivariata metoder har använts inom projektet.

Projektet har två fallstudier och genom anläggningsägare har projektet fått tillgång till mer än 5 års historiska data för respektive damm. Denna data ligger till grund för utveckling och utvärdering av multivariata modeller. Att samla in, bygga upp en gemensam förståelse av och förbehandla dessa data kräver mycket tid och ett gott samarbete mellan dataspecialister och dammägare. Förbehandling, där kända avvikelser utesluts, är ett nödvändigt steg för att kunna utveckla en övervakning baserat på multivariata modeller. De datadrivna analysmetoderna utvecklades till början off-line för att sedan användas on-line.

Tillsammans med dammägare för fallstudierna har pilotinstallationer av modeller samt övervakning av dessa driftsatts. På grund av de krav som säkerhetslagstiftningen ställer kan IVL endast få ett begränsat interface till realtidsdata. Genom implementationer kan dock resultaten analyseras och itereras för att vidareutveckla metoderna för att öka dess nytta.

Projektet har tagit fram multivariata modeller samt tillvägagångssätt för utveckling av dessa modeller. Dessa metoder kan ligga till grund för fortsatt utveckling av kostnadseffektiv multivariat övervakning av dammar samt en kurs för dammägare och dammingenjörer med syfte att öka kompetensen inom området som således kan öka nyttan med övervakning av givarsignaler.

Summary

Dam safety monitoring is a large and important area for Swedish dam owners. It is a challenge to continuously monitor a dam structure with its sensors and be able to quickly react to changes that indicate an anomaly in the dam. This report deals with the detection and identification of sensor and communication errors in dam monitoring as well as the application of multivariate methods for visualizing and capturing anomalous states of a dam.

The report is the result of a research project run by IVL Swedish Environmental Research Institute AB during 2021–2022. The project was funded by Energiforsk and the Foundation for Vatten- och Luftvårdsforskning (SIVL) and aimed to implement monitoring of multivariate methods for the detection of deviations in sensor signals and dam state. This goal was achieved through several steps regarding data processing, signal monitoring, machine learning and method development.

Multivariate models were developed based on established methods in machine learning. The models aim to monitor sensor signals as well as the overarching dam state. Several different types of multivariate methods have been used within the project.

The project has two case studies and through dam owners, the project has gained access to more than 5 years of historical data for each dam. This data forms the basis for the development and evaluation of multivariate models.

Collecting, building a common understanding of and pre-processing this data requires a lot of time and good collaboration between data analysis specialists and dam owners. Pretreatment, in which known deviation is excluded, is a necessary step to be able to develop a monitoring based on multivariate models. The data-driven analysis methods were initially developed off-line and then used online.

Together with dam owners for the case studies, pilot installations of models and monitoring of these have been commissioned. Due to the requirements set by security legislation, IVL can only have a limited interface to real-time data. However, through implementations, the results can be analyzed and iterated to further develop the methods to increase its use.

The project has developed multivariate models and approaches for the development of these models. These methods can form the basis for the continued development of cost-effective multivariate monitoring of dams and a course for dam owners and dam engineers with the aim of increasing competence in the field, which can thus increase the benefits of monitoring sensor signals.

Nyckelord

Datadriven
Dammsäkerhet
Feldetektion
Sensordata
Signalövervakning
Processövervakning
Dataanalys
Maskininlärning

Innehåll

1	Inledning	9
2	Bakgrund	12
3	Data	14
4	Detektion av givar- och kommunikationsfel	19
4.1	Metoder	19
4.2	Resultat från fallstudie II	19
4.3	Anpassningen av gränsvärden för detektion av givare och kommunikationsfel hos fallstudie I	24
5	Arbetsgång för modellering	26
6	Förbehandling av data för modellering	27
7	Multivariata metoder - Teori	31
7.1	Principalkomponentanalys, PCA	32
7.2	Regressionsanalys	33
7.2.1	Multipel linjär regression, MLR	33
7.2.2	Projection to Latent Structures/Partial Least Squares, PLS	33
7.2.3	Hydrostatisk säsong tidsmodell, HST	34
7.2.4	Neurala nätverk, NN	34
7.3	Avvikelse från modellens giltighetsdomän	35
7.3.1	DModX	35
7.3.2	Hotellings T2	36
7.4	Konfidensintervall för predikterade givarsignaler	36
8	Multivariata metoder – Resultat	37
8.1	Övervakning av dammtillstånd	38
8.1.1	Fallstudie I	38
8.1.2	Fallstudie II	39
8.2	Övervakning av enskilda givare	40
8.2.1	Fallstudie I	41
8.2.2	Fallstudie II	44
8.2.3	Multipel linjär regression, MLR	47
8.2.4	Hydrostatisk säsong tidsmodell, HST	48
8.3	Tidsförskjutning i data	51

9	Detektion av avvikelser från normaltillstånd baserat på multivariata metoder	53
9.1	Simulering av tillståndsavvikelse	53
9.2	Resultat vid analys och användning av utvecklade metoder på simulerade data	55
9.3	Modellövervakning av dammtillstånd	57
9.3.1	Konfidensintervall för projektionsvärden (scores), statiska i två dimensioner	57
9.3.2	Konfidensintervall för projektionsvärden (scores), tidsberoende per enskilda dimensioner	62
9.3.3	Metod för att minimera falska larm	67
9.3.4	Avvikelsesummering, CuSum	67
9.4	Övergripande visualisering av enskilda givare respektive dammtillstånd	68
10	Implementation	70
10.1	Använda metoder för att realisera offline- och online- lösningen	70
10.2	Systemtekniska krav på mjukvarulösningen online	71
10.3	Övergripande beskrivning av implementerad lösning	71
10.4	Definitionsfilen ObsVarDef	73
10.5	Program	73
10.5.1	DSPopulateSignals	73
10.5.2	DSModelMaker	73
10.5.3	DSWriteData	74
10.5.4	DSWatchDog	75
10.6	Datakommunikation	78
10.6.1	OPC-kommunikation	78
10.6.2	Filkommunikation	80
10.7	Databas	82
10.7.1	ModelLimits	82
10.7.2	Models	82
10.7.3	NodeServers	83
10.7.4	Signals	83
11	Workshops	85
12	Slutsatser	86
13	Termer och förkortningar	87
14	Referenser	88

1 Inledning

En damm är ett byggnadsverk vars syfte är att lagra, kontrollera och/eller avleda vatten. En damm utgör en barriär över till exempel ett vattendrag för att dämna upp ett magasin eller skydda lägre belägna områden från översvämning. Dammar började byggas i Sverige redan under medeltiden och den kraft som kan utvinnas när vatten släpps igenom dammen har länge utnyttjats för att ersätta muskelkraft vid utförande av tunga sysslor. Dammverksamheten utvecklades sedan för att kunna producera elektricitet. Detta genom att kraften från det indämda vattnet användes för att driva turbiner som i sin tur driver generatorer som producerar elektricitet.

Under 1900-talet byggdes vattenkraften ut för elproduktion och är idag den enskilt största källan till el producerad i Sverige (tätt följd av kärnkraft). År 2019 stod vattenkraften för cirka 39.3 % och kärnkraft för cirka 39.1 % av den totala elproduktionen i Sverige (Ekonomifakta, (2020)).

De effekter som uppkommer vid en omställning till en mer klimatneutral energiproduktion kommer att påverka kraftverksägarnas regleringsstrategier. Till exempel kommer vattenkraften att användas för volatil kraftproduktion vilket medför högre inslag av korttidsreglering. Detta i sin tur kan innebära mer slitage på avbördningsanordningar men även på de dämmande konstruktionerna. Vidare kommer de pågående klimatförändringarna att påverka våra reglerade vattendrag på olika sätt genom förändrade flödesförhållanden. Det är också troligt att flera av klimatförändringens effekter ännu är okända, vilket gör det svårt att förutse om och hur dammsäkerheten påverkas.

Dammätningar, som är en del av dammens tillståndsovervakning, utförs med syfte att följa och utvärdera eventuella förändringar, både kortare förlopp och förändringar på längre sikt, samt ge underlag för en långsiktig bedömning av anläggningens tillstånd. Detta ger ett bra underlag för att kunna bedöma dammkonstruktionens behov av eventuella åtgärder. Installerad mätutrustning ska även ge indikationer för styrning av drift och driftövervakning där även varnings- och larmfunktioner ska ingå.

Vid övervakning av dammar mäts ett antal olika storheter, ofta vid flera olika geografiska positioner i dammen. Den datamängd som sparas från en damm är ofta stor och det är svårt att hantera kvalitet och kvantitet på ett effektivt sätt. I många fall finns både varnings- och larmgränser definierade för varje enskild signal i syfte att uppmärksamma operatörerna på att mätvärden som rapporteras är högre eller lägre än förväntat vid "normalt" beteende. Men vad innebär ett "normalt" beteende? Dammens tillstånd beror dels av förändringar i omgivningen, dels på åldringsprocesser i dammkonstruktionen vilket gör att det i många fall är svårt att bedöma vad som kan betraktas som normalt. Ett sätt att delvis hantera detta är att använda sig av historiska data om vattennivåer och nederbörd i kombination av domänkunskap rörande normala beteenden hos dammövervakningsgivare.

När en avvikelse väl har detekterats måste den analyseras. Avvikande mätvärden kan grovt klassificeras i två orsaksgupper.

- Avvikelser som beror på faktiska förändringar i dammens tillstånd, vilka kan bero på inre processer och yttre påverkan såsom temperatur- eller vattenståndsförändringar. Här antas att alla givarsignaler är korrekta.
- Avvikelser som beror på fel i givare eller störning av signalkedjan

Det är problematiskt när rena mätfel (exempelvis felaktigt placerade eller trasiga givare) ger upphov till avvikande förändringar i mätdata och detta i sin tur påverkar bedömningen av dammens tillstånd. Idag görs denna bedömning erfarenhetsmässigt men den ökande mängden mätdata och ändrade regleringsförhållanden medför att bedömningen allt svårare och även allt mer kritisk. Detta utgör motivationen för att ta fram automatiska metoder för att detektera givar- och kommunikationsfel.

Mycket av den vetenskapliga litteraturen inom datadriven metodik för dammsäkerhet handlar om tillämpning av maskininlärningsmetoder. De tar oftast inte hänsyn i förbehandling och antar att alla givare fungerar, vilket inte alltid är sanningen. Slutsatser från Johansson et al., (2020), belyser att "ett bra förarbete är nödvändigt om metoderna ska vara tillräckligt pålitliga för att kunna användas till dataanalys". Där beskrivs fallgropar vid dataanalys och hur dessa kan undvikas. Ofta kan orsaken till dessa fel ligga tidigt i värdekedjan. Exempel på detta är givarfel, mätfel eller fel vid datainsamling.

I Jacobson et al., (2021) rapporterades hur vi utvecklat och installerat en generell övervakningsapplikation som körs i realtid hos dammägaren med dataöverföring från och till dennes signaldatabas. Vi utvecklade metoder för fristående övervakning av enskilda signaler, univariat övervakning.

Detta projekts syfte har varit att med multivariata metoder upptäcka såväl avvikande beteende hos enskilda signaler orsakade av givarfel, såväl som avvikande trender hos dammtillståndet. Fokus har varit att bygga vidare på den univariata övervakningen genom att utveckla, testa och driftsätta metoder, där flera eller alla signaler ingår, för övervakning av de olika signalerna eller dammtillståndet, multivariat övervakning.

Projektet har arbetat med två fallstudier vid två dammanläggningar. Data från dessa anläggningar har använts för att utveckla multivariata modeller som tagits fram för att övervaka givarsignaler samt dammens tillstånd. Projektet har även fokuserat på att utveckla övervakningsmetoder baserat på dessa modeller för att på ett enklare sätt överblicka den stora mängd givardata som samlas in från dammen. Modellerna samt övervakningsmetoderna har implementerats hos de två fallstudierna. Fallstudierna har genomförts tillsammans med Vattenfall Vattenkraft AB samt Statkraft Sverige AB. Möten har skett till största del digitalt men med ett kompletterande fysiskt möte i Umeå den 3 november. Vid detta tillfälle genomfördes även ett studiebesök vid en av dammarna som använts som fallstudie i projektet.

Rapporten inleds med en bakgrund till studien och en beskrivning av de data som ligger till grund för arbetet med två fallstudier. Därefter beskrivs de metoder för

detektion av givar- och kommunikationsfel som har utvecklats i projektet.

Det närliggande problemet, förbehandling av data för multivariat modellering tas även upp. Resultat vid analys av data från de två fallstudierna används för att illustrera metodernas funktion och prestanda. Implementering i dammägarnas övervakningssystem beskrivs i ett separat kapitel. Därefter följer ett sammandrag av det som diskuterades på ett arbetsmöte som hölls inför implementation av realtidsapplikationen. Rapporten avslutas med slutsatser.

Förklaringar till förkortningar och benämningar förklaras i kapitel 13.

Vi har valt att ta med en stor del av texten från kapitel 1 i Jacobson et al., (2021), projektet som ligger till grund för detta projekt. Detta för att denna rapport skall kunna läsas separat.

2 Bakgrund

Vattenkraftverksamhet och dammsäkerhet styrs av olika lagar, förordningar, föreskrifter och riktlinjer. Här ingår bland annat *Miljöbalken (1998:808)*, (1998) med tillhörande förordningar om dammsäkerhet och verksamhetsutövarens egenkontroll. Ägaren av dammen är skyldig att följa bland annat dessa regelverk, med syfte att så långt det är rimligt undvika dammhaveri med efterföljande negativa verkningar på samhälle och miljö.

Samhällets förväntningar och allmänhetens acceptans påverkar hur vattenkraftsverksamheten bedrivs. Det är av största vikt att säkerhet prioriteras. Energiföretagen Sverige verkar för god dammsäkerhet hos medlemsföretagen genom att tillhandahålla *RIDAS - Energiföretagens riktlinjer för dammsäkerhet*, (2019). Riktlinjerna utgör grund för dammsäkerhetsarbete enligt god praxis för medlemsföretagen vad gäller arbetsmetoder och tekniklösningar.

Uttrycket dammsäkerhet innefattar riskbedömningar vid dammar, minimering av haveririsker samt minimering av de konsekvenser som blir följden om ett haveri ändå skulle inträffa. Dessa konsekvenser beror på ett antal faktorer som till exempel indämd vattenvolym, dammhöjd och skadeobjekt i översvåmningsområdet. I Sverige klassificeras dammar med avseende på de konsekvenser som kan uppkomma i händelse av dammhaveri. Enligt Miljöbalken 11 kap. 24–26 §§ ska indelningen ske i dammsäkerhetsklasserna A, B och C.

- dammsäkerhetsklass A, om ett dammhaveri kan leda till en kris som drabbar många människor och stora delar av samhället samt hotar grundläggande värden och funktioner,
- dammsäkerhetsklass B, om ett dammhaveri kan leda till stora regionala och lokala konsekvenser eller störningar och dammen inte ska vara klassificerad i dammsäkerhetsklass A, och
- dammsäkerhetsklass C, om dammen inte ska vara klassificerad i dammsäkerhetsklass A eller B.

Till Svenska Kraftnäts sammanställning år 2019 rapporterades 427 dammanläggningar med minst en damm i säkerhetsklass A, B eller C. Av dessa har 267 en eller flera dammar i dammsäkerhetsklass A eller B. Det finns även många fler mindre anläggningar som kraftbranschen valt att göra klassificeringar för utöver A, B och C. I *RIDAS - Energiföretagens riktlinjer för dammsäkerhet*, (2019) finns även klasserna D och E. Dessa kan vara till exempel små dammar från tidigare små järnbruk eller mindre spannmålskvarnar.

Grunden för en god dammsäkerhet utgörs av en väldimensionerad, välbyggd och välskött dammbyggnad. I Miljödepartementet, (2015) beskrivs två ytterligare komponenter i en god dammsäkerhet: säkert handhavande (drift och tillståndskontroll) samt beredskap för dammhaveri och allvarliga problem. Tillståndskontroll görs genom bland annat att analysera risker, instrumentera dammen med givare och övervaka dammens tillstånd. Dessa aspekter är idag centrala i dammsäkerhetsarbetet.

Dammar instrumenteras i syfte att erhålla tidig varning (larm) om en hastig förändring är kritisk men även för att kunna följa och utvärdera eventuella förändringar i dammen samt att ge underlag för en långsiktig bedömning av dammanläggningens tillstånd. Detta ger ett bra underlag för att bedöma dammkonstruktionens behov av eventuella åtgärder.

Vi har valt att i detta kapitel, likt inledningen, ta med all text i kapitel 2 från Jacobson et al., (2021), för att ge en god bakgrund. Detta för att denna rapport skall kunna läsas separat.

3 Data

I detta projekt "Optimerade och Anpassade Datadrivna Metoder för Damm-säkerhet II" används data från två dammar, från olika kraftbolag, som fallstudier. Den första dammen tillhör Vattenfall (fallstudie I) och introducerades i Jacobson et al., (2021). I den rapporten presenteras metoder för vilka har använts för att förbehandla data från fallstudie I. Den fallstudien har fortsättningsvis användas för metodutveckling och utveckling av modeller. Den andra dammen tillhör Statkraft (fallstudie II) och har använts för att testa att implementera de metoder för förbehandling och övervakning som tidigare tagits fram. De två dammarna är individuellt instrumenterade. Instrumentationen mellan dammarna skiljer sig i både typ av givare såväl som antal givare, se Tabell 1.

I fallstudie II introducerades nya typer av givare som ej använts i metodutvecklingen såsom till exempel sprickgivare och Thomsonöverfall. För en generell översikt över instrumentering för dammövervakning se Nilsson, (2014).

Tabell 1. Lista på givare som har gjorts tillgängliga i fallstudien

Typ av givare	Storhet	Vattenfall Fallstudie I	Statkraft Fallstudie II
Tryckgivare för nivå	m	33	15
Temperatur	°C	31	15
Flöde, Läckage	l/s	9	3
Flöde, Kraftverk	l/s	-	3
Läckage – Turbiditet	FNU	2	-
Referens – Utetemperatur	°C	1	2
Referens – Tryck	hPa	1	-
Referens - Magasinsnivå	m	1	-
Referens - Nederbörd	mm/d	1	-
Referens – Vindhastighet	m/s	1	-
Referens – Vindriktning	deg	1	-
Thomsonöverfall	mm	-	2
Referens - Övre vattenyta	m	-	2
Referens – Nedre vattenyta	m	-	1
Sprickgivare	mm	-	4
Temperatur vid sprickgivare	°C	-	4
Totalt		81	51

Data för båda fallstudierna har exporterats från mjukvaran PI som är en långtids-databas med verktyg för analys och visualisering (se "OSIsoft," (2020)).

Fallstudie I har exporterat data vid två tillfällen. Initialt exporterades data under 2020 då första studien i denna serie genomfördes. Under detta projekt har data tillhörande fallstudie I uppdaterats och inkluderar nu ytterligare ett år av data.

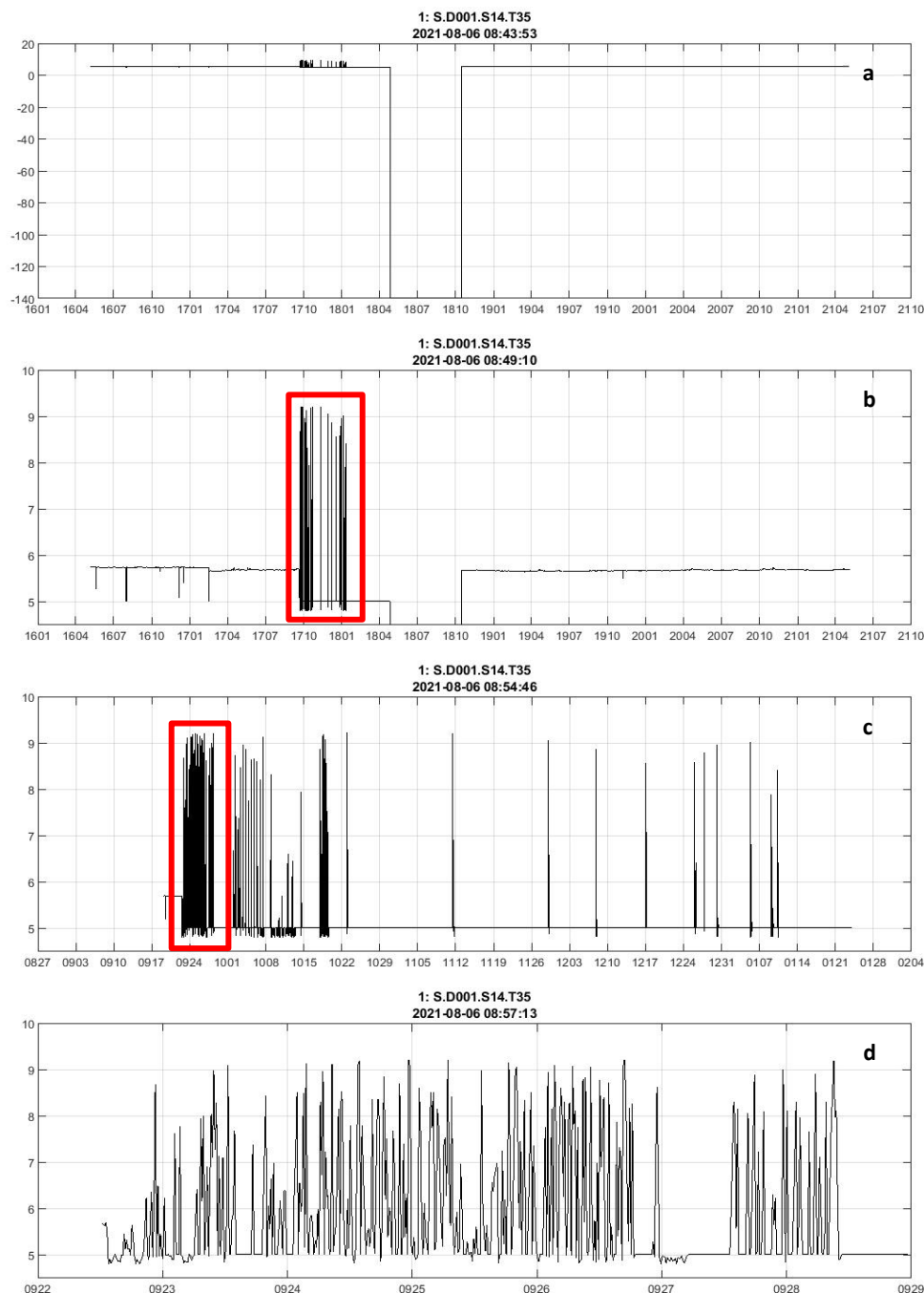
Data från projektets fallstudier representeras med ett samplingsintervall av 15 minuter och är insamlade mellan 1 januari 2015 och 7 september 2021 (de uppdaterade datasetet från fallstudie I) respektive 6 maj 2016 och 6 maj 2021 (fallstudie II). När en eller flera efterföljande observationer saknats för en specifik givare har PI, vid export av data, interpolerat fram värden som saknas till ett samplingsintervall av 15 minuter. Interpolationen är linjär mellan närliggande mätvärden, vilket innebär att förändringshastigheten i mätvärdet antas vara konstant. För fallstudie I har drygt 6,5 års data exporterats för samtliga mät-punkter. Detta motsvarar ca 234 000 observationer för varje mätpunkt. För fallstudie II har drygt 5 års data har alltså exporterats vilket innebär mer än 175 000 observationer från varje mätpunkt.

Fallstudie II som introducerats i detta projekt (från Statkraft) har behandlas enligt samma metoder som utvecklats och används i Jacobson et al., (2021). Datat har analyserats genom att observera historiska rådata för att erhålla en uppfattning av givarsignalernas egenskaper samt förekomsten av fel. Datamängden för denna fallstudie omfattar 55 givare, 5 års data med datapunkter var 15:e minut. All data behöver analyseras och kontrolleras tillsammans med dammägaren för att få en förståelse för specifika givare och identifierade avvikelser. Det senare kan kräva att till exempel en dammsäkerhetsingenjör måste kontrollera med flera aktörer inom sin organisation. Genom att plotta och visuellt studera rådata för samtliga signaler, var för sig, går det att utläsa mycket information om signalen. För att genomföra detta har vi i detta projekt använt oss av "MATLAB," (2021).

Inledningsvis studerades hur samtliga givare i fallstudie II varierade över tid. Avvikelser som identifierades i dessa data var exempelvis:

- Spikar (momentant avvikande värde som snabbt återgår till normalläge)
- Sekvenser med avvikande beteenden
- Tillfällen när flera givare uppvisar liknande beteende
- Trasiga givare

I Figur 1 ser vi rådata från en nivågivare i dammkroppen. Figuren består av 4 grafer med numrering i högra hörnen. Figur 1a visar hela tidserien, med tids-axeln indexerad enligt [år (2 siffror) + månad (2 siffror)]. Figur 1b visar samma data men grafen är beskuren i y-led. Figur 1c och Figur 1d har ett urklipp av kortare sekvens gjorts (beskuren i x-led). Då denna sekvens är betydligt kortare än hela tidsserien har x-axeln indexerats enligt [månad (2 siffror) + dag i månaden (2 siffror)].



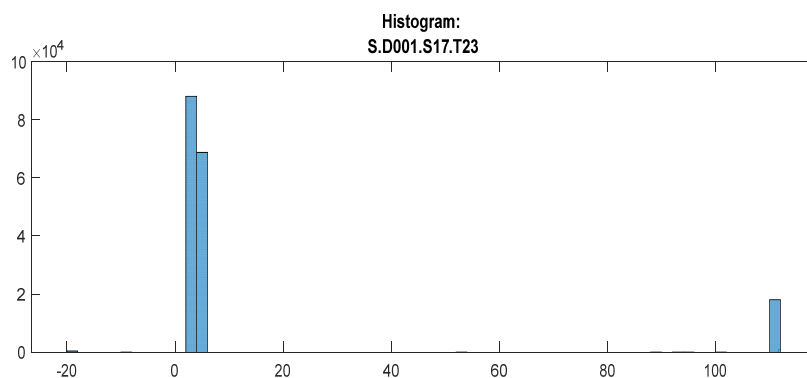
Figur 1. Fyra grafer av samma nivåsignal. Den två första graferna visar signalen över hela tidsserien (5 år) [ååmm] medan de två senare graferna (c-d) visar delmängder av tidsserien [mmdd] där täta spikar identifierats. Detta för att öka kunskapen om avvikelsernas egenskaper

I Figur 1a är det svårt att avläsa någon information av signalen då en sektion i mitten av tidsserien rapporterat ett värde mycket lägre än det som rapporteras i övriga signalen. I Figur 1b har y-axel beskurits för att få en bättre förståelse för vad som sker i signalen. I Figur 1b visar en sekvens, mellan 1710 och 1801, alltså slutet av 2017 fram till början av 2018 (markerat med röd rektangel rött), men flera täta spikar, vilket anses vara ett avvikande beteende.

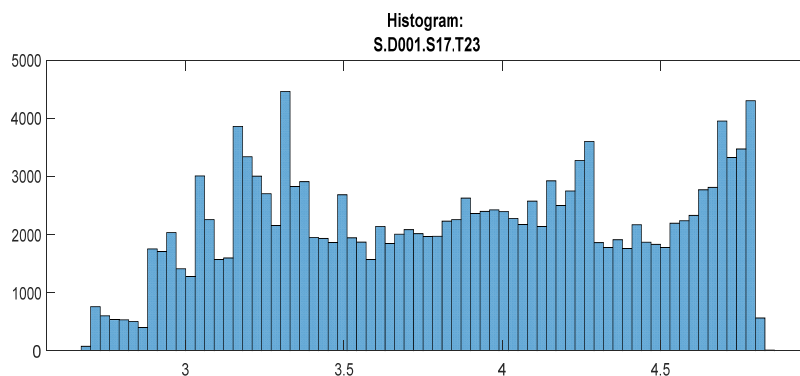
I Figur 1c har figuren beskrivits även i x-led för att försöka få mer information kring avvikelserna. I senare del av Figur 1c identifieras flera enskilda spikar medan tidigare i figuren finns en sekvens med mycket täta spikar. För att få en uppfattning om detta är ett mönster som återkommer i andra signaler har denna signal beskrivits ytterligare en gång, vilket presenteras i Figur 1d. Avvikelsen ser ut att vara oregelbunden, spikarna har olika längd och kommer oregelbundet.

Det finns även andra sätt att visa data för att införskaffa sig mer kunskap om signalen. För att få en tydligare uppfattning om signalens omfång och fördelning kan till exempel histogram (frekvensdiagram) användas. I denna fallstudie finns flera olika typer av givare (temperatur, tryckgivare mätandes nivå, sprickgivare, Thomsonöverfall etc.) som kan förväntas uppvisa histogram med olika utseenden.

Flera temperaturgivare har längre perioder där givaren rapporterat ett kraftigt avvikande mätvärde, som är fysiskt orimligt. Orsaken till detta är troligtvis givar- eller kommunikationsfel. Därav ser histogrammen för rådata från flera temperaturgivare iögonenfallande ut på grund av en kraftig överrepresentation av ett visst mätvärde, se Figur 2. För att få mer information av histogrammet har signalen justerats och kompenserats så att dessa överrepresentationer vilket ger ett mer informativt histogram, se Figur 3. Histogrammen över justerade data för enskilda temperatursignalerna uppvisar ett, till största delen, likartat beteende. Fördelningen är förhållandevis jämn över "normala" intervall (en rektangulär fördelning).

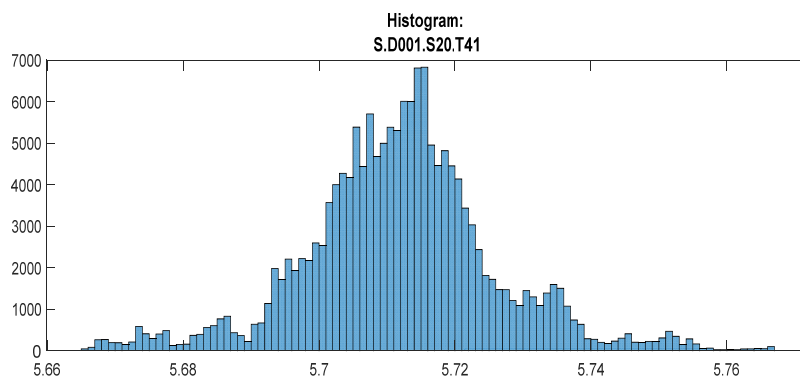


Figur 2. Histogram av rådata för en utvald temperaturgivare.



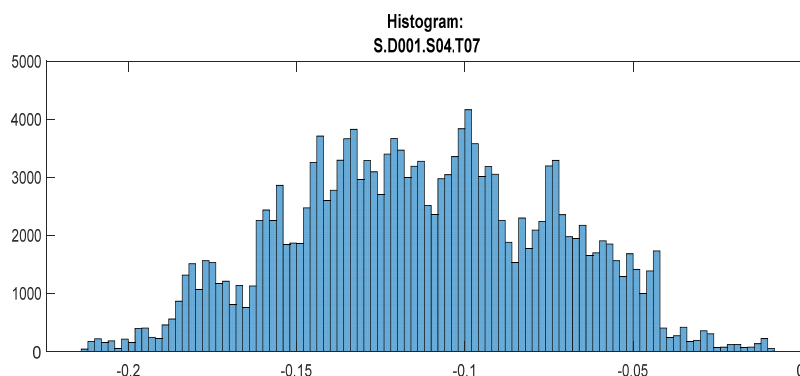
Figur 3. Histogram av justerade data för en utvald temperaturgivare.

Histogrammen för tryckgivare för mätning av nivå varierar mer. Generellt tenderar dessa histogram att vara mer eller mindre normalfördelade, men i vissa fall med en viss övervikt av de lägsta rapporterade nivåerna. Detta har även uppmärksammats vid observationer av rådata. Flera av nivågivarna "bottnar" under en viss del av året. Vattennivån antas, under dessa perioder, vara lägre än nivån som givaren är installerad på.



Figur 4. Modifierat histogram över en nivågivare i dammkroppen. Ofta finns det avvikande mätvärden som urskilljer sig kraftigt från den huvudsakliga fördelningen. Dessa extrema värden har filterats bort innan histogrammet skapats. Signalen har anonymiserats genom att en offset på signalen har introducerats.

Dammen är även instrumenterad med fyra sprickgivare vilka installerats för att bedöma belastningen i dammens betong. Dessa givares histogram ser likartade ut och har ett utseende liknande en normalfördelning med stor standardavvikelse. Den stora standardavvikelsen påverkar grafens utseende genom att "klockan" eller "kullen" i grafen minskar och kurvan blir mer jämnt fördelad, se Figur 5.



Figur 5. Modifierat histogram över sprickgivare i dammkroppen.

Likt fallstudie I är data från fallstudie II generellt årsvis cyklisk. Men cykliciteten inte är lika tydlig som för fallstudie I. Större andel av givarna har en mindre tydlig cyklicitet samt datamängden från fallstudie II innehåller en större mängd bortfall.

4 Detektion av givar- och kommunikationsfel

I projektet "Detektion av givar- och kommunikationsfel vid dammövervakning", Jacobson et al., (2021) togs ett antal metoder för univariata givar- och kommunikationsfel fram. Dessa metoder syftar till att identifiera när givare rapporterar avvikande mätvärden. Nedan beskrivs metoderna kortfattat och de resultat som uppnåtts på data från fallstudie II

4.1 METODER

Detta delkapitel beskriver ett antal metoder som utvärderats på data från fallstudie II.

- **Min- och maxbegränsningar**

En enkel, men effektiv, metod för att undersöka om den nya observationen befinner sig inom givarens fysikaliskt förväntade min- och maxvärden.

- **Identifiera "frusen" signal**

En metod som undersöker om givaren har rapporterat samma värde under en längre period.

- **Begränsningar i differens**

Genom att jämföra differensen mellan successiva mätvärden kan snabba förändringar mellan observationer upptäckas. Om skillnaden är ovanligt stor mellan närliggande observationer kan detta tyda på att det rapporterade värdet är avvikande. Metodiken kan användas för att upptäcka stora hopp eller spikar i data.

4.2 RESULTAT FRÅN FALLSTUDIE II

Här sammanfattas våra observationer vid användning av metoderna som beskrivs i avsnitt 4.1 på data från fallstudie II. I varje figur presenteras två grafer.

Den övre visar historiska rådata från fallstudien, oftast beskuren i y-led för att få en bättre överblick av hur signalen varierar under normalt beteende.

Den nedre visar ett antal binära signaler som presenterar om signalen, utifrån varje tests villkor, befinner sig i ett normalt läge, eller om funktionen har uppmärksammat en avvikelse. Gränsvärden för signalerna är individuellt framtagna baserat på historiska data. Dessa värden kan justeras om bedömningen görs att funktionerna reagerar på beteenden som bör anses vara normala.

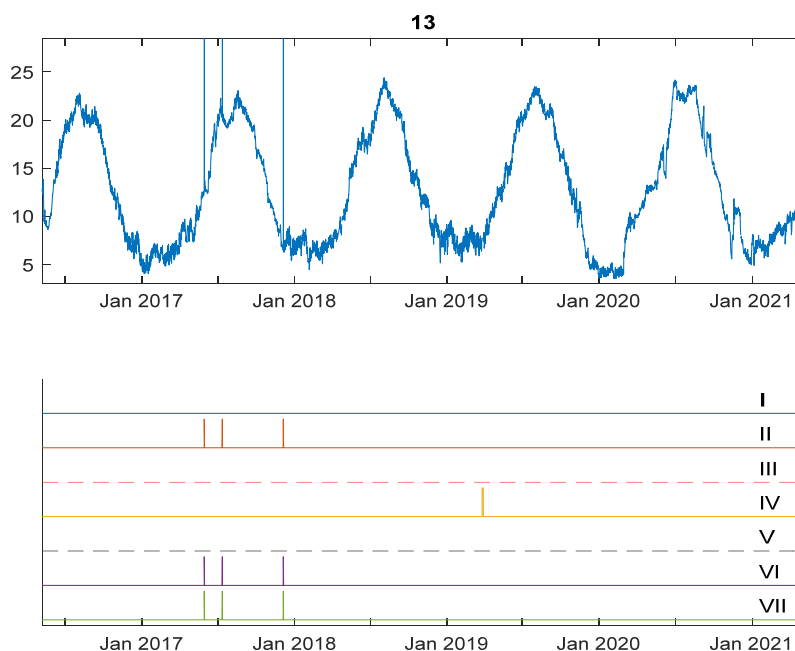
Sju olika tester finns representerade i figurerna, varav två ej implementerades. En offset har adderats till vardera av testsignalerna för att underlätta tydlig presentation av samtliga testresultat.

Följande tester är representerade i figurer sedan, och är numrerade enligt följande i samtliga figurer:

- Mingränsvärde
- Maxgränsvärde
- Givardrift (ej implementerat)
- Frusen signal
- Glidande varians (ej implementerat)
- Andra ordningens differens
- Första ordningens differens

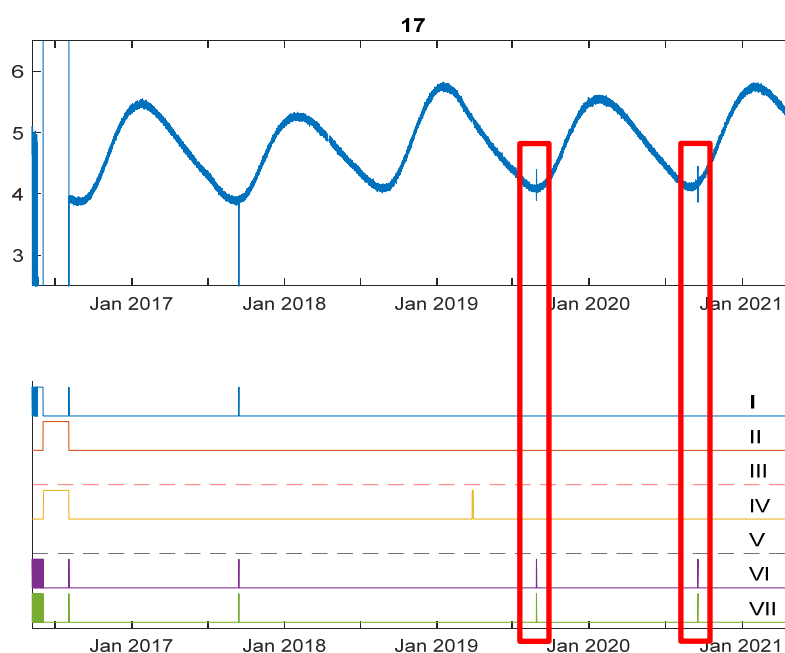
Två av testen är ej implementerade för denna fallstudie. Testet för givardrift exkluderades då slutsatsen i Jacobson et al., (2021) var att detta test är mycket svårt att utveckla univariat. I rapporten beskrivs olika angreppssätt för att övervaka drift, men resultaten var inte tillräckligt tillförlitliga för att användas vid realtidsövervakning. Rörande test för glidande varians implementerades detta för fallstudie I då ett specifikt beteende identifierades av flertalet efterföljande observationer som växlade mellan två värden. Detta beteende har ej identifierats i fallstudie II vilket reducerar detta tests funktion.

Figur 6 visar en temperaturgivare som är placerad intill en sprickgivare i dammkroppen. Signalen innehåller ett fåtal spikar. Notera att dessa upptäcks av flera test (maxgräns, första samt andra differens) när de uppkommer. Spikarna är kraftiga och rapporterar värden större än vad som anses vara "normalt" för givaren, därav markerar maxtestet dessa. Förändringen går fort och därför markeras dessa förändringar även av differenstesterna.



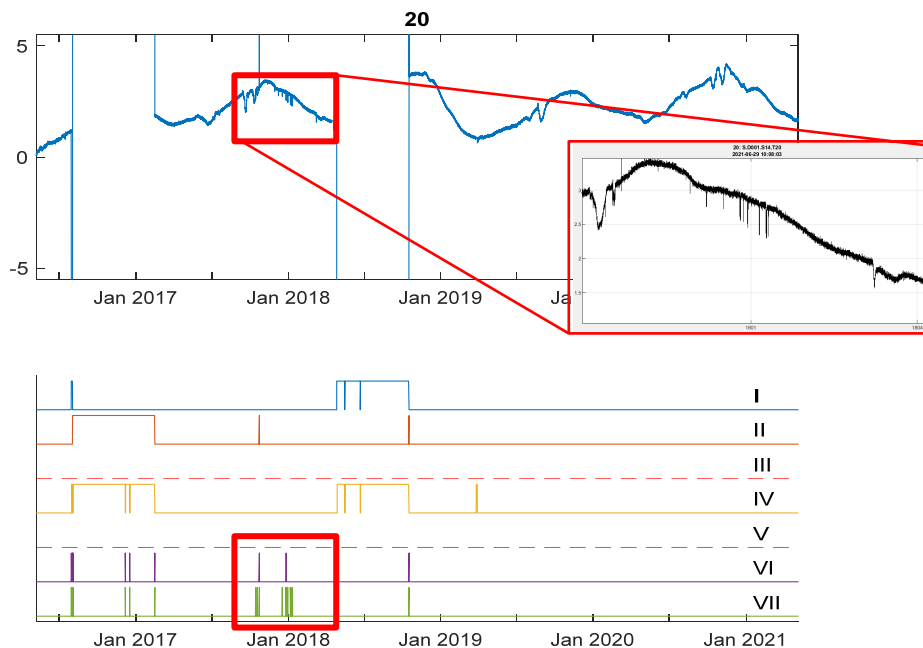
Figur 6. Sprickgivartemperatur från dammen i denna fallstudie, med tillhörande övervakningssignaler. Givarens signal innehåller ett antal stora spikar. Ej implementerade metoder visas med streckad linje.

Figur 7 visar en temperatursignal vars beteende har ett antal avvikelser. Under början av tidssekvensen, 2016, uppvisar signalen ett tydligt avvikande beteende. Inledningsvis innehåller signalen flera stora spikar innan ett konstant felaktigt värde rapporteras. Detta orsakar samtliga test att markera under den första sekvensen i tidsserien. Under resterande tid är felen få. En större spik markeras av min-testet såväl som differenstagare (hösten 2017). Signalen innehåller också ett fåtal mindre spikar under hösten 2019 och 2020 (markerade i rött). Dessa spikar rapporterar värden som anses vara inom signalens normala värden vilket gör det omöjligt för min- och maxtester att markera dessa. Dock är differensen stor vid dessa tidpunkter vilket gör det möjligt för differenstagare att markera.



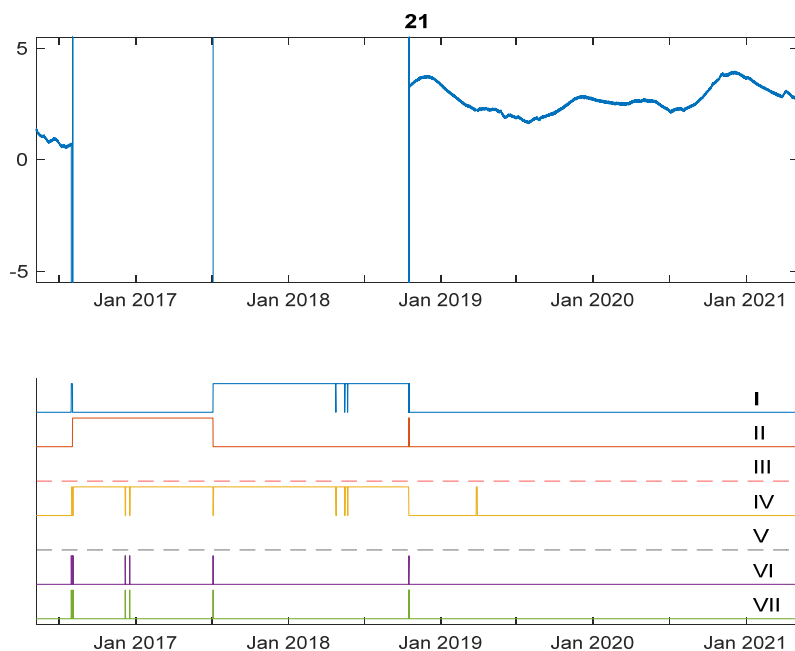
Figur 7. Temperatursignal samt dess övervakningssignaler från dammen i denna fallstudie, signalen inleds med en sekvens av fel och innehåller sedan ett fåtal spikar under resterande tid.

Figur 8 visar även den en temperatursignal från samma damm. Signalen innehåller flertalet fel. Delvis två perioder med värden utanför normalt beteende (årsskiftet 2016–2017 samt sommaren 2018). När dessa sekvensen börjar och slutar markerar differenstagarna och under det att felet pågår markeras detta av både max- och min-test samt frusen signal. Runt årsskiftet 2017 och 2018 (se röda markeringar) får vi samtidigt flera markeringar från differenstagare. Vid en närmare blick på denna sekvens ser vi att ett antal mindre spikar rapporterats där. Det är dessa spikar som markeras av differenstagarna.



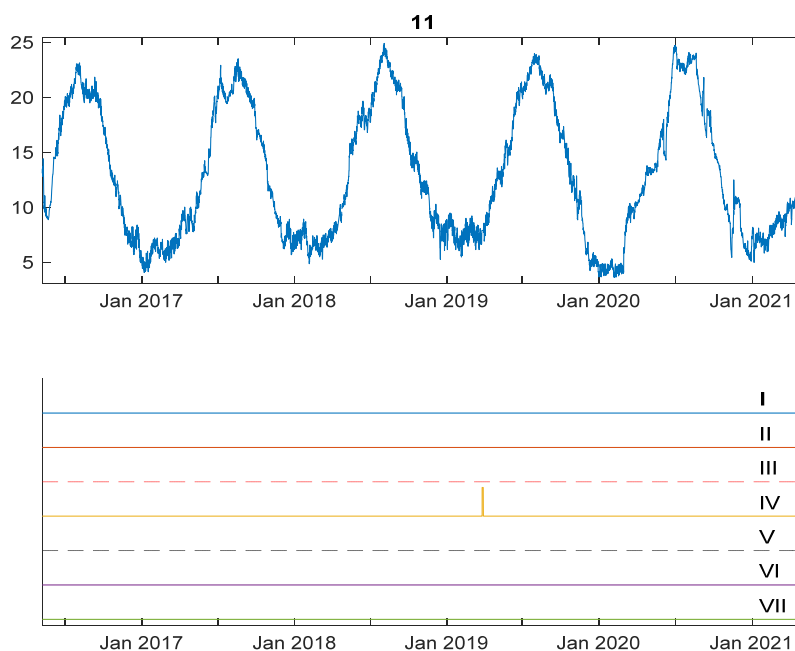
Figur 8. Temperatursignal och dess övervakningssignaler från dammen i denna fallstudie. Signalen innehåller flera mindre spikar samt två sektioner med konstant felaktiga värden.

Flera signaler från denna fallstudie innehåller fel under tidsseriens första 2,5 år. I Figur 9 presenteras ett exempel av detta.



Figur 9. Temperatursignal från dammen i denna fallstudie, tillsammans med övervakningssignalerna. Signalen rapporterar felaktigt värde under halva tidsserien.

Det finns 5–10 signaler som endast uppvisar ett fåtal beteenden som kan tolkas som givar- eller kommunikationsfel. Inget av testen (utom en kort sekvens av frusen signal, vilket kommenteras nedan) reagerar vid något tillfälle i denna tidsserie.



Figur 10. Sprickgivartemperatur från dammen i denna fallstudie, tillsammans med dess övervakningssignaler.

Den uppmärksamma läsaren kan ha noterat att samtliga av de signaler som presenterats ovan har en kort sekvens under våren 2019 där markering av frusen signal sker simultant. Detta är svårt att se när hela tidsserien presenteras i figuren, men det är korrekt att samtliga signaler rapporterar ett konstant värde under ett fåtal dagar vid denna tidpunkt. Markeringen är således korrekt och funktionen markerar en faktisk avvikelse. Tidpunkten har diskuterats med dammägaren, men orsaken till denna tillfälliga störning i signalerna har ännu inte fastställts.

4.3 ANPASSNINGEN AV GRÄNSVÄRDEN FÖR DETEKTION AV GIVARE OCH KOMMUNIKATIONSFEL HOS FALLSTUDIE I

Inom ramarna för Jacobson et al., (2021) gjordes en implementation av univariata övervakningsfunktioner hos Vattenfall. Att sätta gränser för dessa metoder är en iterativ process och har utvärderats ett par gånger under våren 2021.

Metoderna implementerades december 2020, och har varit i drift sedan dess.

Utfallet från metoderna har utvärderats ett antal gånger. Inspektionen har bestått av att kontrollera markeringar som gjorts av signalövervakningsfunktionerna samt även identifiering av avvikande beteenden från signaltrenderna (utan markering av övervakningsfunktionerna). Dessa tillfällen spänner över tiden från december 2020 till juni 2021 och har klassats enligt Tabell 2.

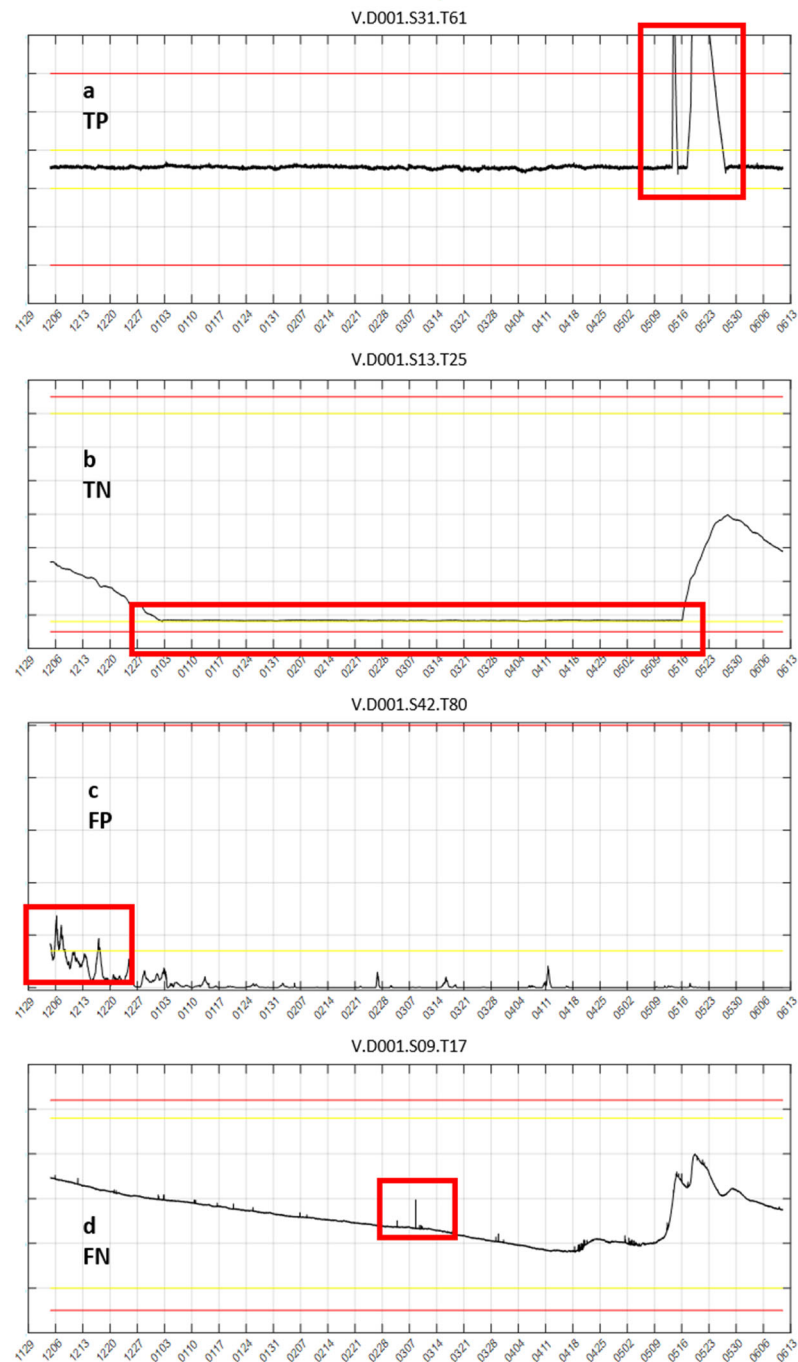
Tabell 2. Klassificering av tillfällen för utvärdering av implementerade övervakningsfunktioner

Klassning	Förklaring	Bedömning	Antal klassade event
True positive (TP)	Sanna positiva dvs. en korrekt markering	Bra	19
True negative SvenskaKraftnät, ()	Sanna negativa dvs. korrekt att inte markera	Bra	5
False positive (FP)	Falska positiva dvs. felaktig markering	Mindre bra, behöver ses över	2
False negative (FN)	Falska negativa dvs. missad markering	Mindre bra, behöver ses över	3

Notera att vi tittade närmare på 29 av 81 givarsignaler, övervakningen av övriga givarsignaler har analyserats utan anmärkningar. Av de 29 så visade 24 korrekt beteende enligt Tabell 2 (TP + TN). Gränsvärden behövde således uppdateras för 5 av 81 signaler.

Utfallet visar att vissa gränser för ett fåtal signaler behöver ses över och troligtvis uppdateras något. I Figur 11 exemplifieras detta med ett exempel av varje klass. Delfigurerna består av rådata (svart), två horisontella linjer för min- och maxgränsvärden för varning (gula) samt två horisontella larmgränser för min- och maxvärden (röda). I Figur 11a markerar övervakningsfunktionerna faktiska avvikelser, ett korrekt beteende. Markering för överstigande av maxvärde rapporteras då signalen rapporterar värden över den röda linjen. I Figur 11b ges ingen markering, detta anses vara korrekt då nivågivaren bottnar och inte understiger larmgränser. Den rapporterar inte heller exakt samma värde som skulle orsakat att frusen signal markerat. Figur 11c markerades av derivator i början av tidserien.

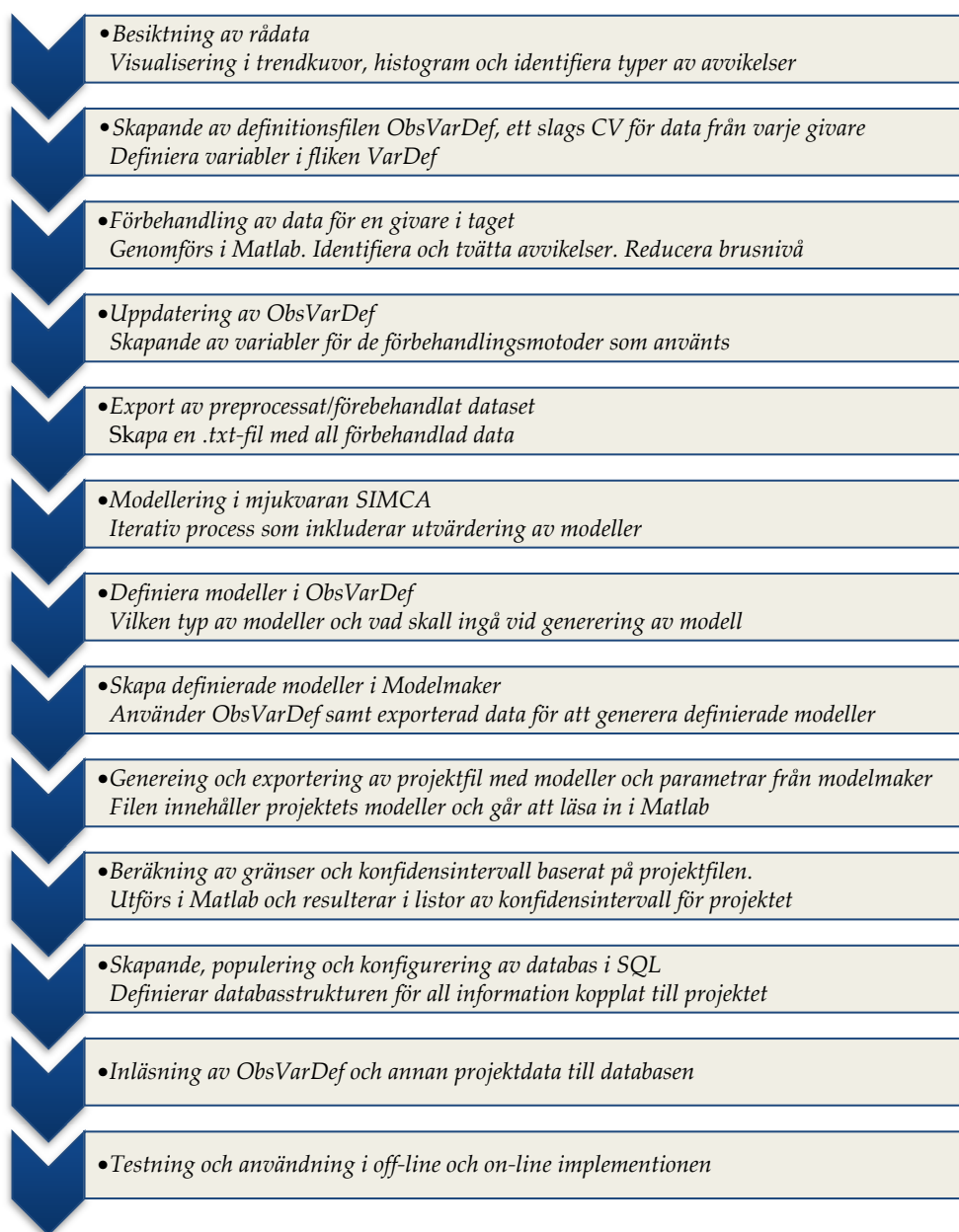
Skillnaderna mellan mätpunkter är relativt stor mot vad som observerats i den historiska data, men detta anses inte vara ett givarfel vilket således kan leda till en justering av gränsdragningen. Figur 11d innehåller en spik i mitten av tidserien. Det ser ut att vara en felrapportering (om än kort och liten). Att se över denna gränsdragning för derivator kan bli aktuellt.



Figur 11. Exempel av events från implementation hos vattenfall. I figurens olika plottar (a-d) visas exempel av de utfall som diskuteras i Tabell 2

5 Arbetsgång för modellering

I detta kapitel presenteras den arbetsgång vi använt för modelleringsarbetet och hur vi går från en första modellering i en programvara med omfattande grafiskt gränssnitt till att modellerna körs online för en viss fallstudie. Arbetsgången följer i riktlinjer som tog upp i den första rapport vi tog fram om metoder för datadriven dammsäkerhet, Johansson et al., (2020). För att vara mer specifika och mer i detalj beskriva arbetsgången i detta projekt så har ett flödesschema gjort, se Figur 12.



Figur 12. Arbetsgång vid modelleringsarbetet

6 Förbehandling av data för modellering

Projektets syftar till att utveckla och implementera multivariata modeller för dammövervakning. För att kunna genomföra detta behöver förbehandling av data göras innan modellerna identifieras. Metoderna för förbehandlingen är likartade med de univariata metoderna för identifiering av givare- och kommunikationsfel som presenterats ovan. I förbehandlingen av datat ersätts avvikande värden med bortfall i dataserien för att dessa avvikande värden inte skall påverka modellen och dess robusthet. Datat behöver alltså "tvättas" från onormala tillstånd eftersom modellerna skall användas för att indikera onormala tillstånd.

Det finns ett antal fallgropar relaterat till dataanalys som är viktigt att ta i beaktande. I Johansson et al., (2020) diskuteras sex vanliga fallgropar och föreslagna motmedel. Dessa är viktiga att ta i beaktande vid analys och förbehandling av data.

Förbehandling av data sker enskilt för varje enskild givare. Målsättningen med förbehandlingen är att få en ren och tydlig datamängd för modellering. De tydligaste behoven som identifierats i dessa data är

- Ta bort stora spikar (som inte härrör till ett normaltillstånd)
- Minska kortvariga variationer i givarsignalerna (så att brusnivån synbart minskas)
- Identifiera "trasiga" givare

Det finns en mängd av förbehandlingsmetoder som kan appliceras för att nå önskat resultat vid denna förbehandling. Metoder finns beskrivna i Jacobson et al., (2021). I denna studie har följande förbehandlingsmetoder använts:

- Min och maxgränser

Flera givare visar mätvärden som skiljer sig avsevärt från den förväntade årliga variationen. Dessa ogiltiga värden kan rensas bort genom att förkasta mätvärden över och under vissa gränsvärden.

- Filtrering

Många givare uppvisar olika nivåer av brus i signalen. Bruset kan reduceras i förhållande till en långsamt varierande signal genom lågpass-filtrering. Brus kan uppstå i en givare eller under transmission från givare till datalagring, och yttrar sig som en slumpartad variation mellan mätvärden (Bulyaculov et al., (2017)). Genom att använda sig av tidigare mätvärden kan beräkningar göras för att få en brusreducerad signal. Filtreringar som används i denna förbehandling är till exempel median-filtrering eller exponentiellt viktat glidande medelvärdesfiltrering (EWMA), se Jacobson et al., (2021).

- Differensbegränsningar

Genom att undersöka skillnaden mellan konsekutiva mätvärden kan signalens momentana förändringshastighet övervakas. Detta görs för att se när avvikande stora och snabba förändringar sker, till exempel när mätvärdet tillfälligt gör ett hopp (en så kallad spik). Det kan också användas för att identifiera när förändringen mellan flera mätvärden i rad är konstant vilket kan tyda på att observationen inte är ett mätvärde utan har interpolerats fram med hjälp av närliggande mätvärden.

Detta beteende kan vara intressant att identifiera, då sådana interpolerade mätvärden eventuellt behöver tas bort för att kunna ta fram tillförlitliga modeller av data.

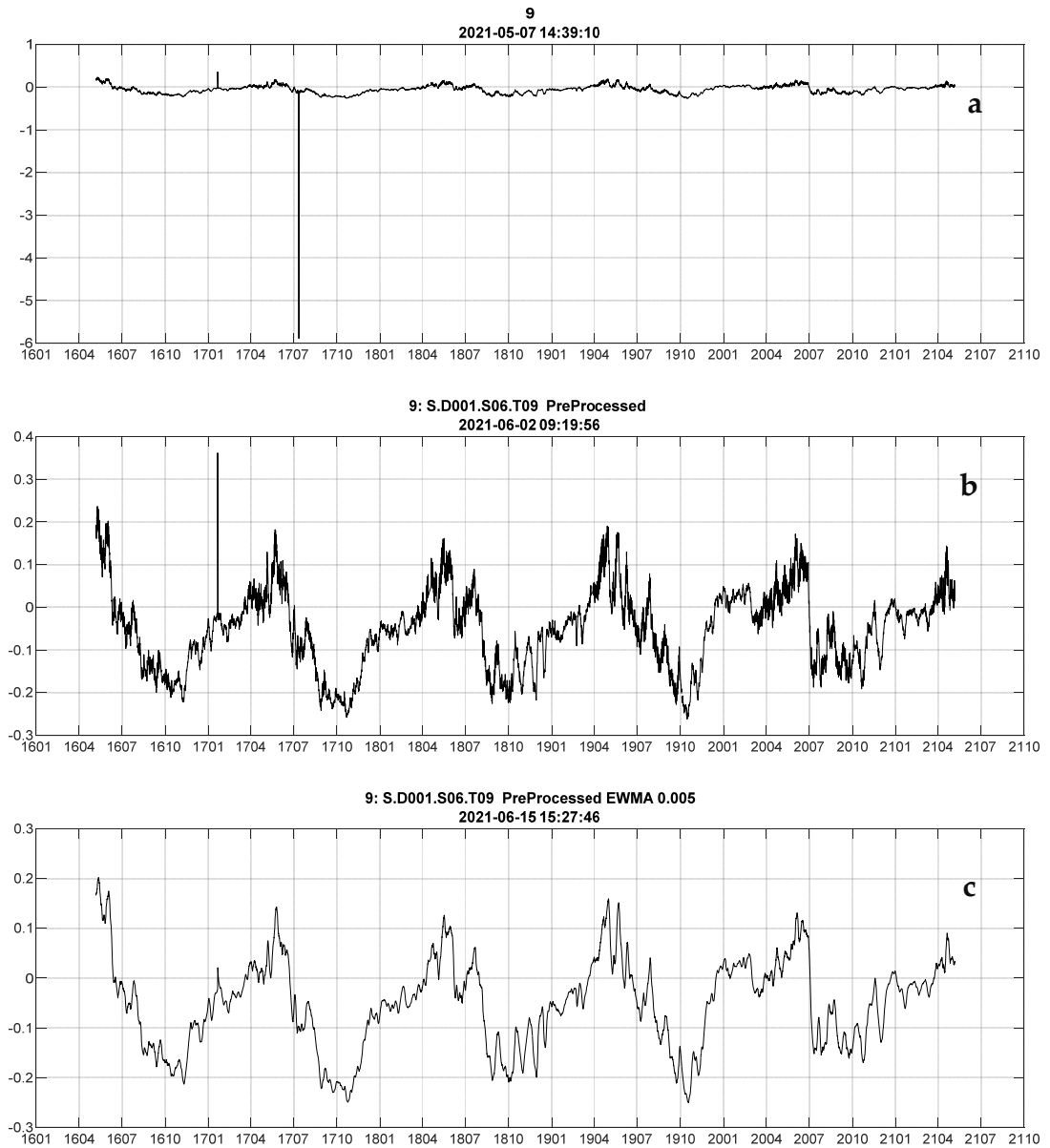
De vanligaste felen som upptäckts i den historiska datamängden är spikar, där en eller flera mätvärden i följd rapporterats utanför den årliga variationen och bryter mönstret i signalen, se nedan. Ett antal givare uppvisar även en hög brusnivå, vilket gör det svårt att utläsa mer information om den naturliga variationen.

Nedan visas resultat av förbehandling av ett antal signaler från fallstudien. Varje figur innehåller tre grafer,

- a. Rådata
- b. Graf av rådata beskuren i y-led (endast i vissa figurer)
- c. Data efter förbehandling

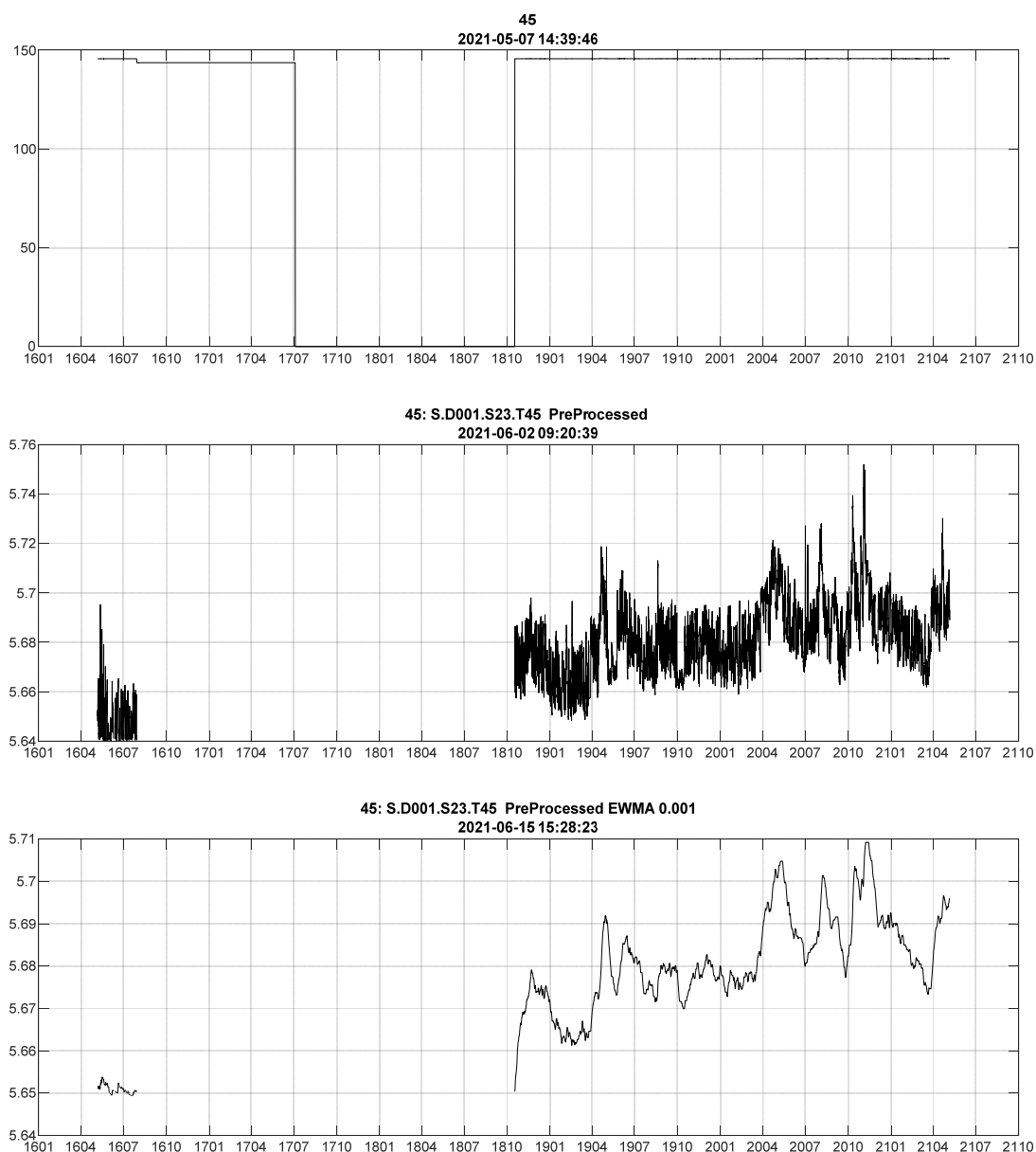
Den nedersta grafen i varje figur presenterar alltså resultatet efter genomförd förbehandling enligt ovan.

Figur 13 visar en signal från en sprickgivare från fallstudie II. Spikar som förekommer i både den övre och mittersta grafen har filtrerats bort i förbehandlingen av signalen. Signalen har även filtrerats för att reducera brusnivån och få en mer stabil signal.



Figur 13. Data från en sprickgivare från fallstudie II där X-axeln beskriver tid [ååmm].

Figur 14 visar en nivågivare. Grafen över rådata är svår att tolka, mer än att givaren rapporterat felaktiga värden under en stor del av denna tidsserie. Redan vid 1608 görs ett skifte och signalen ligger konstant fram till 1707 där den byter rapporterat värde, men fortsätter att rapportera felaktigt. Detta filtreras bort i förbehandlingen och det som återstår finns presenterat i sista grafen. Vilket betyder att data från ungefär halva tidsserien från denna givare inte kommer att användas för utveckling av multivariata metoder. Det finns andra givare liknande beteende vilket kommer göra denna utveckling svårare och kan komma påverka resultatet och kvaliteten på modellerna som ska användas för övervakning.



Figur 14. Data från en nivågivare i dammen där X-axeln beskriver tid [ååmm].

7 Multivariata metoder - Teori

I detta kapitel introduceras de multivariata metoder som använts för utveckling av de multivariata modellerna för fallstudie I och II. Den data som ligger till grund för detta kapitel är samma som presenterats ovan. Datat från båda fallstudierna har förbehandlats och förberetts för multivariat modellering enligt Jacobson et al., (2021) och tidigare i denna rapport. Det mesta av modellutvecklingen har skett interaktivt med programmet SIMCA, version 17 ("SIMCA - Master your data!", (2021)). De färdiga modellerna har därefter genererats för exekvering med programmet VD001ModelMaker, se avsnitt 10.5.2.

Ett ytterligare förbehandlingssteg vid multivariat modellering är autoskalning av alla ingående variabler. Det är en standardiserad skalning där varje signal centreras till medelvärdet noll och skalas till variansen ett. På så sätt ges alla variabler samma betydelse i det fortsatta modellerandet.

Ett generellt dataset utgörs av ett antal observationer (N) och därmed sammanhängande variabelvärden (K). På samma sätt som ett dataset med N observationer och 3 variabler kan representeras av N punkter i ett 3-dimensionellt koordinatsystem, kan N observationer med K variabler alltid representeras som N punkter i ett K-dimensionellt koordinatsystem.

Vi vet att data alltid innehåller både struktur och brus (stokastisk variation kring mätvärden),

$$Data = Struktur + Brus.$$

Syftet med datamodellering är att endast modellera den struktur som finns i data, och inte bruset.

Det som inte modellen modellerar är resten, också kallad residualen,

$$Data = Modell + Residual.$$

Med en bra modell är

$$Modell = Struktur$$

och där

$$Residual = Brus.$$

Måttet $R^2 = (\text{modellerad variation}) / (\text{initial variation})$, vilket alltid är minde eller lika med ett, används för att uttrycka en modells modelleringsgrad.

Därmed finns det två varianter av mindre bra modeller: under- och överanpassade. Vid en underanpassad modell, R^2 för liten, finns det struktur kvar i Residualen. För en överanpassad modell, R^2 för stor, finns det brus i modellen.

De multivariata metoderna PCA, Principalkomponentanalys, och PLS, Partial Least Squares, som använts här, kallas ibland underrumsmetoder eller projektionsmetoder. Det innebär att man approximerar punkternas position i det K-dimensionella koordinatsystemet med punkter i ett A-dimensionellt koordinatsystem. Eftersom A är mindre än K kallas detta koordinatsystem ett underrum, också kallat ett A-dimensionellt hyperplan. Observationernas läge i K-rymden approximeras av observationerna i A-rymden, vilket också kallas att lägena i K-rymden projiceras ned i A-rymden. Ofta vill man att underrums-

approximationen skall fånga så mycket av observationernas utbredning i K-rymden som möjligt.

Ett exempel på detta är att jämföra ett foto av ett objekt, till exempel en buss. Fotot är en 2-dimensionell underrepresentation av bussen som är ett 3-dimensionellt objekt

Avståndet mellan en observation i K-rymden och i A-rymden kallas Distance to Model, DMod.

Axlarna i underrummet kallas komponenter och beskrivs matematiskt som riktningskoefficienter, loadings i förhållande till K-axlarna.

Observationernas lägen längs A-axlarna kallas Scores, T. En sammanfattning av en observations läge i underrummet ges av Hotelling's T², HT², som bildas av scores i alla A dimensionerna.

Valet av antal dimensioner i A-rymden är ett viktigt steg i både PCA och PLS modellering:

- Om för få PCA/PLS komponenter (A-dimensioner) används kommer modellen att underanpassas och modellen kommer ej att prediktera all struktur i data.
- Om för många PCA/PLS komponenter används blir modellen överanpassad vilket resulterar i att även bruset modelleras.

Inom PCA och PLS modellering används uttrycket "Sparsamhet är en dygd" vilket syftar till att en modell inte skall inkludera mer modellparametrar än nödvändigt, vilket kan appliceras vid valet av PCA/PLS komponenter. Normalt väljs antalet PLS komponenter genom att optimera modellens prediktionsförmåga för Y (Q²).

I verktyget SIMCA beräknas även konfidensintervall för de två första dimensionerna i A-rymden. I rapporten finns ett antal figurer som exporterats från SIMCA, de visar datasetet i dessa två dimensioner (t1 och t2). I figurerna uttrycks konfidensintervallet som en elips runt punktsvärmen av observationer utifrån antagandet att scores är normalfördelade med väntevärdet i origo, vilket inte är fallet i dessa fallstudier. Elipserna kan därför bortses ifrån i dessa figurer.

Metoderna MLR och HST, som också använts, utgår från det fulla K-dimensionella koordinatsystemet.

7.1 PRINCIPALKOMPONENTANALYS, PCA

PCA är ett sätt att matematiskt beskriva de mest dominanta korrelationsstrukturerna i ett dataset. Datasetet består av K variabler och N observationer, där en observation representeras av en punkt i koordinatsystemet definierat av en PCA. Med N observationer kommer alltså datasetet representeras med N punkter i koordinatsystemet, en svärm av observationer. Koordinatsystemet definieras av ett antal axlar, A, latent variabler eller principalkomponenter, som representerar de mest dominanta korrelationsstrukturerna i datasetet, där:

- Den första principalkomponenten i detta koordinatsystem fångar den största utbredningen eller variabiliteten i punktsvärmen/datasetet.

- Den andra principalkomponenten fångar den näst största utbredningsriktningen, vinkelrätt mot den första principalkomponenten.

Koordinatsystemet kan utökas med fler dimensioner vinkelräta mot tidigare, upp till maximalt samma antal som det finns variabler (K) i datasetet. Men då varje komponent beskriver den största kvarvarande utbredningsriktningen brukar en till tre komponenter räcka för att få en god visualiserbar överblick av en datamängd.

Läs mer om PCA i Johansson et al., (2020) , Varmuza & Filzmoser, (2008) och "Principal Component Analysis," (2019).

7.2 REGRESSIONSANALYS

Regressionsanalys är ett sätt att beskriva responsvariabler, Y, med hjälp av prediktionsvariabler, X. Med hjälp av regression kan en modell skapas som beskriver relationen mellan X och Y och som kan användas för att prediktera Y från X. Inom regression finns flertalet metoder för att identifiera sådana modeller.

I sektionerna som följer beskriver vi metoderna: multipel linjär regression (MLR), hydrostatisk säsongs tidsmodell (HST), partial least squares (PLS) och artificiella neutrala nätverk (ANN).

7.2.1 Multipel linjär regression, MLR

Multipel linjär regression är en utvidgning av linjär regression (en x-variabel predikterar en y-variabel) till att inkludera flera x-variabler för att modellera och prediktera en y-variabel. Med flera y-variabler görs en MLR per y-variabel.

$$Y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + \dots + \epsilon$$

MLR finner de b som minimerar residualen ϵ . Metoden bygger bland annat på antaganden om att alla x är exakta (utan brus), att de är okorrelerade samt utan bortfall (data finns för alla variabler vid alla tidpunkter). MLR bygger även på att antalet observationer är många fler än antalet variabler. Om dessa antaganden inte är uppfyllda, vilket de sällan är i verkligheten, resulterar MLR i överanpassade modeller med hög förklaringsgrad men som har dålig prediktionsförmåga vid användning med nya data.

7.2.2 Projection to Latent Structures/Partial Least Squares, PLS

PLS är en regressionsmetod som kan ses som en utvidgning av PCA till att omfatta data i två rymder, X och Y. Metoden identifierar den underrumsmodell i PLS-komponenter mellan X och Y som maximerar kovariansen mellan observationerna i dessa två rymder. Modelleringen sker stegvis, PLS-komponent för PLS-komponent med ökande prediktionsförmåga. För att minimera risken för överanpassning, för många komponenter, används inre korsvalidering där delar av data iterativt utesluts och predikteras för att minimera denna risk. En PLS-modell går mot samma lösning som en MLR-modell om maximalt antal komponenter inkluderas i modellen. PLS fungerar bra i kontrast till MLR, trots brus i data, med korrelerade X/Y, med bortfall och även med fler variabler än observationer.

Vid utveckling av PLS finns ett antal godhetstal, som till exempel Q^2 . Detta tal uttrycker modellens prediktionsförmåga (Eriksson et al., (2003)). Under den iterativa processen jämförs Q^2 mellan modellerna för att säkerställa att modellens prediktionsförmåga ökar. Andra godhetstal kopplat till en modell är R^2X och R^2Y . R^2X beskriver hur stor delen av X-variationen som är involverad i modellen och R^2Y hur stor del av Y-variationen som är modellerad.

En PLS-modell kan uttryckas med en regressionsekvation,

$$Y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + \dots + \epsilon$$

på samma sätt som vid MLR. Koefficienterna, b_n kan användas för att exkludera givarsignaler som ger ett litet eller inget bidrag till prediktionen.

För PLS-modellerna kan också Variable Importance of Projection (VIP) (Chong & Jun, (2005)) beräknas. VIP är ett annat mått på hur viktig varje variabel i X är i modellen. Måttet handlar dels om hur viktiga variablerna är för att beskriva X, dels hur väl de korrelerar med Y.

För mer information mer om PLS se Johansson et al., (2020), Eriksson et al., (2003) och Varmuza & Filzmoser, (2008).

7.2.3 Hydrostatisk säsongstidsmodell, HST

HST är en standardmetod i dammsäkerhetssammanhang som ofta använder MLR för att modellera en damms respons baserat på miljöparametrar. Metoden har ofta använts som referens i studier med andra metoder för modellering.

HST är en linjärisering av en olinjär relation mellan miljövariabler (magasinets vattennivå, magasinets ålder och tid på året/säsongvariabler) och mätvärden med hjälp av tillagda olinjära termer.

Läs mer om HST i Johansson et al., (2020), Gamse & Oberguggenberger, (2017) och Bonelli et al., (2003).

7.2.4 Neurala nätverk, NN

Neurala nätverk är ett samlingsnamn på ett antal självlärande algoritmer som försöker efterlikna den mänskliga hjärnans funktion. De flesta neurala nätverk byggs upp av ett antal neuroner som kan liknas vid mänskliga nervceller. En neuronnod kan kombinera ihop flera indata med hjälp av olika vikter och applicerar en icke-linjär s.k. aktiveringsfunktion på dem för att beräkna ett utdata-värde.

Ett neuralt nätverk delas in i två faser. Den första fasen består av träning, eller inläring där nätverket presenteras för uppgiften som skall genomföras samt med utfallet som förväntas. Den andra fasen består av tillämpning och verifiering där nätverket använder sig av resultatet av träningen och jämför utfallet med det förväntade resultatet. Dessa två faser kan, men fördel, genomföras i en iterativ process för att nå den precision som eftersträvas. Läs mer om neurala nätverk i Johansson et al., (2020), Haykin, (1994) eller Duda et al., (2006).

7.3 AVVIKELSE FRÅN MODELLENS GILTIGHETSDOMÄN

Alla modeller är lokala approximationer av verkliga samband i data.

En del modeller är baserade på fysikaliska eller kemiska samband, dessa kallas teoretiska modeller. Sambanden uttrycks som naturvetenskapliga formler. Andra modeller är baserade på erfarenhet, historiska data. Sådana modeller kallas för empiriska och uttrycks som matematiska samband. Det finns också modeller som är en kombination av teoretiska och empiriska samband.

Alla modeller har en giltighetsdomän, ju mindre giltighetsdomänen är desto enklare modell: den teoretiska allmänna gaslagen gäller endast för torra gaser, empiriska modeller gäller endast i det signalvärdesintervall som finns representerat i underliggande historiska data. Därför är det viktigt att en modell endast används för data som ligger i dess modellens giltighetsdomän.

Projektionsmetoderna PCA och PLS har inbyggd diagnostik för att avgöra om nya data ligger i modellernas giltighetsdomän eller ej. Metoderna kallas DModX, avstånd till hyperplanet i X-rymden, och Hotelling's T², avstånd i hyperplanet till centrum, se nedan. Om data ligger utanför giltighetsdomänen är modell-användningsresultaten inte tillförlitliga. För MLR-modeller saknas denna diagnostik.

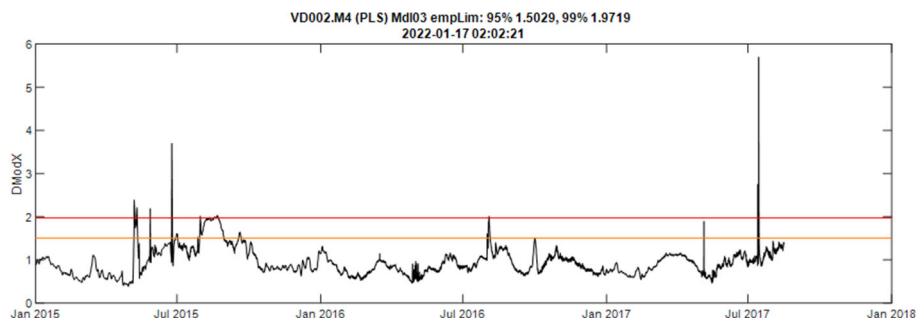
7.3.1 DModX

Avståndet från en observation i X-rymden vinkelrätt till en modells hyperplan i X-rymden kallas DModX, s_i i Ekvation 1. Läs mer i Eriksson et al., (2003).

Ekvation 1. Grundformeln för DModX för observation i . e_{ik} är residualvärdet för variabel k för observation i , K = antal X variabler i modellen, A = antalet komponenter i modellen.

$$s_i = \sqrt{\sum \frac{e_{ik}^2}{(K - A)}}$$

För alla observationerna kan ett fördelningsmått beräknas. I Figur 15 ses en gräns för 95 % (brandgul) och en för 99 %. Dessa gränser kallas DModXCrit. Nya observationer som har DModX som överstiger ett visst DModXCrit anses som avvikare. Försök har gjorts att beräkna DModXCrit för en viss signifikansnivå på teoretisk väg men ingen sådan har stämt i praktiken. Vi har istället valt att använda det empiriska DModXCrit, beräknat på modellens fördelning för DModX.



Figur 15. Exempel på DModX, 95% av observationerna ligger innanför den brandgula linjen, 99% av observationerna ligger innanför den röda linjen.

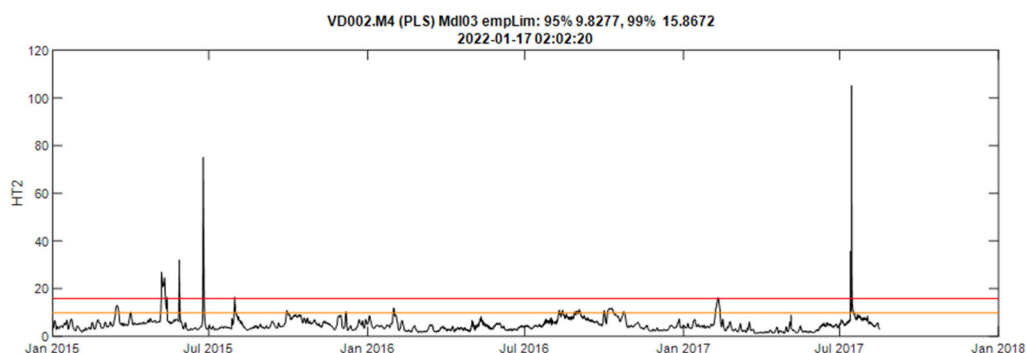
7.3.2 Hotellings T2

Positionen för en observation i modellens hyperplan ges av dess scores, ett värde per komponent. Måttet Hotelling's T2 väger samma dessa koordinater till ett mått, T_i^2 i Ekvation 2, läs mer i Eriksson et al., (2003).

Ekvation 2. Formel för beräkning av Hotelling's T2, s_{ta}^2 är variansen för scores för komponent a .

$$T_i^2 = \sum \left(\frac{(t_{ia} - t_{avg})^2}{s_{ta}^2} \right)$$

För alla observationerna kan ett fördelningsmått beräknas. I Figur 16 två gränser, vilka kallas HT2Crit. En gräns för 95 % (brandgul) och en för 99 %. Observationer som har HT2 som överstiger en viss HT2Crit anses som avvikare. Försök har gjorts att beräkna HT2Crit för en viss signifikansnivå på teoretisk väg, men stämmer dåligt i praktiken när vi undersökt detta, varför vi har valt att använda det empiriska HT2Crit, beräknad på modellens HT2.



Figur 16. Exempel på Hotelling's T2

7.4 KONFIDENSINTERVALL FÖR PREDIKTERADE GIVARSIGNALER

För varje signalgivare som övervakas multivariat finns en PLS-modell som predikterar givarens värde (Y) som funktion av de ingående predikterande signalgivarnas värden. Med prediktionen följer också ett konfidsintervall som beror av önskad konfidensnivå. En högre konfidensnivå innebär ett större konfidsintervall. Om det observerade värdet ligger utanför prediktionens konfidsintervall anses en avvikelse föreligga med ansatt konfidensgrad.

I projektet har vi beräknat konfidsintervallet baserat på ett 2-sidigt klassiskt t-test för modellens Y-residual på önskad konfidensnivå.

PLS modellerna kan, liksom PCA modellerna, också övervakas genom tidsberoende konfidsintervall av principalkomponenterna på samma sätt som beskrivs i 9.3.2.

8 Multivariata metoder – Resultat

Målsättningen för detta projekt är att ta fram modeller för de två olika fallstudierna. Tabell 3 listar de typmodeller som projektet har som målsättning att utveckla. Då instrumenteringen i fallstudierna skiljer sig åt behövs individuell utvecklingen av modeller, baserat i de listade typmodellerna (Se Tabell 1 för mer information rörande instrumentering). Genom att använda samma metodik vid utvecklingen av modellerna för respektive fallstudie möjliggörs en jämförelse mellan dem.

Förkortningar använda i Tabell 3:

PCA	Principalkomponentanalys
PLS	Partiell minsta kvadraters projektion
ÖVY	Övre vattenyta
MSPC	Multivariat statistisk processkontroll
MLR	Multipel linjär regression
NN	Neurala nätverk
HST	Hydrostatisk säsong tidsmodell
SPC	Statistiskt processkontroll

Tabell 3. Projektets målsättning rörande typmodeller för implementation i de två fallstudierna.

Nr	Typmodell	Modell-typ	Gruppering/ dataurval	Användning
1	Övergripande dammtillstånd	PCA	Alla variabler	Visualisera + utanför normal variation detektera
2	Övergripande dammtillstånd	PCA	Alla variabler + $\sin(t) + \cos(t)$ + högre ordning av $\sin$ och $\cos$	Visualisera + utanför normal variation detektera
3	För två utvalda nivåör*	PLS	ÖVY + korrelerade givarsignaler + $\sin(t) + \cos(t)$ + högre ordning av $\sin$ och $\cos$	SPC avvikelsen modell/verklig nivå MSPC
4	För två utvalda nivåör*	MLR	ÖVY + korrelerade givarsignaler + $\sin(t) + \cos(t)$ + högre ordning av $\sin$ och $\cos$ (samma X och Y som PLS)	Jämförelse av modeller med PLS
5	För två utvalda nivåör*	NN	ÖVY + korrelerade givarsignaler + $\sin(t) + \cos(t)$ + högre ordning av $\sin$ och $\cos$ (samma X och Y som PLS)	Jämförelse av modeller med PLS
6	För två utvalda nivåör*	HST	ÖVY + dammens ålder + $\sin(t) + \cos(t)$ + högre ordning av $\sin$ och $\cos$	Jämförelse av modeller med PLS
7	För två utvalda nivåör*	PLS	Som 4 + laggade på timme, på vecka, på månad (att testas)	SPC avvikelsen modell/verklig nivå MSPC

* Innehållande en temperatur och en nivågivare

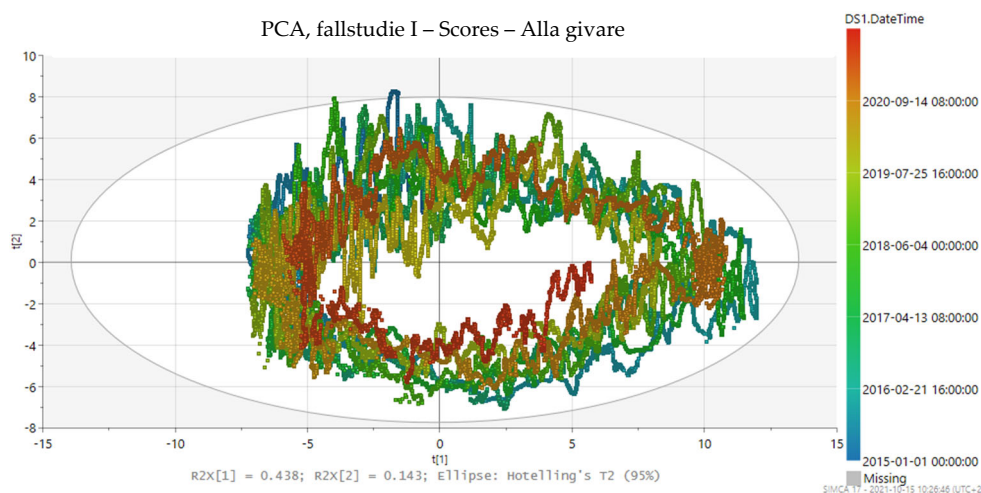
Med dammtillstånd avser vi den information som kan extraheras om dammen genom de installerade signalgivarna

8.1 ÖVERVAKNING AV DAMMTILLSTÅND

8.1.1 Fallstudie I

För att få en översikt över alla dammens signaler har principalkomponentanalys, PCA, använts. Genom denna metod representeras tillståndet för alla dammens givare i PCA-rymden. För att få en tillförlitlig PCA-modell är det av högsta vikt att data är väl förbehandlat.

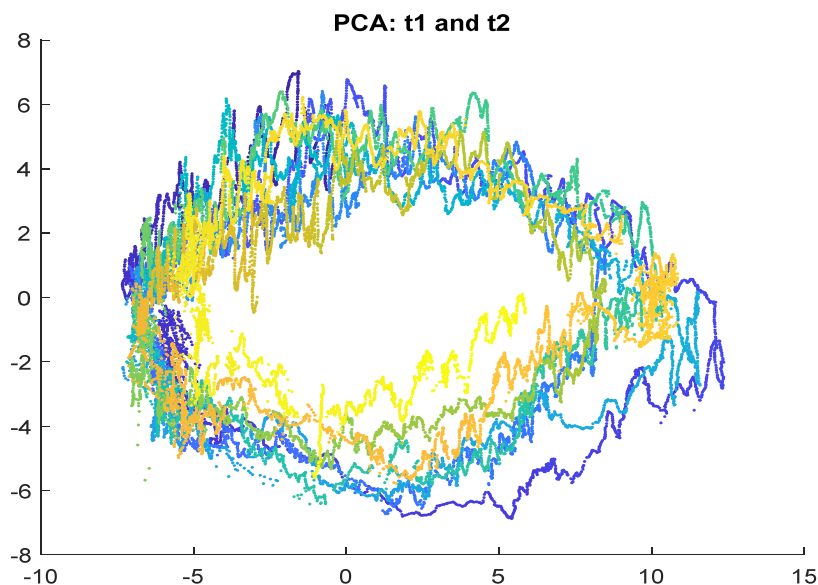
Majoriteten av givarna i fallstudie I uppvisar ett cykliskt beteende med tidsintervall på ett år. Detta medför att den PCA analysen av dammens tillstånd också förväntas att vara cykliskt med samma tidsintervall. I Figur 17 presenteras resultatet av PCA-modellen från fallstudie I. I denna figur presenteras drygt 5 års data (1 januari 2015 – 15 mars 2020).



Figur 17. En PCA-modell för dammens tillstånd i fallstudie I

Figuren visar ett tydligt cykliskt beteende som rör sig motsols i fem varv. Varje varv representerar ett år med start nära Y-axeln mellan första och andra kvadranten. Den datamängd som presenteras har sin start den 1 januari 2015, vilket medför att varje kvadrant representerar ungefär ett kvartal under året. Med den vetskapen kan slutsatser om vilka delar av året som dammens tillstånd visar större skillnad mellan åren dras. Då grafens årliga cirklar skiljer sig som mest i fjärde kvadranten (högra nedre) antas att den största skillnaden för dammens tillstånd mellan åren är under sensommaren och hösten. När en PCA-modell implementeras och nya observationer analyseras kommer det resultera i en ny punkt som rör sig i enlighet med det cirkulära mönstret.

PCA-modellen som utvecklats har byggts enligt datasetet i Figur 17. Projektet har fått tillgång till nya data som sträcker sig fram till 7 september 2021. De nya data som har använts för att verifiera modellen presenteras i Figur 18.



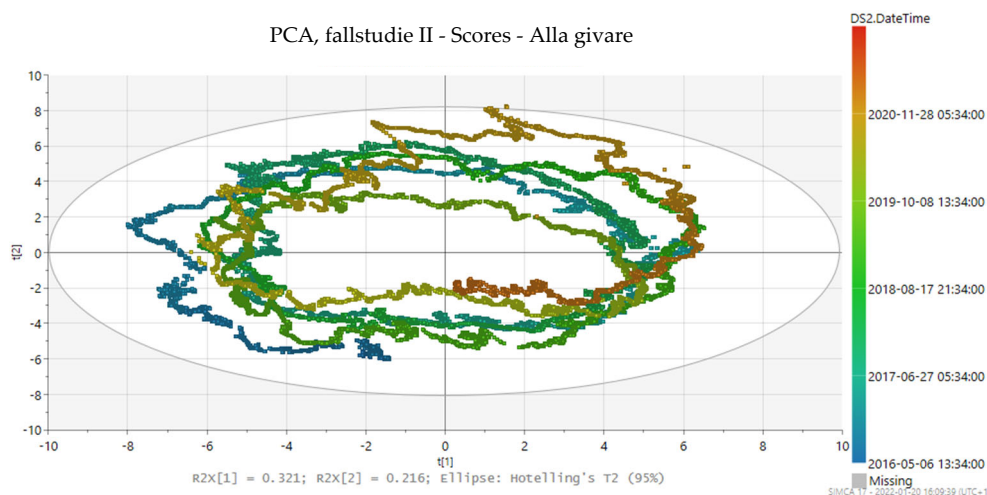
Figur 18. PCA-modell över dammtillstånd i fallstudie I. Datasetet som visas i figuren sträcker sig mellan 15 mars 2015 till 7 september 2020. Detta är en prediktion gjord av modellen som byggts på en delmängd av denna data.

Det kan observeras i de data som adderats till modellen är att det också följer samma cirkulära mönster som identifierats för tidigare år. De mest nyligen adderade datat har färgen gul i Figur 18. Under sommar och höst 2020 (slutet av datasetet) ser vi att detta års "ellips" rör sig närmare origo än vad processen gjort tidigare. Men detta behöver inte betyda att dammen lämnar ett normalt tillstånd, utan att datamängden som använts från fallstudie I inte innehåller data som täcker in hela utfallsrummet av data som bör betraktas som normala tillstånd.

Den metrologiska påverkan är stor. Det kommer varmare och torrare år samt kallare år då modellen ger resultat som kan röra sig utanför den ellips som presenteras i Figur 18 men som ändå bör betraktas som normaltillstånd.

8.1.2 Fallstudie II

PCA används även för visualisering av dammtillståndet för fallstudie II, Figur 19. Tidigare har data från fallstudie II beskrivits som årsvis cykliskt, om än inte lika tydligt som fallstudie I. Datasetet från fallstudie II har även visat sig vara av sämre kvalitet än från fallstudie I. Detta medför att PCA-modellen för fallstudie II inte har ett lika tydligt och unisont cirkulärt beteende. Det är även enkelt att upptäcka att flera tidssekvenser i dessa data har ett tydligt avvikande beteende det cirkulära beteendet, vilket orsakats av brister i datakvalitet vilket i sin tur orsakat att stora mängder data i dataserien ersatts med bortfall.



Figur 19. PCA av dammens tillstånd i fallstudie II

8.2 ÖVERVAKNING AV ENSKILDA GIVARE

Vid utveckling av multivariat övervakning av enskilda givare har historiska data från fallstudie I & II använts för modellbyggandet. Dessa data har förbehandlats enligt Jacobson et al., (2021) och kapitel 6.

Inom projektet är målsättningen att utveckla ett antal PLS modeller som skall övervaka enskilda givare (Y kommer därför, oftast, att bestå av endast signal från dammkroppen). Detta kommer göras genom att jämföra modellens predikterade värde (modellens prediktion av Y) med det observerade värdet (rapporterat från givaren) för samma signal.

När en modell skapas definieras vilka x-variabler (givare) som skall ingå som X. Hur dessa variabler väljs och hur representativa de är för Y påverkar modellens prediktionsförmåga.

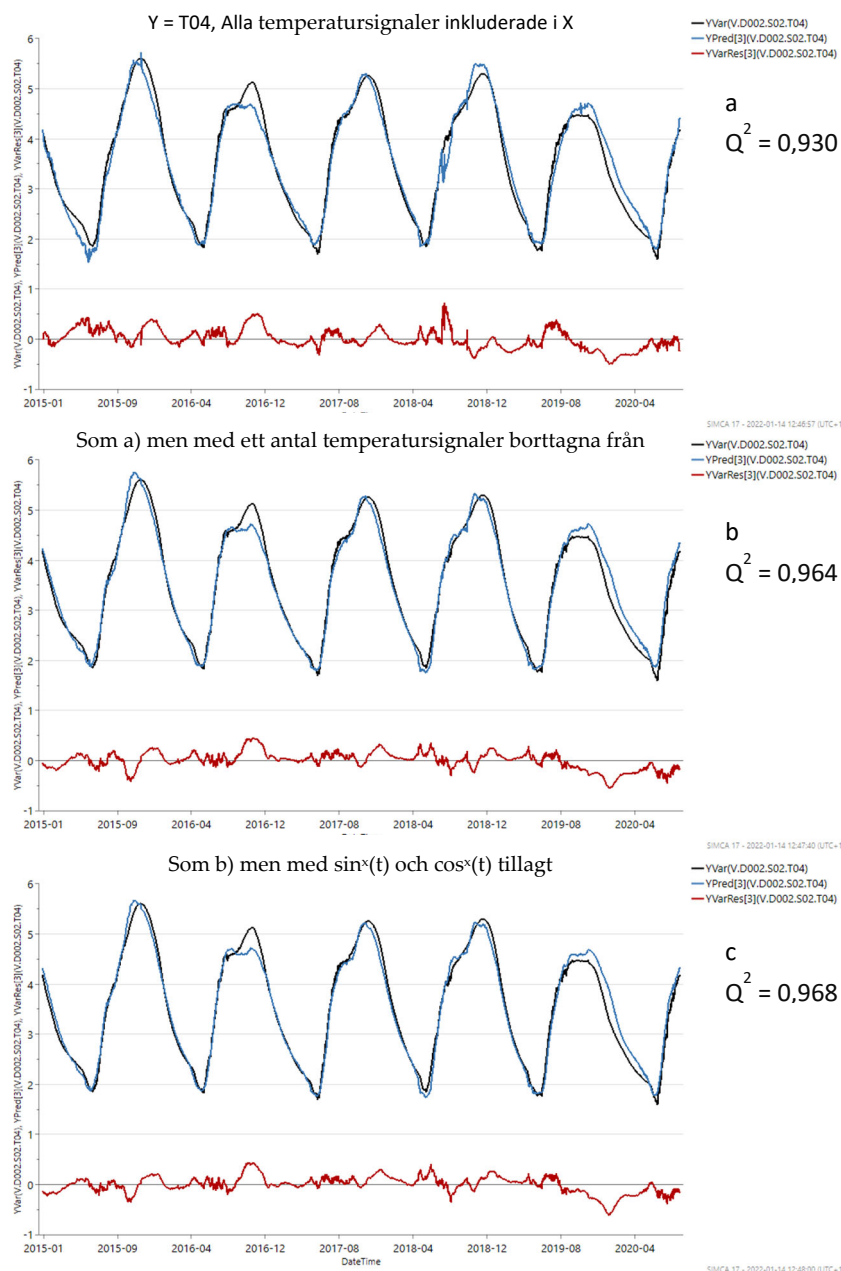
Vid modellbyggande har en utgångspunkt för urvalet av X varit att ha ett stort antal signaler av samma typ (till exempel temperatur eller nivå). Ibland har en iterativ process använts där variabler från datamängden tas bort och läggs till för att nå den bästa modellen, ur prediktionssynpunkt. En kartläggning av givares geografiska position samt givarnas korskorrelation har gjorts för att få en förståelse för hur givare relaterar till varandra på flera plan.

Efter att ha studerat data från fallstudie I från Jacobson et al., (2021) har nästan alla trendkurvor i denna fallstudie identifierats som årvis cykliska. Flera av temperatursignalerna har en form som liknar en sinuskurva. För att få med denna cyklicitet bland X-variablerna har även $\sin(t)$ och $\cos(t)$, inklusive dess kvadrat- och kubiska varianter, inkluderats som X-variabler i modellerna. Då både sinus och cosinus adderats som X-variabler kommer detta underlätta hanteringen av periodicitetens tidsskifte i datat, vilket även ingår i HST-modellens struktur. Även i fallstudie II identifieras de flesta trendkurvor som cykliska, även om cykliciteten inte är lika påtaglig eller övertygande för många av givarna i fallstudie II. Det handlar delvis om att fallstudie II instrumenterats med flera nya typer av givare, vars under-

liggande processer inte är lika årsvis cykliska. Multiplar av $\sin(t)$ och $\cos(t)$ kommer introduceras även till fallstudie II.

8.2.1 Fallstudie I

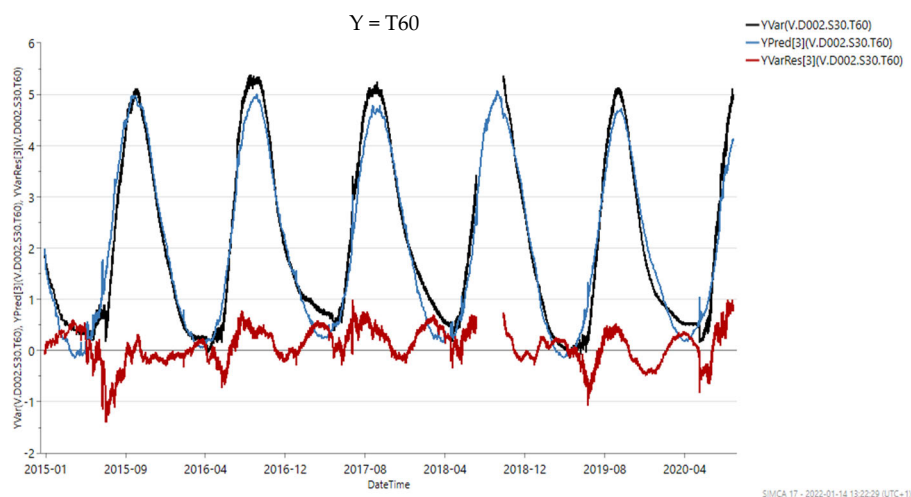
I Figur 20 presenteras ett exempel av modellframtagning för en temperaturgivare i dammkroppen hos fallstudie I. Figuren är uppdelad i tre grafer med tre linjer i var graf som representerar det observerade värdet (svart), det predikterade värdet (blå) och signalens differens mellan det predikterade och observerade värdet, vilken även kallas residual.



Figur 20. Framtagande av PLS-modell för en temperaturgivare i fallstudie I. Figuren är uppdelad i tre olika grafer som representerar den iterativa steg som utförts vid framtagningen av modellen

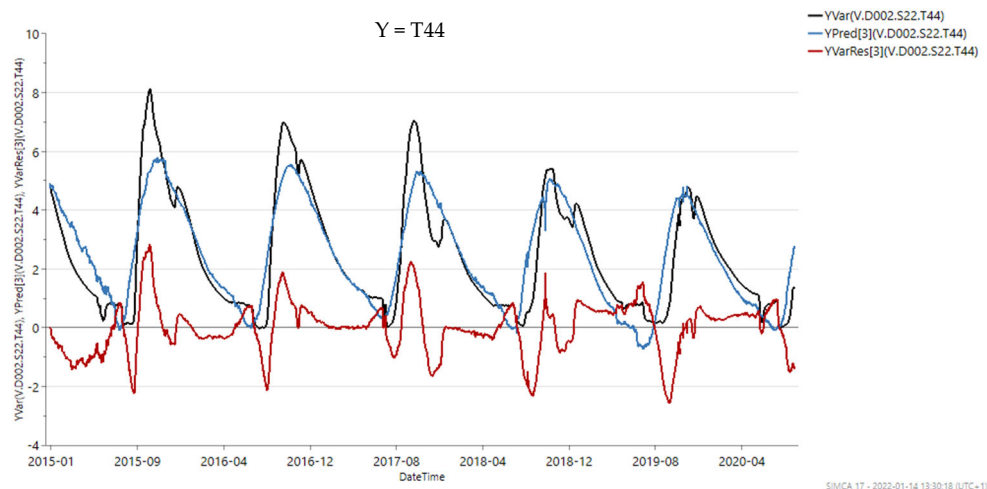
Figur 20a visar modellen där alla övriga temperaturgivare inom dammen inkluderats som X. I Figur 20b har ett antal temperaturgivare tagits bort, då bidraget från dessa signaler hade en negativ alternativt liten påverkan på modellens prediktionsförmåga. I Figur 20c, som är den bästa modellen, har även sinus och cosinus av flera ordningar introducerats i X, för att hantera fasförskjutningen.

Figur 21 är ytterligare ett exempel från fallstudie I av modeller för temperaturgivare som Y med mycket hög grad av prediktionsförmåga.



Figur 21. PLS modell med en temperaturgivare i fallstudie I som Y

Figur 22 presenterar en modell som inte når samma höga grad av prediktionsförmåga och behöver förbättras innan denna kan användas. Värt att notera är att residualen i denna figur uppvisar ett årsvis återkommande mönster. Mönstret tyder på att modellen och signalen inte är i synk och dess årliga högsta punkt inte sker vid samma tidpunkt. Det betyder att modellen är osynkroniserad med signalen, vilket diskuteras vidare i kapitel 8.2.3.

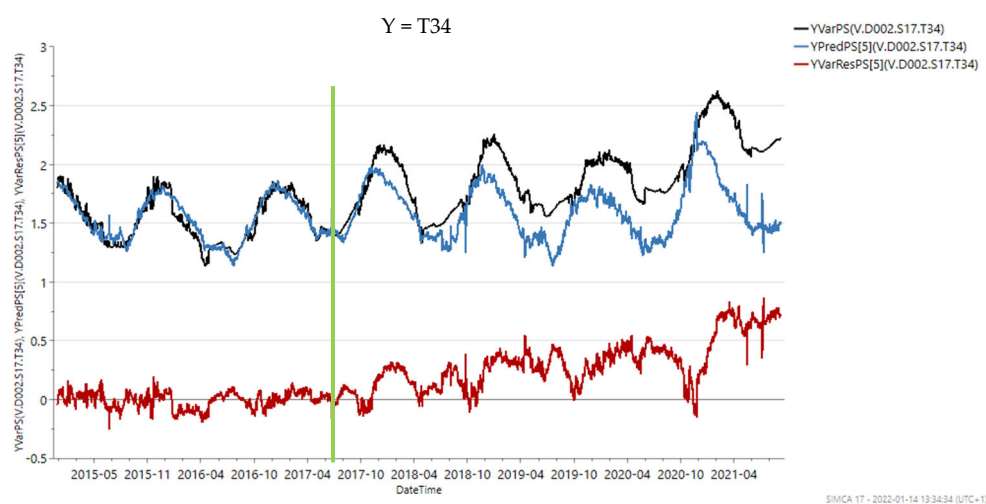


Figur 22. PLS modell med en temperaturgivare som Y

Figur 23 visar en temperatursignal som i Jacobson et al., (2021) identifierats som drivande. PLS modellen för T34 byggdes på data mellan januari 2015 – september 2017 vilket markerats med en grönt lodrätt linje.

Därefter användes modellen för prediktion på data som ej använts i modellbyggandet (augusti 2017 – mars 2020).

I Figur 23 följer prediktionen det observerade värdet fram till den lodräta linjen, därefter skiljer sig dessa signaler åt. Den predikterade signalen rör sig runt samma medelvärde som den första sekvensen av grafen, medan de observerade värdet långsamt driver uppåt. Detta går även att se på residualen som tenderar att öka och hållas till största delen positiv. Detta möjliggör att modellen skulle kunna användas för att identifiera denna drift.



Figur 23. PLS modell med en temperaturgivare som Y, byggd på data mellan januari 2015 till och med september 2017 och använde för att prediktera värde signalens värde mellan augusti 2017 och mars 2020.

I Tabell 4 redovisas godhetstal för de PLS-modeller som presenterats i figurer ovan.

Tabell 4. Nyckelvärden kopplat till modeller från fallstudie I som presenterats som figurer i kapitel 8.2.1.

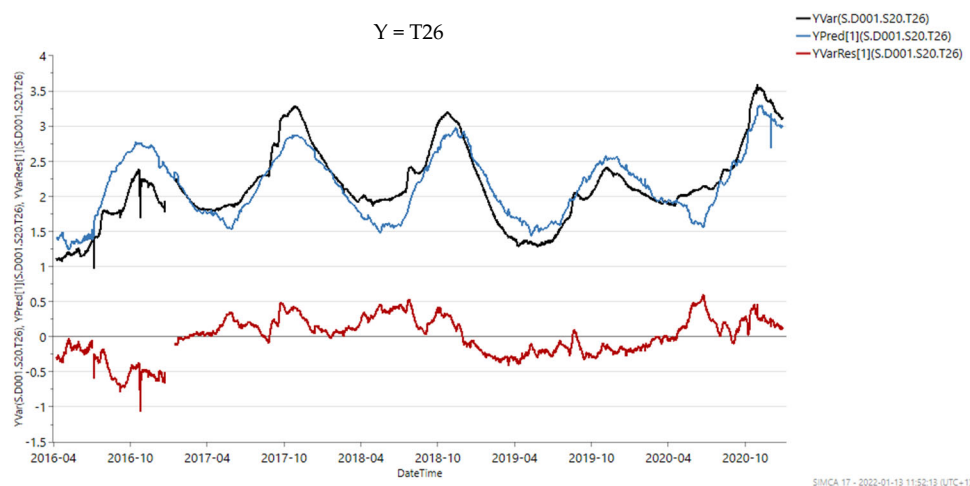
Givare	Figur	R ² X	R ² Y	Q ²
T04	Figur 20	0,823	0,968	0,930
T04	Figur 20	0,929	0,976	0,964
T04	Figur 20	0,924	0,977	0,968
T60	Figur 21	0,871	0,971	0,947
T44	Figur 22	0,907	0,828	0,750
T34	Figur 23	0,946	0,875	0,645

Innan en multivariat modell skall driftsättas bör modellens prestanda utvärderas. Detta har testat genom en implementering off-line vilket beskrivs i kapitel 10.

8.2.2 Fallstudie II

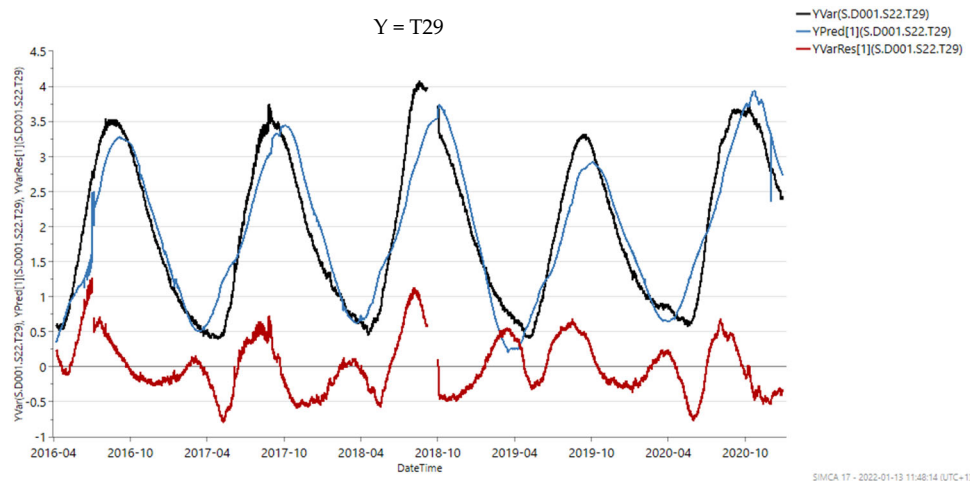
I detta kapitel presenteras exempelmodeller från fallstudie II. Då fallstudie II använder flera typer av givare presenteras här exempel på dessa typer. Generellt sett har färre givare använts i X, detta beroende på att kvaliteten på data generellt sett har en sämre kvalitet vilket påverkar dess inverkan på Y.

I Figur 24 och Figur 25 presenteras modeller för två temperaturgivare i fallstudie II. Som i kapitel 8.2.1 består figurerna av observerat värde (svart), predikerat värde (blå) och residualen (differensen) mellan observerat och predikerat värde.



Figur 24. PLS modell med en temperaturgivare i fallstudie II som Y

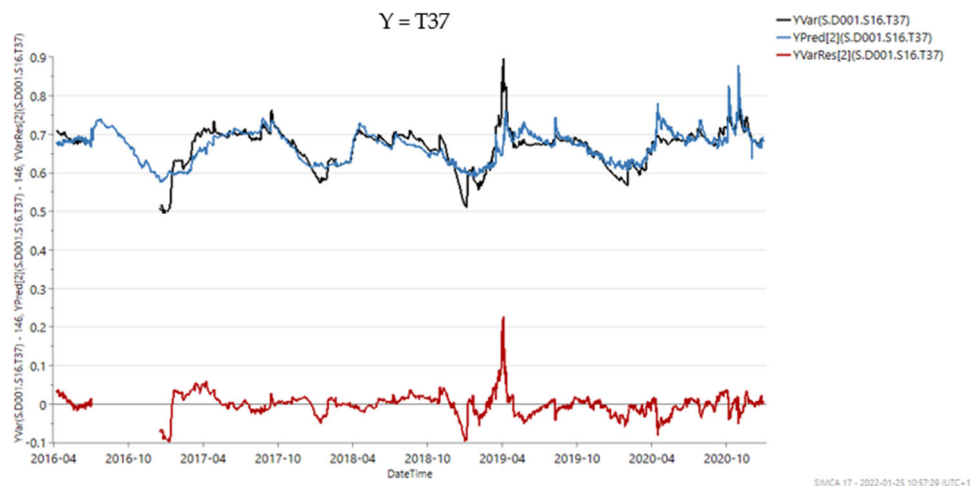
Figur 24 visar att modellens prediktion följer de årliga variationerna samt skillnad mellan åren, även om det finns en viss skillnad i observationen och prediktionen. Residualen är väl fördelad mellan positiva och negativa värden.



Figur 25. PLS modell med en temperaturgivare i fallstudie II som Y

I Figur 25 har residualen ett sinusliknande beteende. Detta tyder på att det finns rörelser i signalen som inte lyckats modellerats korrekt. När prediktionen och observationen studeras kan en viss tidsförskjutning identifieras. Detta skulle kunna motverkas genom att använda sig av kompensations för tidsförskjutning i data som presenteras i kapitel 8.2.3.

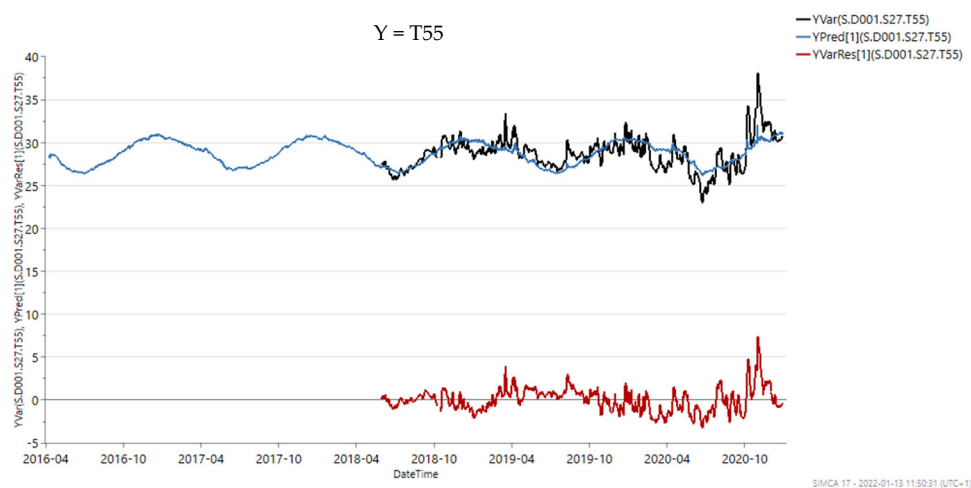
Figur 26 presenterar en PLS modell för en av dammens nivågivare. Signalen för nivå redovisas i meter och för att residualen och samt nivån skall vara överskådlig i samma figur har en offset introducerats för nivågivaren.



Figur 26. PLS modell med en nivåsignal i fallstudie II som Y. Den övre delen visar observerad (svart) och predikterad (blå) tidserie. Den undre delen visar skillnaden mellan de båda, residualen.

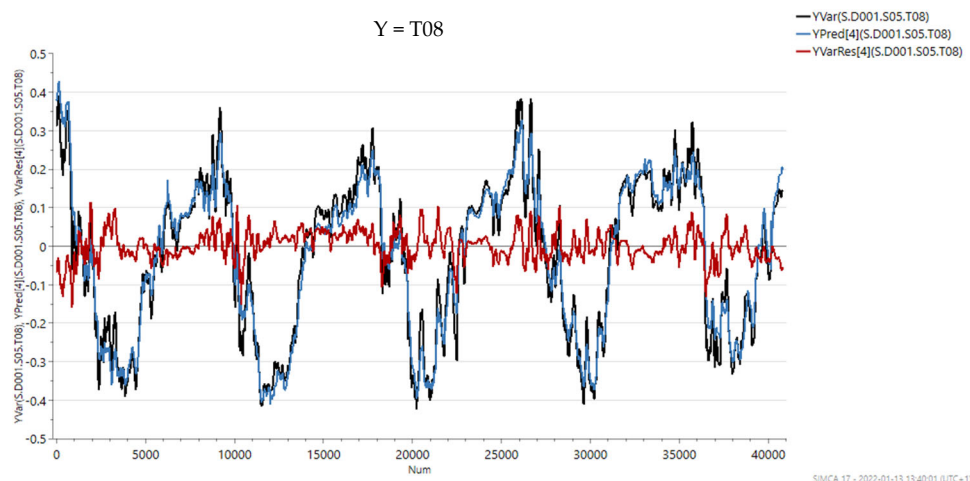
Även i Figur 26 är fluktuerar residualen kring 0. Lägg märke till att det finns en spik i de observerade värdet vid 2019-04. Denna spik saknas i den predikterade signalen vilket också resulterar i att den också blir tydlig i residualen. Det beror på att beteendet inte återfinns i signalerna som används i X. Vid vidare analys är denna observation troligtvis avvikande och något som görs möjligt att upptäcka med hjälp av denna modell.

Figur 27 presenterar en modell för en givare på ett Thomson-överfall. Det observerade värdet har ett stort bortfall under första halvan av tiden. Detta beror på att signalens saknas under denna tidssekvens.

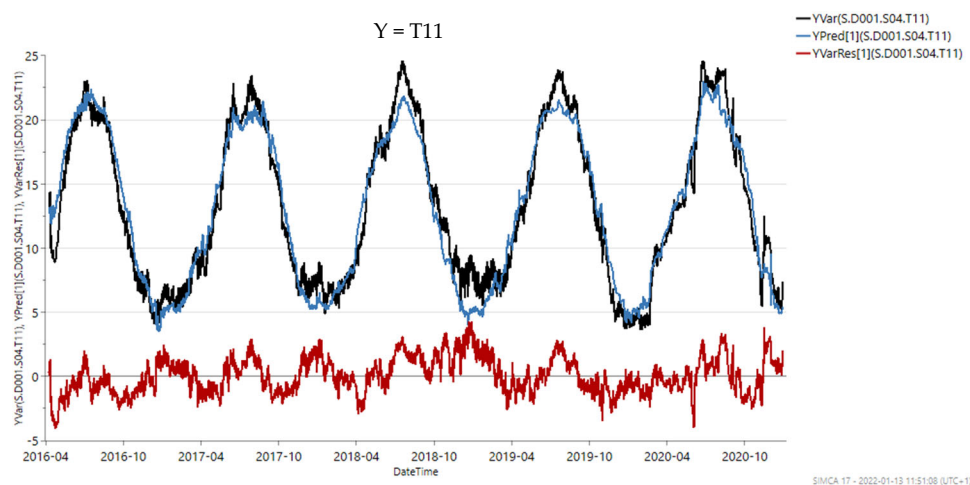


Figur 27. PLS modell med ett Thomson-överfall i fallstudie II som Y

I Figur 28 och Figur 29 presenteras modeller för sprickgivare som mäts i millimeter samt temperatur mätt i betong nära sprickgivare.



Figur 28. PLS modell med en sprickgivare [mm] i fallstudie II som Y



Figur 29. PLS modell för en temperaturgivare placerad i betongen nära en sprickgivare i fallstudie II som Y

Avslutningsvis presenteras godhetstal för modellerna för fallstudie II som presenterats i kapitel 8.3.1. Tabell 5 presenteras godhetstal för modellerna för fallstudie II som visats ovan.

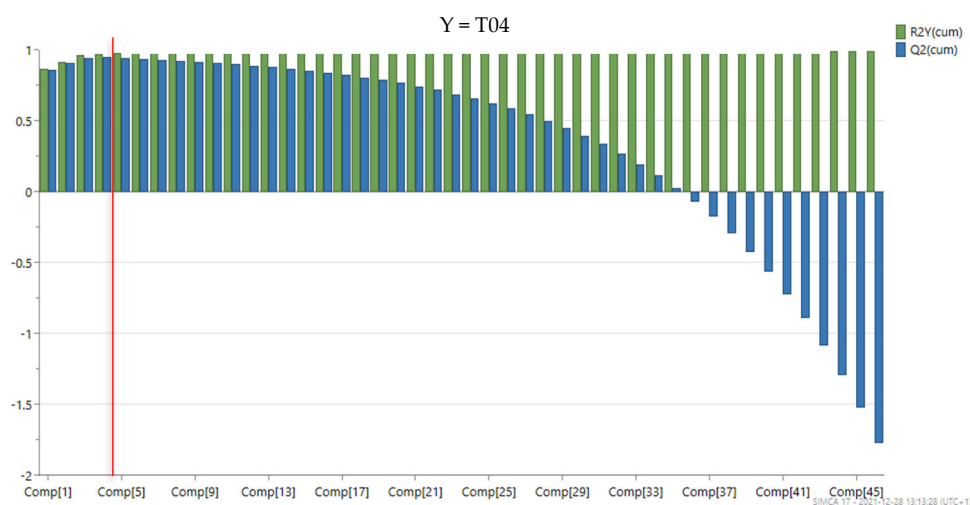
Tabell 5. Nyckelvärden kopplat till modeller från fallstudie II som presenterats som figurer i kapitel 8.3.1.

Givare	Figur	R ² X	R ² Y	Q ²
T26	Figur 24	0,330	0,745	0,683
T29	Figur 25	0,944	0,948	0,954
T37	Figur 26	0,546	0,677	0,571

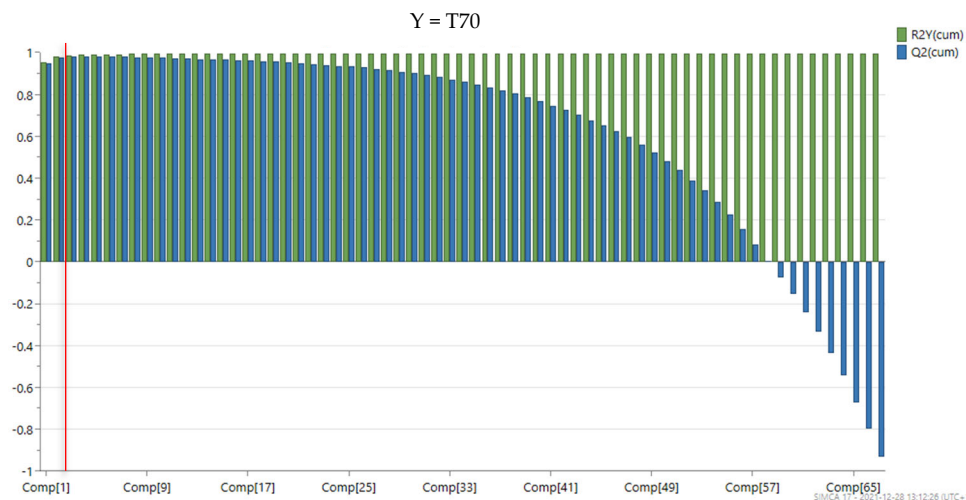
Givare	Figur	R^2X	R^2Y	Q^2
T55	Figur 27	0,656	0,622	0,355
T08	Figur 28	0,851	0,964	0,950
T11	Figur 29	0,690	0,976	0,966

8.2.3 Multipel linjär regression, MLR

I kapitel 7.1.2 nämndes att PLS även kan användas för att beräkna MLR-modeller. Man driver PLS-komponentberäkningen vidare långt efter det att korsvalideringen indikerat överanpassning. För att exemplifiera detta har två givare från fallstudie I använts. MLR och PLS modellerna innehåller samma uppsättning signaler i X och Y. För att illustrera hur MLR med korrelerade x-variabler leder till överanpassade modeller med sammanhängande dålig prediktionsförmåga visas i Figur 30 och Figur 31. Anpassningen av X till Y (minimeringen av Y-residualen) (R^2_{cum} , grön i figurerna) kommer att öka kontinuerligt men validiteten (prediktionsförmågan, Q^2_{cum} , blå i figurerna) kommer att sjunka till katastrofalt negativa värden. I det första fallet stannar PLS genom korsvalidering vid fyra komponenter och vid det andra fallet vid två komponenter. Nyckelvärden för de fyra modeller som presenteras i Figur 30 och Figur 31 presenteras i Tabell 6.



Figur 30. Kumulativ summering av R^2Y och Q^2 för en temperaturgivare från fallstudie I. MLR visar på en överanpassad modell med dålig prediktionsförmåga. Den röda linjen markerar var PLS modellen har stannat genom korsvalideringen



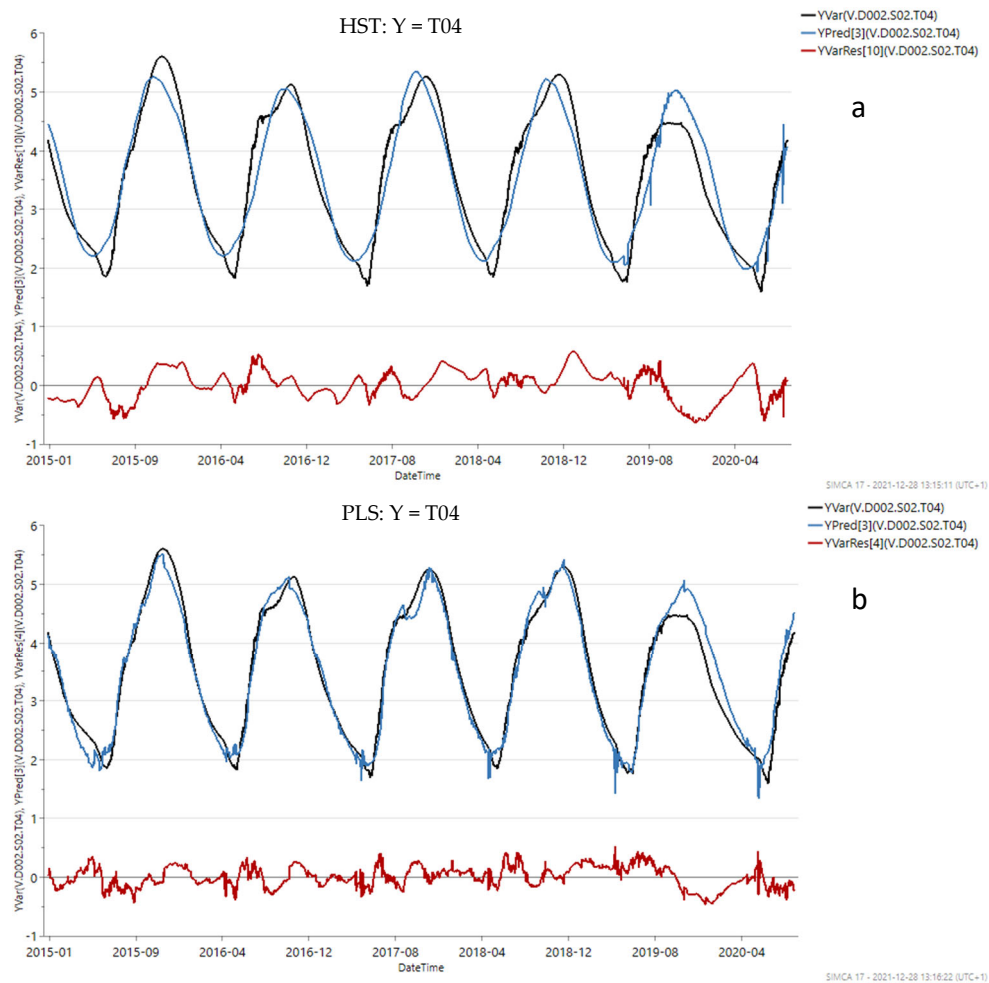
Figur 31. Kumulativ summering av R^2Y och Q^2 för en flödesgivare från fallstudie I. MLR visar på en överanpassad modell med dålig prediktionsförmåga. Den röda linjen markerar var PLS modellen har stannat genom korsvalideringen.

Tabell 6. Nyckelvärden kopplat till modeller från fallstudie I som presenterats som figurer i kapitel 8.2.3

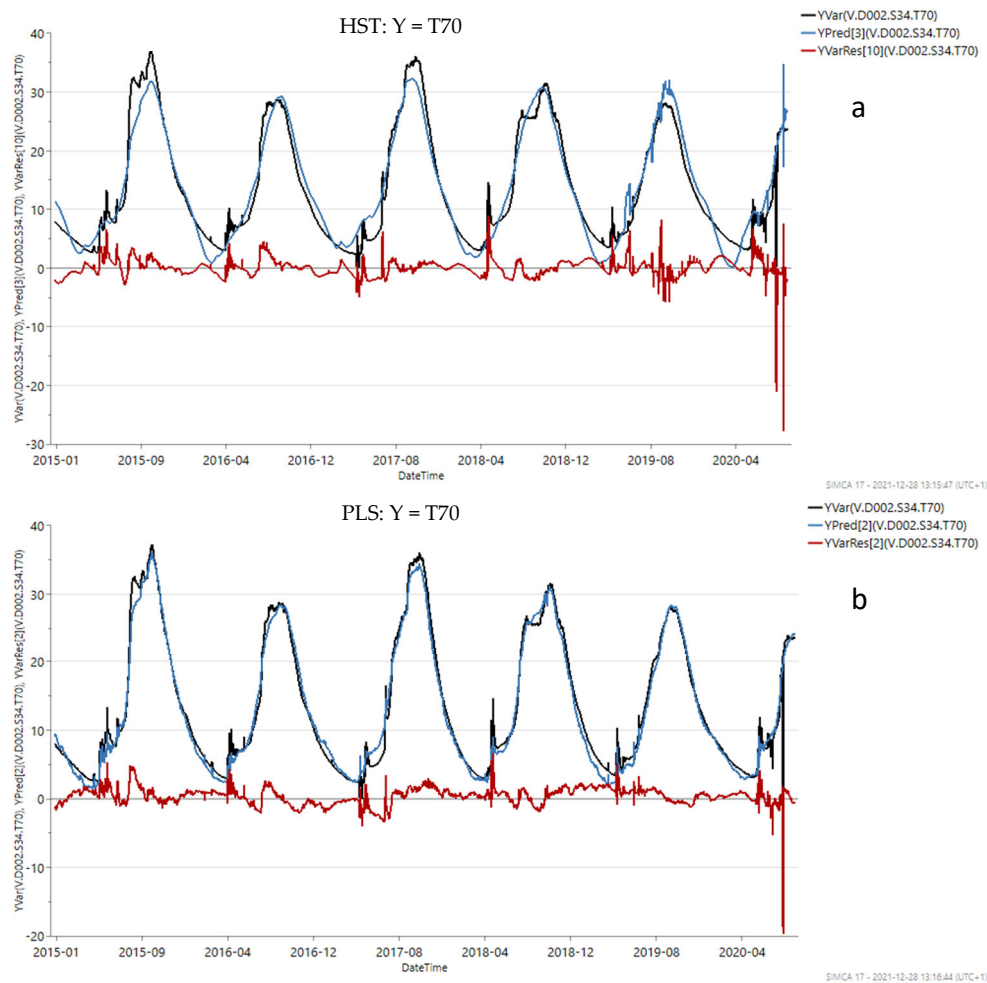
Givare	Figur	Metod	Komponenter	R^2X	R^2Y	Q^2
T04	Figur 30	MLR	46	0,999	0,993	-1,770
T04	Figur 30	PLS	4	0,924	0,972	0,968
T70	Figur 31	MLR	64	0,998	0,995	-0,932
T70	Figur 31	PLS	2	0,706	0,980	0,975

8.2.4 Hydrostatisk säsong tidsmodell, HST

HST modeller har tagits fram för ett par givare från fallstudie I. Dessa modeller bygger endast på miljöparametrar kopplat till dammen. Dessa modeller inkluderar dammens ålder, säsongsb beroende samt övre vattenyta (ÖVY). HST använder alltså inte korrelerade givarsignalen från dammen. Antalet signaler som utgör X är relativt få och de svängningar som finns representerade är ÖVY samt säsongsvariabler. HST modellerna blir därav något trubbiga och variationen mellan år beror endast på ÖVY och på den linjära signalen för dammens ålder. Modellen är alltså begränsad i att modellera snabba förändringar i signalen då dessa sällan återfinns i signalen av ÖVY. I Figur 32 & Figur 33 presenteras en jämförelse mellan en HST och motsvarande PLS för två olika givarsignaler i fallstudie I. Nyckeltalen för dessa modeller presenteras i Tabell 7.



Figur 32. Två modeller över samma temperatursignal i fallstudie I. Figur 32a visar HST modellen och Figur 32b visar motsvarande PLS modell



Figur 33. Två modeller över samma flödesgivarsignal i fallstudie I. Figur 32a visar HST modellen och Figur 32b visar motsvarande PLS modell.

Tabell 7. Nyckelvärden kopplat till modeller från fallstudie I som presenterats som figurer i kapitel 8.2.4

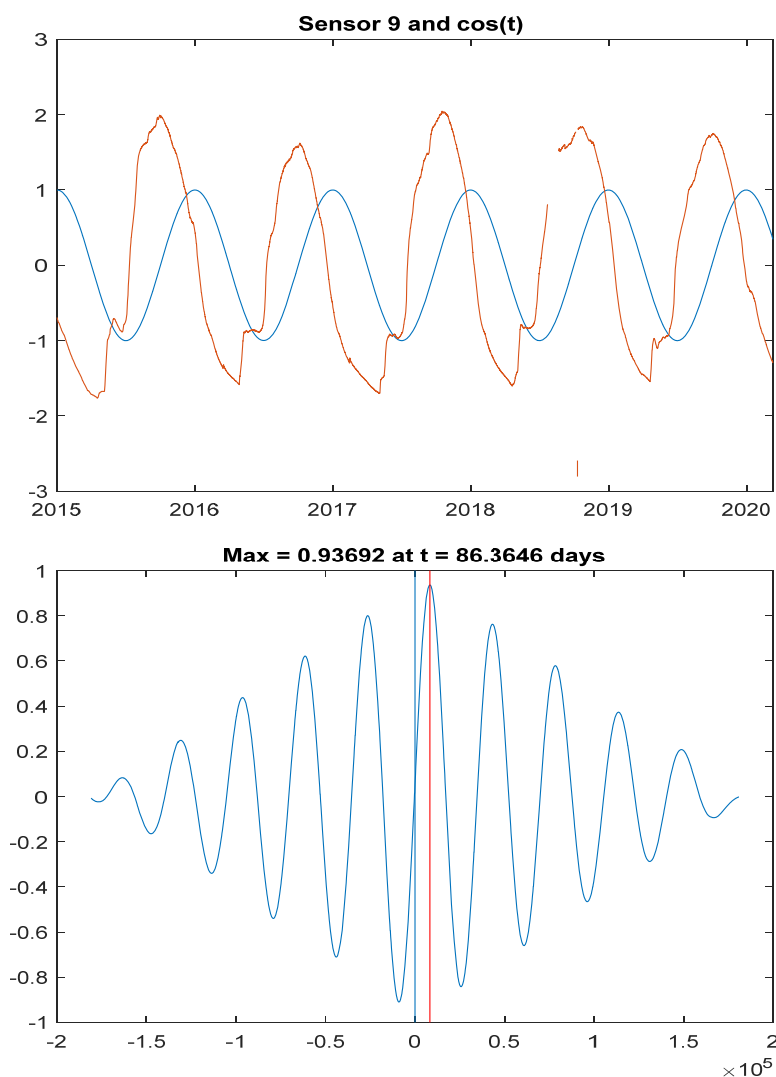
Givare	Figur	Metod	R^2X	R^2Y	Q^2
T04	Figur 32a	HST	1	0,947	0,833
T04	Figur 32b	PLS	0,924	0,972	0,968
T70	Figur 33a	HST	1	0,976	0,959
T70	Figur 33b	PLS	0,706	0,980	0,975

8.3 TIDSFÖRSKJUTNING I DATA

Ett resultat som upptäckts under arbetet med de multivariata modellerna är det som visas i Figur 22, där modellen och observationerna är ur fas.

Detta kan härledas till att det årliga cykliska beteendet givarna sensorerna uppvisar inte har en synkroniserad årlig max- och mintidpunkt. Det kan till exempel bero på givarens position i dammen samt på vilket djup givarna är placerade. Vilket innebär, exempelvis för temperaturgivare, att det tar olika lång tid för värmen och kylan att sprida sig till de olika djupen i dammen.

För att försöka minska inverkan av detta kan givarnas data förskjutas i tid för att synkronisera deras årliga max och minvärde. En metod för detta är att korskorrelera, Johansson et al., (2020), givarsignaler mot en gemensam referens, undersöka vid vilken tidsförskjutning som givarsignalerna korrelerar som mest med varandra och sedan justera givarsignalerna (positionen i datatabellen) enligt den maximala korrelationen.



Figur 34. Korskorrelation mellan en temperaturgivare (röd-linje) och $\cos(t)$ (blå-linje). Den övre bilden visar de två signalerna plottade i samma graf, den nedre visar dess korrelation och en markering där de korrelerar som bäst.

I Figur 34 finns exempel på hur en temperatursignal korrelerar med en cosinus-funktion. Denna signal korrelerar som bäst med en tidsförskjutning av 86 dagar relativt cosinusfunktionen. Relationen till cosinus är inte intressant i sig, men förhållandet mellan olika givare och samma cosinus blir intressant och underlag för tidsförskjutning.

Efter att ha korrelerat samtliga givare på detta sätt har det visat sig att tidsförskjutningen kan skilja över tre månader mellan givare i dammen. Detta har en påverkan vid skapandet av modellerna och kan behöva hanteras för att kunna bygga än bättre modeller.

Några försök med tidsförskjutna data vid modelleringen gjordes men resultaten motiverade inte olägenheten med att behöva vänta upp till tre månader för att ha observationer utan nödvändiga bortfall. Att implementera detta off-line (där data kommer in på samma sätt som on-line) kräver en betydligt mer omfattande datahantering. Driftsättningen on-line blir ännu mer komplext.

9 Detektion av avvikelser från normaltillstånd baserat på multivariata metoder

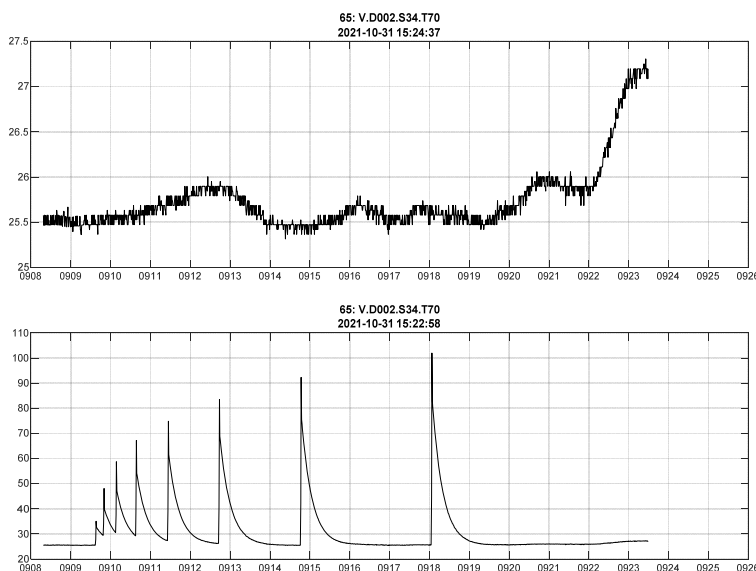
Detta kapitel diskuteras hur de multivariata metoder och modeller som utvecklats skall användas för att identifiera avvikelser från normaltillståndet i dammar.

9.1 SIMULERING AV TILLSTÅNDSAVVIKELSE

En stor svårighet som identifierats under projektet är bristen på data när tillståndet faktiskt avviker från normaltillståndet. Detta är förståeligt då dessa avvikelser är mycket sällsynta. Men då projektet syftar till att upptäcka avvikelser från normalbeteendet måste detta verifieras genom data från en sådana avvikelse.

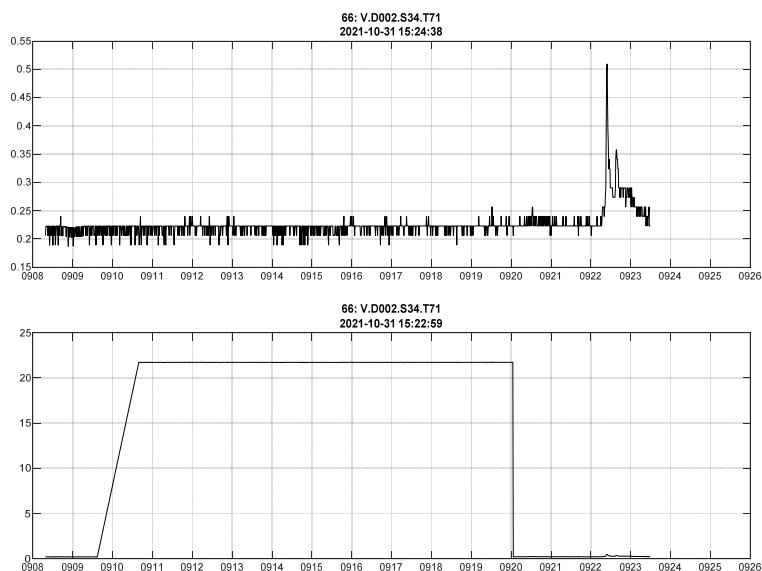
För att möjliggöra verifierande och testande av " trovärdiga " avvikelser diskuterades tillsammans med sakkunniga inom dammsäkerhet inom referensgruppen hur en avvikelse skulle kunna representeras i datamängden. Det finns flera tänkbara orsaker och olika typer av tillståndsavvikelser hos en damm.

Inom projektet beslutades att en avvikelse som kallas "piping" skulle simuleras för att skapa avvikande data. I det simulerade datasetet introduceras piping för en viss position i dammen. Detta resulterar i ett ökat flöde i de flödesmätare som är placerade nedströms den punkt där läckaget placerats. Förändringen i flöde antas ske i pulser då dammen antas återhämta sig mellan pulserna av ökat flöde. Exempel på denna simulering presenteras i Figur 35.



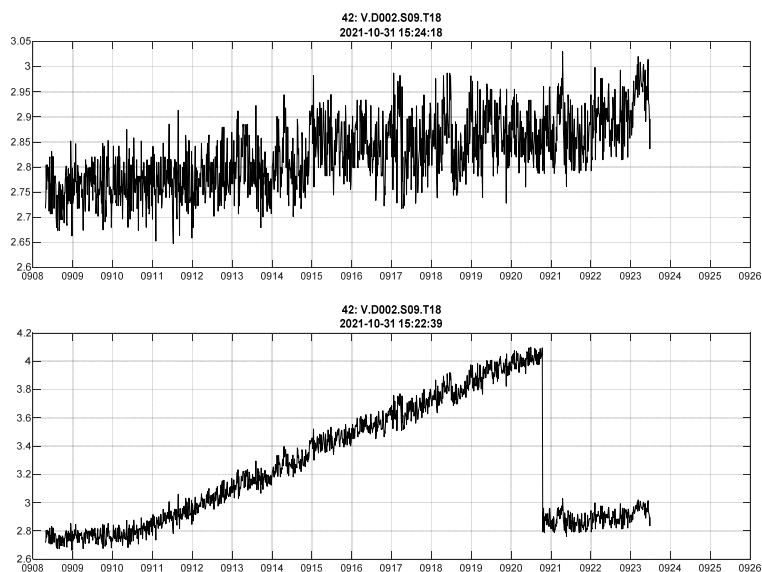
Figur 35. Observerad (övre) och simulerad (nedre) signal för flödesgivare placerad nedströms ett simulerat dammläckage. Lägg märke till skillnaden i Y-axelns storlek för att få en uppfattning av storleken på det introducerade felet. I den observerade signalen ser vi i slutet av sekvensen en viss ökning. Den går även att identifiera i den simulerade sekvensen även om den är mycket liten i jämförelse med de introducerade pulserna.

Om turbiditeten övervakas i samband med flödet kommer även denna öka då det ökade flödet rör upp partiklar från botten, vilket ökar turbiditeten. Detta exemplifieras i Figur 36.



Figur 36. Här visas observerad (övre) och simulerad (nedre) turbiditet för samma tidssekvens av data

Teoretiskt bör även de temperaturgivare som är lokaliserade nedströms läckagepunkten att påverkas av avvikelser. Då genomströmningen genom dammen ökar så förkortas den tid från att vattnet lämnar magasinet tills det når temperaturgivaren. Då denna tid minskar förväntas den observerade temperaturen nedström närma sig temperaturen i magasinet. Om temperaturen i magasinet är högre än temperaturen som observerats nedströms förväntas denna öka för att minska temperaturskillnaden, och om temperaturen i magasinet är lägre bör temperaturen nedströms således minska. Ett exempel på hur detta ser ut för simulerade data presenteras i Figur 37.



Figur 37. Observerad (övre) och simulerad (nedre) temperatur nedströms den simulerade dammavvikelsen

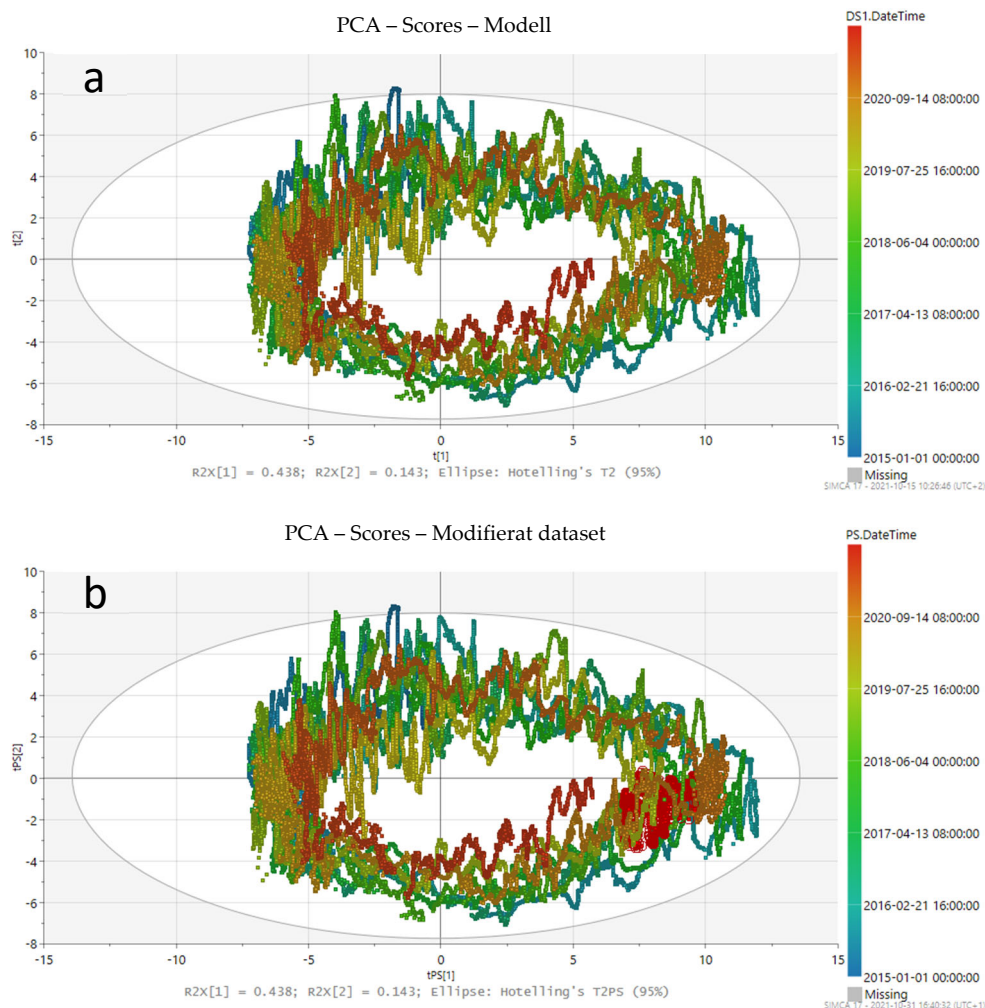
Antalet signaler som modifierats i simuleringen av dammavvikelsen är totalt åtta och presenteras i Tabell 8. Resterande 73 signaler i fallstudie I har inte modifierats vid simulering av dammavvikelse, utan inkluderas i datasetet med sin ursprungliga förbehandlade data.

Tabell 8. Antalet givarsignaler av olika typ som modifierats vid simuleringen av dammavvikelse

Typ av givare	Antal
Flöde	2
Turbiditet	1
Temperatur	5

9.2 RESULTAT VID ANALYS OCH ANVÄNDNING AV UTVECKLADE METODER PÅ SIMULERADE DATA

Simuleringen av piping, som presenterats i 9.1 introducerades vid en slumpmässigt utvalt tillfälle i dataserien. Ett nytt dataset innehållande detta event användes med den PCA-modell som byggs baserat på de ursprungliga data. Resultatet visas i Figur 38.



Figur 38. Figur över PCA-modell för dammens tillstånd. Figur 38a visar modellens ursprungliga resultat efter modellskapandet. Figur 38b visar resultaten när den simulerade dammavvikelsen introducerats. I fjärde kvadranten i Figur 38b kan ett område med röda markeringar identifieras. Dessa punkter representerar den sekvens där dammavvikelsen introducerats.

Figuren består av två plottar. Figur 38a visar modellens ursprungliga resultat, när den byggdes på preprocessad data från fallstudie I. Figur 38b visar resultatet datasetet innehållande en simulering av dammavvikelse. I fjärde kvadranten i Figur 38b kan ett fält innehållandes en rödmarkerad svärm med observationer noteras. Dessa representerar den tidssekvens då dammavvikelsen introducerats. Vid närmare studerande går det att se avvikelens genomslag i PCA-modellen. Men vid besiktning av helheten är det svårare att upptäcka. Den simulerade tidssekvensen rör sig väl inom vad som kan betraktas som normaltillstånd. Detta betyder att en avvikelse liknande denna, som påverkar en liten del av givarna i dammen (1/10) kommer vara svår, beroende på vid vilket tillfälle det sker, att upptäcka på denna globala nivå där dammtillståndet övervakas. Dock kommer dessa beteenden markeras dels av univariata metoderna dels av modeller övervakande dessa givare enskilt.

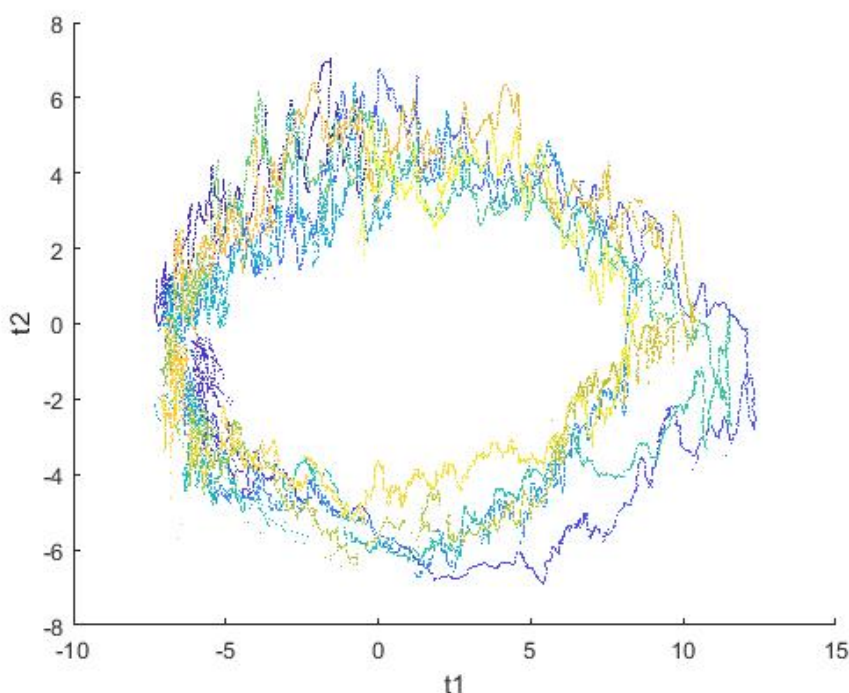
9.3 MODELLÖVERVAKNING AV DAMMTILLSTÅND

För att övervaka ett dammtillstånd har detta projekt fokuserat på de PCA modeller som beskrivits ovan. Under projektet har litteratursökningar på cykliska processer och dammsäkerhet genomförts. Dessa sökningar har inte resulterat i några ansatser att behandla konfidensintervall för dessa processer i litteraturen.

Olika utvärderingsmodeller för PCA-resultatet har diskuterats inom projektet. Utvärdering kan göras mot statiska eller dynamiska (tidsberoende) konfidensintervall. Övervakningen kan även göras i en eller flera dimensioner samtidigt. I detta kapitel beskrivs de huvudsakliga idéerna på konfidensintervall som kan användas för övervakning av cykliska processer.

9.3.1 Konfidensintervall för projektionsvärden (scores), statiska i två dimensioner

För att hitta ett statistiskt konfidensintervall för de två första principalkomponenterna behöver ett polynom beskriva dessa gränsvärden. För fallstudie I, vilket presenteras i Figur 39, skulle dessa kunna bestå av två ellipsliknande former, en som befinner sig nära origo, innanför-, och en större som sträcker sig utanför den årliga variationen. Alltså ett ellipsformat konfidensintervall som avgör om dammens tillstånd befinner sig inom sitt normaltillstånd.



Figur 39. PCA för dammövervakning. samma som figur X men denna gång skapad i MATLAB

Arbetsgången för att beräkna detta konfidensintervall har bestått av flera steg, vilka beskrivs nedan.

- Identifiera ett antal punkter som skall användas vid polynom Anpassning
 - Gruppera data beroende på tid på året
 - Identifiera antalet grupper
 - Identifiera punkter vilka skall representera varje grupp
- Anpassa polynom utifrån framtagna punkter.

För att identifiera polynomen som beskriver de ellipsliknande gränserna behöver ett antal punkter som kan ligga till grund för polynomet identifieras. Genom att gruppera samtliga observationer i Figur 39 i mindre data delmängder (svärmar) kan punkter representerades dessa delmängder beräknas.

För varje delmängd kan ett separat medelvärde och en standardavvikelse beräknas för de två första principalkomponenterna, som också utgör axlarna i Figur 39.

För varje delmängd kan medelvärden för t_1 och t_2 bilda medelvärdeskoordinater, och med hjälp av standardavvikelserna kan en rektangel kring medelvärdeskoordinaterna bildas. De fyra hörnen som begränsar varje rektangel beräknas enligt Ekvation 3.

Ekvation 3. Beräkning av rektangelns hörn som potentiellt skall användas för anpassning av konfidensintervall

$$\begin{aligned}
 a) (p_{1_{t_1}}, p_{1_{t_2}}) &= (\bar{t_1} + 3 * std(t_1), (\bar{t_2} + 3 * std(t_2)) \\
 b) (p_{2_{t_1}}, p_{2_{t_2}}) &= (\bar{t_1} - 3 * std(t_1), (\bar{t_2} + 3 * std(t_2)) \\
 c) (p_{3_{t_1}}, p_{3_{t_2}}) &= (\bar{t_1} - 3 * std(t_1), (\bar{t_2} - 3 * std(t_2)) \\
 d) (p_{4_{t_1}}, p_{4_{t_2}}) &= (\bar{t_1} + 3 * std(t_1), (\bar{t_2} - 3 * std(t_2))
 \end{aligned}$$

För att avgöra vilken av dessa punkter som skall representeras i de olika konfidenspolynomen beräknas avståndet mellan varje rektangelns hörn och origo. Den punkten med minst, respektive störst, distans kommer att inkluderas i inre, respektive yttre polynomet. En svaghet med dessa beräkningar, som också tas upp i (Johansson et al., (2020)) är att data ofta antas vara normalfördelade.

När standardavvikelsen används för att beräkna konfidensintervall antas data vara normalfördelade, vilket inte är fallet för t_1 och t_2 i varje delmängd. Detta har diskuterats inom projektet vid ett flertal tillfällen och igen bra lösning på problemet har ännu identifierats.

Två frågeställningar som behöver besvaras innan punkterna för polynom Anpassningen kan beräknas är

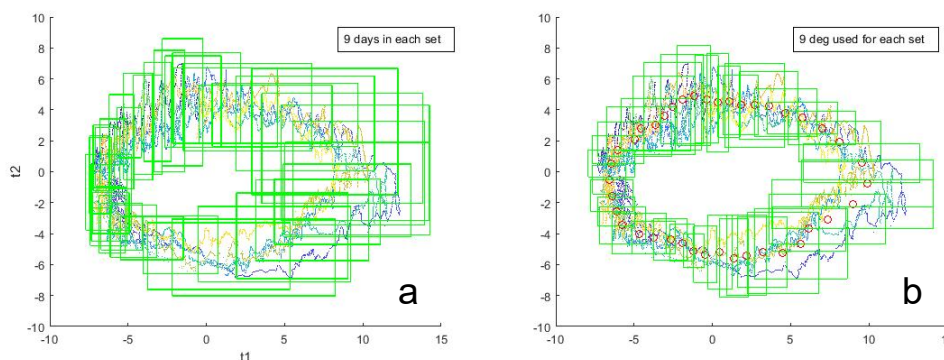
1. Hur många grupper skall året delas upp i (hur många punkter skall användas för att definiera polynomen)
2. Hur skall grupperna definieras?

Antalet delmängder som data skall delas upp i avgör mängden observationer i varje delmängd. I praktiken betyder det hur lång tid som skall inkluderas i varje delmängd. Genom en iterativ process har flera delmängdsstorlekar undersökts. Initialt testades delmängder baserat på dagar, kalenderveckor och kalendermånader.

Om varje dag på året definierar en grupp bildas totalt 365 eller 366 grupper. Medan om grupperingen baseras på veckor motsvarar det 52 eller 53 och för månader 12 grupper. Efter testning och utvärdering har vi valt att använda oss av 40 grupper. Denna slutsats har nåtts genom att testning med 52 grupper blivit en för stor mängd med stora individuella skillnader mellan närliggande grupperingar.

Att antalet delmängder valts till 40 är inte matematisk optimerat utan framtaget genom testning av flertalet delmängdsstorlekar. Dessa resultat har sedan studerats och jämförts för att hitta den mängd som ser ut att ha den bästa anpassningen till datamängden.

Den andra frågan behandlar hur grupperna skall definieras. De två huvudsakliga spår som utforskats är dels baserat på vilken dag på året som observationen är rapporterad, dels baserat på den beräknade vinkeln från origo till observationen i t_1 & t_2 rymden. Om alternativ ett används kommer storlekarna på alla delmängder vara så när lika stora, medan det inte behöver vara fallet för alternativ två. I Figur 40 presenteras dessa två olika alternativ. Figur 40a presenterar alternativ ett och Figur 40b alternativ två.

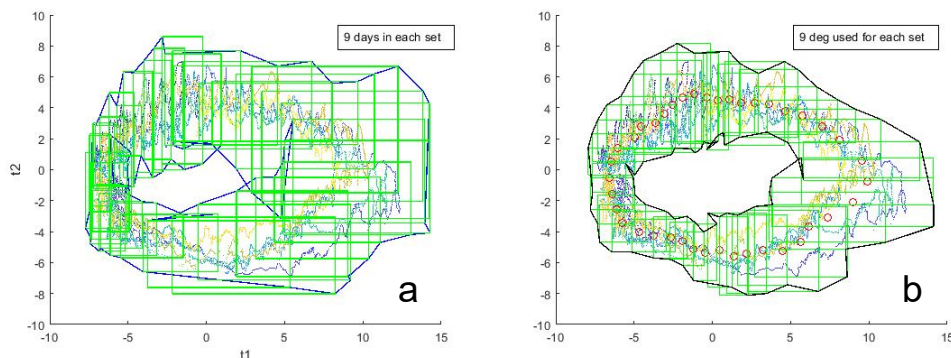


Figur 40. Två olika varianter på gruppering av data. I Figur 40a är grupperna baserat på dag på året medan i Figur 40b baseras grupperna på observationens vinkel. De gröna rektanglarna baseras på de beräknade medelvärdeskoordinaterna samt $3 \cdot \text{standardavvikelse}$ i varje grupp.

De gröna rektanglars hörnen i Figur 40 har beräknats enligt Ekvation 3.

Notera att rektanglarna i Figur 40a ofta är större än vad det är i Figur 40b. Detta beror på att då observationer delats in efter kalendertid medför detta att spridningen på observationen vid en viss tid mellan åren kan skilja sig relativt mycket, vilket kan orsaka en stor standardavvikelse. I Figur 40b, där vinkeln definierar grupperna, är punkterna i varje delmängd mer samlad vilket också resulterar i mindre standardavvikelser.

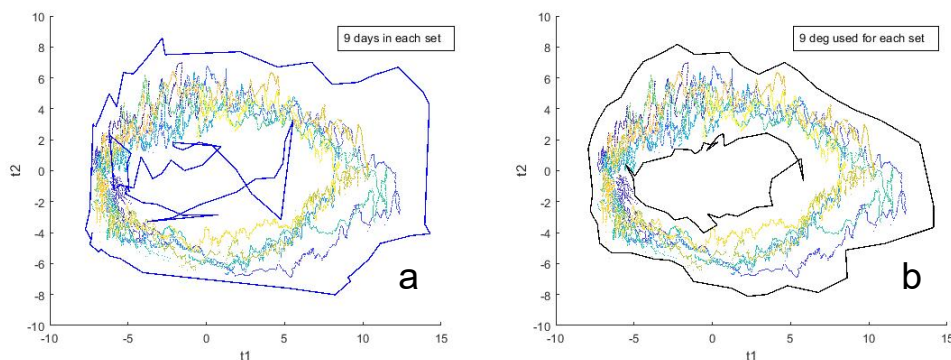
Figur 41 presenterar samma bilder som Figur 40 men därtill har även två svarta linjer dragits mellan de hörn i varje delmängd som har minst respektive störst distans till origo.



Figur 41. Samma bilder som Figur 40 men med svarta linjer dragna mellan rektanglarnas hörn som har minst respektive längst distans till origo.

Notera att den inre svarta linjen korsar sig själv vid ett antal tillfällen. Detta är ett fenomen som uppkommer när medelvärdeskoordinaterna mellan närliggande grupperingar korsar en axel. Detta blir tydligt när rektanglarnas centrum rör sig från kvadrant 4 till kvadrant 1 i Figur 41a. Dessa rektanglar har en relativt stor utbredning i t_2 vilket har en stor påverkan på distansen till origo. För de rektanglar som har ett litet negativt medelvärde i t_2 förväntas det hörnet där standardavvikelse adderas vara närmast origo (Ekvation 3b), medan när medelvärdet korsar x-axel och blir positivt kommer den närmsta punkten istället höra till när standardavvikelse subtraherats från medelvärdet (Ekvation 3c). Samma fenomen sker även för de yttre gränserna, men resulterar då i att distansen mellan två efterföljande punkter blir avsevärt längre än andra. Med detta sagt är påverkan på den yttre gränsdragningen relativt liten då förhållandet mellan dessa punkter följer samma riktning som förhållandet mellan gruppernas medelvärde.

I Figur 42 presenteras samma figurer men endast med linjer som sammankopplar rektanglarnas hörn med den minsta respektive den kortaste distansen. Alltså punkterna som skall användas för att definiera polynomen för begränsningarna.



Figur 42. Samma plottar som i Figur 41, men endast med de linjer som kopplar samman rektanglarnas hörn med minst och störst distans till origo

Notera att Figur 42b ser ut att ha en betydligt bättre anpassning till svärmen av observationer i bilderna. Slutsatsen är att grupperna baserad på vinkeln kan vara att föredra vid denna typ av gränsdragningen

Nästa steg i att beräkna begränsningspolynom är att använda dessa punkter för att göra en polynomanpassning.

Det finns olika sätt att anpassa ett antal punkter till ett polynom. De två varianter denna studie utvärderat är minsta kvadraters ellipsanpassning samt Savitzky-Golay filtrering.

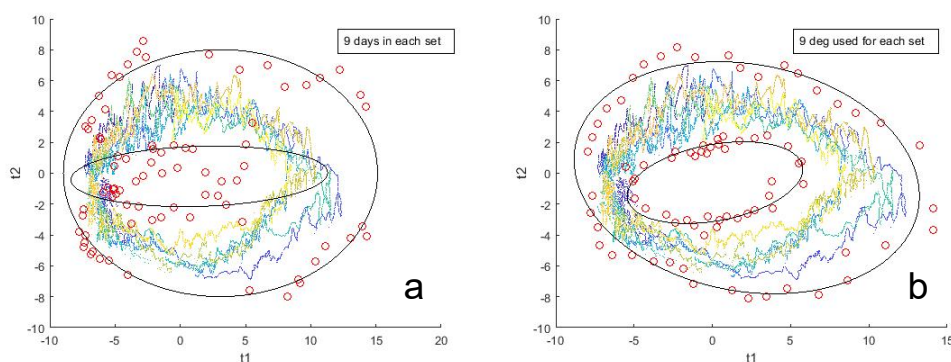
- Ellipsanpassning – minsta kvadraternas ellipsanpassning

Funktionen anpassar en ellips baserat på ett antal punkter. Anpassningen minimerar summan av kvadraterna för de ortogonala distanserna från ellipsen till det specificerade punkterna. Lär mer om ellipsanpassning i Ahn et al., (2000)

- Savitzky-Golay-filter

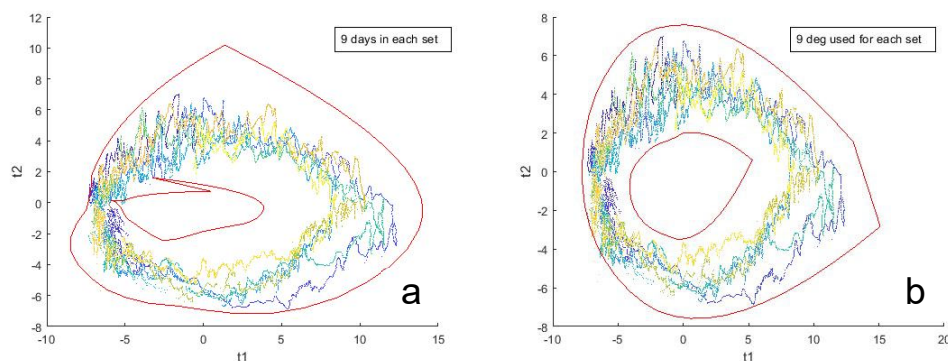
Savitzky-Golay-filtrering används för utjämning (smoothing) av data utan att data skall förlora sin trend. Den baseras på faltning av en låggradigt polynom beräknat på närliggande punkter, vilket sedan anpassas med en linjära minstakvadratmetoden (linear least squares). Läs mer om detta filter i Savitzky & Golay, (1964)

Dessa två metoder implementerades och resultaten visas i Figur 43 respektive Figur 44.



Figur 43. Samma modell som visats tidigare men med minsta kvadraternas ellipsanpassning beräknad på de punkter vilka ligger till grund för linjerna i Figur 42. Dessa punkter är representerade av röda cirklar i denna figur.

I Figur 43 presenteras resultaten av minsta kvadraternas ellipsanpassning på de punkterna som presenterats tidigare. För Figur 43a är utbredningen av den inre ellipsen i t_1 stor vilket gör att vissa delar av modellens resultat ligger innanför denna ellips. Denna ellips är därmed inte användbar vid övervakning av dammtillståndet. Samma gäller även för den inre ellipsen för Figur 43b även om resultatet för denna figur är bättre anpassad än Figur 43a. De yttre ellipserna för båda figurerna anpassat sig bra efter modellens utbredning och skulle eventuellt fungera som begränsning.



Figur 44. Samma modell som plottats tidigare men med Savitzky-Golay filtrering av punkter vilka ligger till grund för linjerna i Figur 42

Figur 44 liknar Figur 43 men gränserna har här beräknats med hjälp av Savitzky-Golayfilter. Resultaten för Figur 44b ser ut att skapa ett relativt bra konfidsensintervall för att bedöma dammtillståndet.

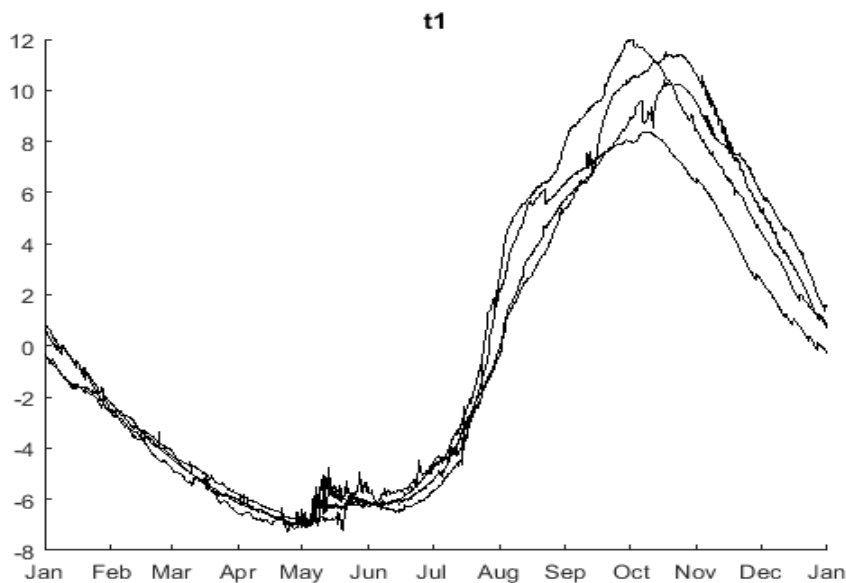
Metoden att övervaka dammtillståndet med hjälp av statistiska begränsningar i fler dimensioner har visat sig vara svårt, vilket också kan förklara den begränsade forskning som finns tillgänglig i litteraturen.

9.3.2 Konfidsensintervall för projektionsvärden (scores), tidsberoende per enskilda dimensioner

En annan metod för att övervaka dammens tillstånd är att bevaka modellernas principalkomponenter (dimensioner) var för sig. Komplexiteten i denna övervakning är mindre än vad den är med övervakningen presenterad i 9.3.1, även om metodiken liknar varandra, men med vissa avgörande skillnader.

Den stora skillnaden mot föregående metod är att föregående kapitel resulterade i förslag till polynom vilka dammens tillstånd förväntade sig ligga inom, oavsett när på året observationen är gjort. I denna, tidsberoende, metod kommer istället ett konfidsensintervall definieras utifrån vilken tid på året som observationen är gjord.

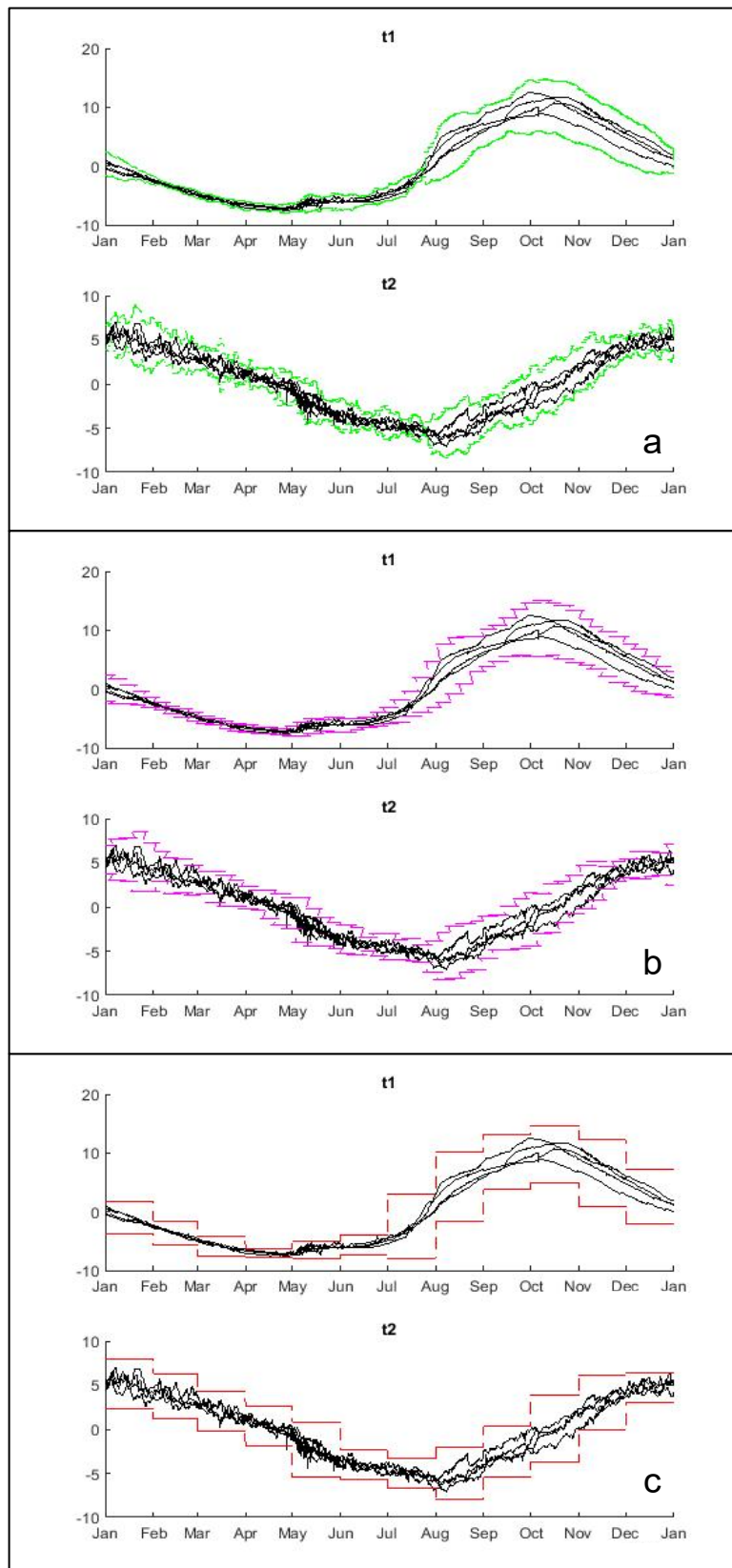
Som i föregående kapitel kommer ett antal grupper behöva definieras för att användas som grund till de konfidsensintervall som skall beräknas. De historiska data kommer sedan fördelas i dessa grupper och ligga till grund för beräkningarna av konfidsensintervallen, för respektive period. I Figur 45 exemplifieras detta med den första principalkomponenten. Figuren visar flera års data synkroniserad på årssdag. Det blir då tydligt att se likheter och skillnader mellan året. All data fördelas i grupperna vilka kommer användas för beräkning av medelvärde och en standardavvikelse, som ligger till grund för konfidsensintervallen. Liksom i föregående kapitel finns det en tydlig svaghet i användandet av standardavvikelse för konfidsensintervall för icke normalfördelade data.



Figur 45. PCA modellens första principalkomponent (t1) för fallstudie I. I denna figur har data synkroniserats och plottats årsvis i samma figur. De olika åren skiljer sig till viss del men skillnaden är förhållandevis liten om den jämförs med den årliga variationen.

PCA-modellens första principalkomponent (t1) är plottad årsvis i Figur 45. Det är enkelt att se årens skillnader och likheter. Notera att under första halvan av året (januari – juli) är kurvorna mer lika än under andra halvan av året. Detta är ett annat sätt att visualisera det vi ser i Figur 17 och som diskuteras i kapitel 8.1.1.

Den tidsberoende övervakningen behöver fördela året i ett antal grupper, vilka kommer definiera gränserna beroende på när på året observationen är rapporterad. I Figur 46 har grupper definierats på kalenderdagar, kalenderveckor och månader.



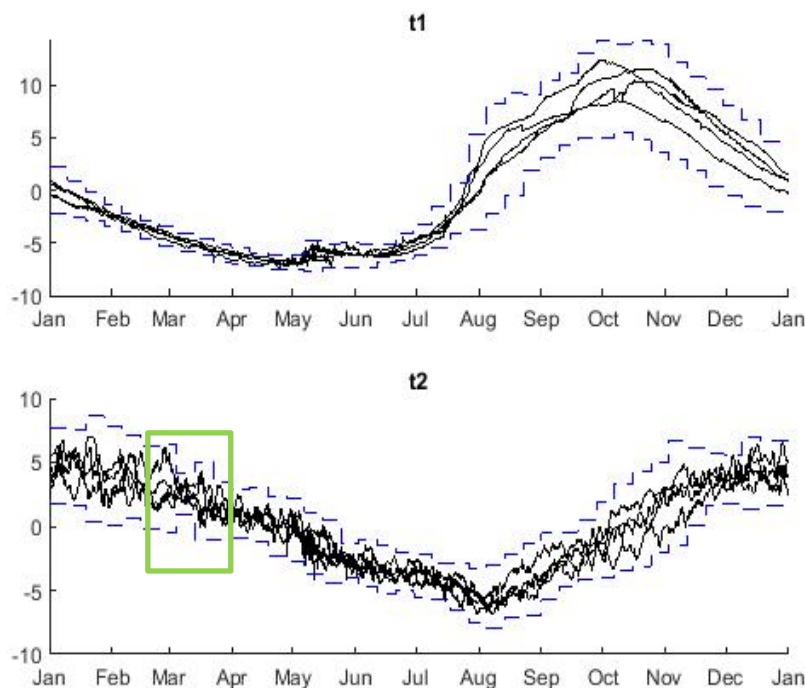
Figur 46. PCA modellens två första principalkomponenter (t1 och t2) för fallstudie I. Bild a har konfidensintervall baserat på kalenderdag, bild b på kalendervecka och bild C på kalendermånad

I Figur 46a ser vi att ju mindre data (kortare tidsintervall) som utgör varje grupp, desto mer följsam bild konfidensintervallet mot data. Konfidensintervallet i Figur 46a också mycket snäva, framförallt under årets första halva. I Figur 46b, där grupperna baserats på kalenderveckor, kan noteras att gränsdragningen i vissa tidpunkter har två olika värden. Detta beror på att olika kalenderveckor inte innehåller samma datum varje år vilket resulterar i att många datum inkluderas i olika grupper. Detta medför att kalenderveckor inte blir en användbar fördelning. Rörande gruppering på månader i Figur 46c är gränserna ofta väl tilltagna men då dessa inkluderar ett långt tidsintervall så kan den naturligt årliga variabiliteten få ett tydligt genomslag och sannolikheten att gränserna överskrids är mycket större i början och slutet av månaderna. Detta betyder att dessa grupperingar är för stora och bör vara mindre än en månad för att bygga ett bra konfidensintervall.

Genom testning av gruppstorlekar med olika tidsintervall har grupper baserats på nio dagars ansetts vara en bra fördelning av data. Vid denna bedömning har två huvudsakliga riktlinjer använts.

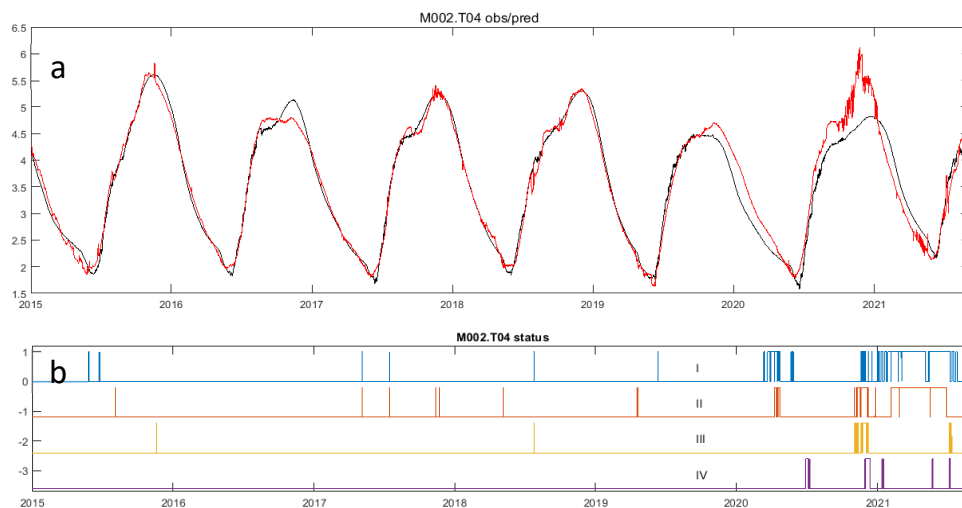
1. Gruppernas storlek måste hållas kortare än att den årliga variationens rörelse. Väntevärdet för alla observationer inom gruppen skall vara ungefär samma. Om medelvärdet av observationerna från gruppens första dygn skiljer sig avsevärt från medelvärdet för gruppens sista dygn kan den årliga variationens struktur fångats i en och samma grupp. Detta kommer resultera i mycket stora standardavvikelser för gruppen, vilket inte är önskvärt. I Figur 46c kan t1 för månaden juli studeras. Vid början av månaden beräknas medelvärdet till ungefär -5, medan i slutet av månaden är medelvärdet närmare +2. Väntevärdet i gruppen beror alltså på var inom gruppen observationen görs. Detta resulterar i mycket stora konfidensintervall, vilket inte är önskvärt.
2. För att grupperna inte skall vara för små och att korta kraftiga skillnader därmed har en stor inverkan på konfidensintervallen skall skillnaden mellan två efterföljande gruppers standardavvikelse vara liten om skillnaden av medelvärdet är litet. Detta oönskade beteende förekommer vid ett fåtal tillfällen i Figur 47 (se grön markering). Under februari för t2 kan vi se att konfidensintervallet ökar i både positiv och negativ riktning vid samma punkt, vilket visar på den stora förändringen i standardavvikelse. Detta sker även i augusti för t1, men vid det tillfället ökar även medelvärdet kraftigt vid samma tidpunkt vilket inte medför samma resultat där den undre gränsen minskar, medan den över gränsen ökar.

Notera att de svarta linjerna i Figur 46 och Figur 47, vilka presenterar modellens värden för t1 och t2 över flera år, inte är exakt samma i dessa figurer. Detta beror på att olika modeller har använts till grund för dessa plottar. Dock är modellerna nära på identiska och dessa små skillnader påverkar inte slutsatserna dragna utifrån figurerna.



Figur 47. PCA modellens två första principalkomponenter (t1 och t2) för fallstudie I. Figuren visar ett konfidensintervall (streckad blå) baserat på grupper om 9 dagar.

Vid utvärdering av denna övervakning testades detta i Matlab. Ett exempel av detta presenteras i Figur 48.



Figur 48. En modell från fallstudie I där övervakningsmetoder för t1-t4 utvärderas. I Figur 48a presenteras observerat- (svart) och predikterat värde (rött). I Figur 48b presenteras resultaten av övervakningen av t1-t4 där de romerska siffrorna berättar vilken dimension som avses (t1 = I, etc.)

Modellerna är tränade på data fram till årsskiftet 2019–2020. Data från 2020 och framåt är det som kallas verifieringsset. I Figur 48a ser vi att prediktionen följer observationen väl under träningsdelen av datat, medan flera avvikelser sker i verifikationsdelen. Dels ser vi att från november 2019 till mars 2020 är det en

relativt konstant differens mellan observerat och predikterat värde. Detta kan härledas till att den observerade temperatursignalen börjar sjunka tidigare under vintern 2020 jämfört med andra år, i förhållande till de givare som utgör X. Detta blir svårt att bygga in i en modell då detta är ett nytt förhållande mellan X och Y.

Från oktober 2020 till februari 2021 ser vi att det predikterade värdet går långt över det observerade värdet. Det predikterade värdet har dessutom en hög brusnivå jämfört med tidigare i datasetet. Detta orsakas av något beteende i X som skiljer sig från tidigare år och detta är något som behöver analyseras vidare.

Figur 48 presenteras resultaten från de dynamiska övervakningssignalerna för t1-t4. Fram till årsskiftet 2019–2020 är markeringarna små och mycket korta. Då denna data utgjort grunden för beräkningarna av konfidensintervallen så är detta beteende väntat. Under 2020 ökar frekvensen av markeringar i samtliga övervakningsmetoder. Detta beteende är inte önskvärt och kan grundas i en rad orsaker vilka behöver analyseras närmare.

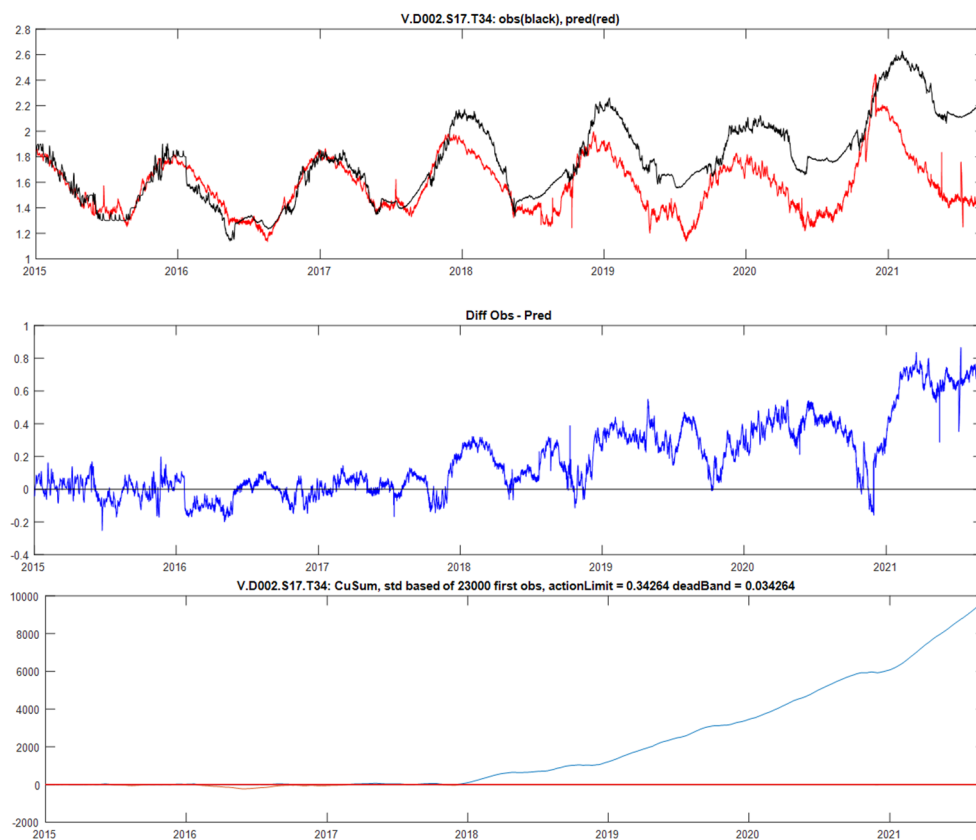
9.3.3 Metod för att minimera falska larm

De konfidensintervall som används baseras på procentuell gränssättning, antingen baserat på standardavvikelse alternativt empiriska gränser. Att använda sig av procentuella gränser medför att en delmängd av de värden som använts vid beräkningarna skulle överstiga gränsvärdena. Vid användandet av tre standardavvikelser skulle, vid en normalfördelad datamängd, statistiskt sett 3 av 1000 värden att avvika från konfidensintervallet. Då 4–7 parametrar övervakas per modell och en damm är instrumenterad med över 50 givare ger detta minst 200, och upp mot 500, signaler som övervakas multivariat. Detta skulle medföra att om enskilda signalövervakningar skulle markeras, med vad som kan betraktas som falska larm, mycket ofta skulle dessa markeringar med tiden ignoreras.

För att minska mängden falska markerade avvikelser kan tidsgränser implementeras, vilket ännu inte har genomförts. Dessa tidsgränser är räknare som håller reda på hur lång tid som en signal har varit utanför sitt konfidensintervall. Då ett avvikande beteende förväntas vara en långsam process skulle en tidsbegränsning således minska mängden korta och felaktiga avvikelser och istället hitta avvikelser som pågår över en längre tid och således kan tyda på faktiska förändringar i antingen givarens prestanda eller signalens beteende.

9.3.4 Avvikelsesummering, CuSum

Om små skift i en signals medelvärde, 1 – 4 standardavvikelser, skall detekteras är CuSum-metoden (**K**umulativ **S**umma) effektiv, Montgomery, (2013). Normalt beräknas två summor, en för positiva och en för negativa avvikelser från medelvärdet. I projektet har vi använt CuSum för att övervaka skillnaden mellan observerat och predikterat signalvärde, som ju har ett förväntat medelvärde = 0.0. Se Figur 49 för ett exempel.



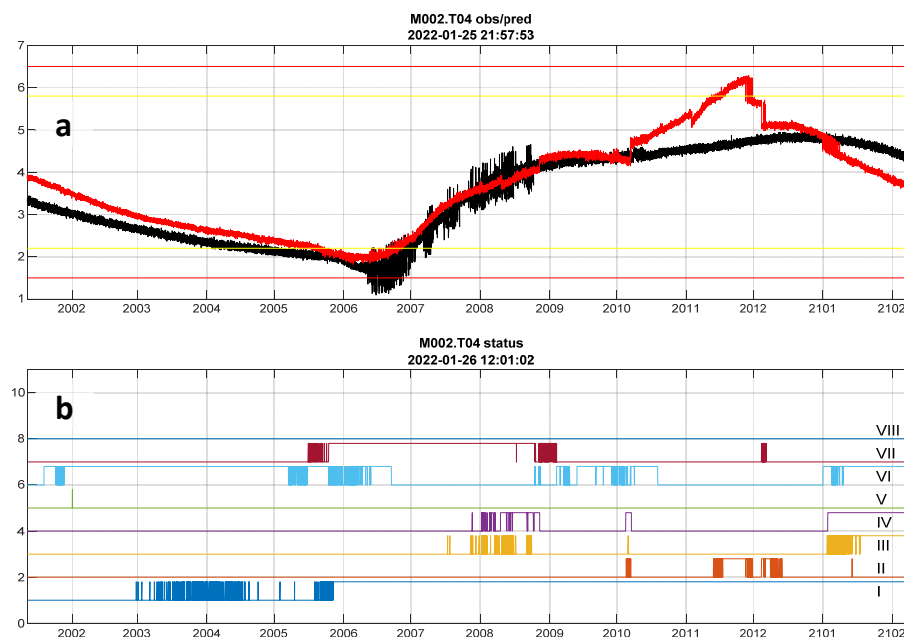
Figur 49. Exempel på CuSum. Den översta bilden visar observerat och predikerat signalvärde för modellen för T34. Den mellersta visar skillnaden mellan observerat och predikerat värde. Den undre bilden visar CuSum för samma period. CuSum ligger still fram till början av 2016, där den indikerar att skillnaden avviker från 0.0.

9.4 ÖVERGRIPANDE VISUALISERING AV ENSKILDA GIVARE RESPEKTIVE DAMMTILLSTÅND

Ett antal av övervakningsmetoderna har implementerats för testning och utvärdering vilka använder prediktioner från de multivariata modellerna för de två fallstudierna. De övervakningsmetoder som implementerats finns listade i Tabell 9 och ett exempel de första resultaten från den implementerade lösningen presenteras i Figur 50.

Tabell 9. Lista över implementerade övervakningsmetoder

Övervakad variabel	Representeras i Figur 50
DModX	I
HotellingsT2	II
Residual Ypred	III
Övre CuSum	IV
Nedre Cusum	V
t1	VI
t2	VII
t3	VIII



Figur 50. Utdrag från övervakning av en utvald modell från fallstudie I. X-axeln presenterar tid [ååmm].

Figur 50 sträcker sig över drygt ett års data. I Figur 50a presenteras det observerade värdet (svart) och det predikterade värdet (rött), Figur 50b visar resultaten från övervakningsmetoderna. Dessa är markerade med romerska siffror för att den skall gå att härleda vilket test som hör till vilken signal, vilket presenteras i Tabell 9.

Den predikterade signalen i Figur 50a följer det observerade värdet väl större delen av året. Mellan oktober 2020 till februari 2021 ser vi att de predikterade värdet går långt över det observerade värdet. Detta orsakas av något beteende i X som skiljer sig från tidigare år och detta är något som behöver analyseras vidare.

Figur 50b presenteras resultaten från de dynamiska övervakningssignalerna listade i Tabell 9. Frekvensen av markeringar i Figur 50b är hög och sker under längre perioder. Detta beteende är inte önskvärt och kan grundas i en rad orsaker vilka behöver analyseras närmare.

10 Implementation

Den här implementationen, version 2, bygger på version 1, som gjordes i det tidigare projektet och rapporterades i motsvarande kapitel i Jacobson et al., (2021), med tillägg av den multivariata ML-delen. Programmen är användbara både offline och online, och enkelt flyttbara till olika datorer. Allt som är dator- och användningsspecifikt administreras i databastabeller och konfigureringsfiler.

10.1 ANVÄNDA METODER FÖR ATT REALISERA OFFLINE- OCH ONLINE-LÖSNINGEN

Metoderna från 4.1 som utvecklats i "MATLAB," (2021) har implementerats och driftsatts för användning i realtid. Programvaran är skriven i språket "C-sharp," (2021) med hjälp av utvecklingsmiljön "Microsoft Visual Studio," (2021) version 2019 och använder i huvudsak ".NET Framework," (2020) version 4.0. Detta är en stabil miljö som används för de flesta realtidsapplikationerna hos IVL. Den första implementationen var förberedd för det fortsatta projektet med multivariat signalövervakning vilket pågår sedan våren 2021.

Detta projekt återanvänder vissa mjukvarudelar från två tidigare IVL-projekt. I projektet Miljöväder (se "Miljöväder ska visualisera utsläpp och buller," (2020)) implementerade vi en liknande applikation med realtidsexekvering av flera multivariata modeller, datahämtning, datalagring och automatisk modelluppdatering. I ett annat projekt med Sandviken Energi kördes olika modeller med ett omfattande användargränssnitt byggt med den kommersiella mjukvaran "GENESIS64," (2021) från "Iconics," (2021). I båda fallen, liksom i många andra, använde vi OPC (se "What is OPC?," (2021)) (standard för industriell signalöverföring) för datakommunikation och presentation.

För att undvika att bygga in hårdkodade data i programmen läggs alla mjuka data i databaser eller konfigureringsfiler. Vi använder tabeller för administrativ information såsom signal- och modelldefinitioner i en databas, "SQL Server Express," (2020), från Microsoft. För historiska signaldata använder IVL historikdatabasen "Hyper Historian," (2021), medan Vattenfall och Statkraft använder databasen PI från "OSIsoft," (2020). Med data i databaser och definitionsfiler är det enkelt att flytta applikationen till andra användare: man ändrar bara signal- och modelldefinitionerna.

Programutvecklingen har gjorts i två versioner:

- Version 1, endast för univariat signalövervakning, med metoder enligt 4.1, fram till år 2020.
- Version 2, applikationen utökad med multivariata delar, vilket utvecklats under 2021.

Båda versionerna kan användas offline eller online, användningen styrs via konfigureringsfilerna.

10.2 SYSTEMTEKNISKA KRAV PÅ MJUKVARULÖSNINGEN ONLINE

Vattenkraftsproduktion är kritisk infrastruktur vilket ställer krav på företagets IT-infrastruktur och informationssäkerhetsarbete. Detta påverkar de krav anläggningsägarnas IT-funktioner ställer på mjukvarulösningar och val av arkitektur. Dessa krav har även påverkat det här projektet. Ett krav var att absolut inte använda någon molntjänst för den lösning vi utvecklat. Frågan prövades men av olika skäl var beslutet att en sådan inte fick användas. Ett annat önskemål var en enkel och transparent datatransport mellan det projektet utvecklat och de befintliga system kraftbolagen använder för att lagra och visa data från givare för dammsäkerhet. Vidare är metadata för signaler och modellresultat från och till den utvecklade mjukvarulösningen anonymiserade. Vidare är tillgången till de system där IVL:s lösning installerats ytterst begränsad för IVL:s del och sker endast genom kraftföretagens personal.

10.3 ÖVERGRIPANDE BESKRIVNING AV IMPLEMENTERAD LÖSNING

Ett blockschema för implementationens komponenter visas i Figur 51.

Inom ramen för projektet är 4 program utvecklade:

- DSPopulateSignals (hantering av signaldefinitioner och gränsvärden),
- DSModelMaker (ML-modeller),
- DSWriteData (för uppspelning av historiska data offline), samt
- DSWatchDog (det egentliga övervakningsprogrammet).

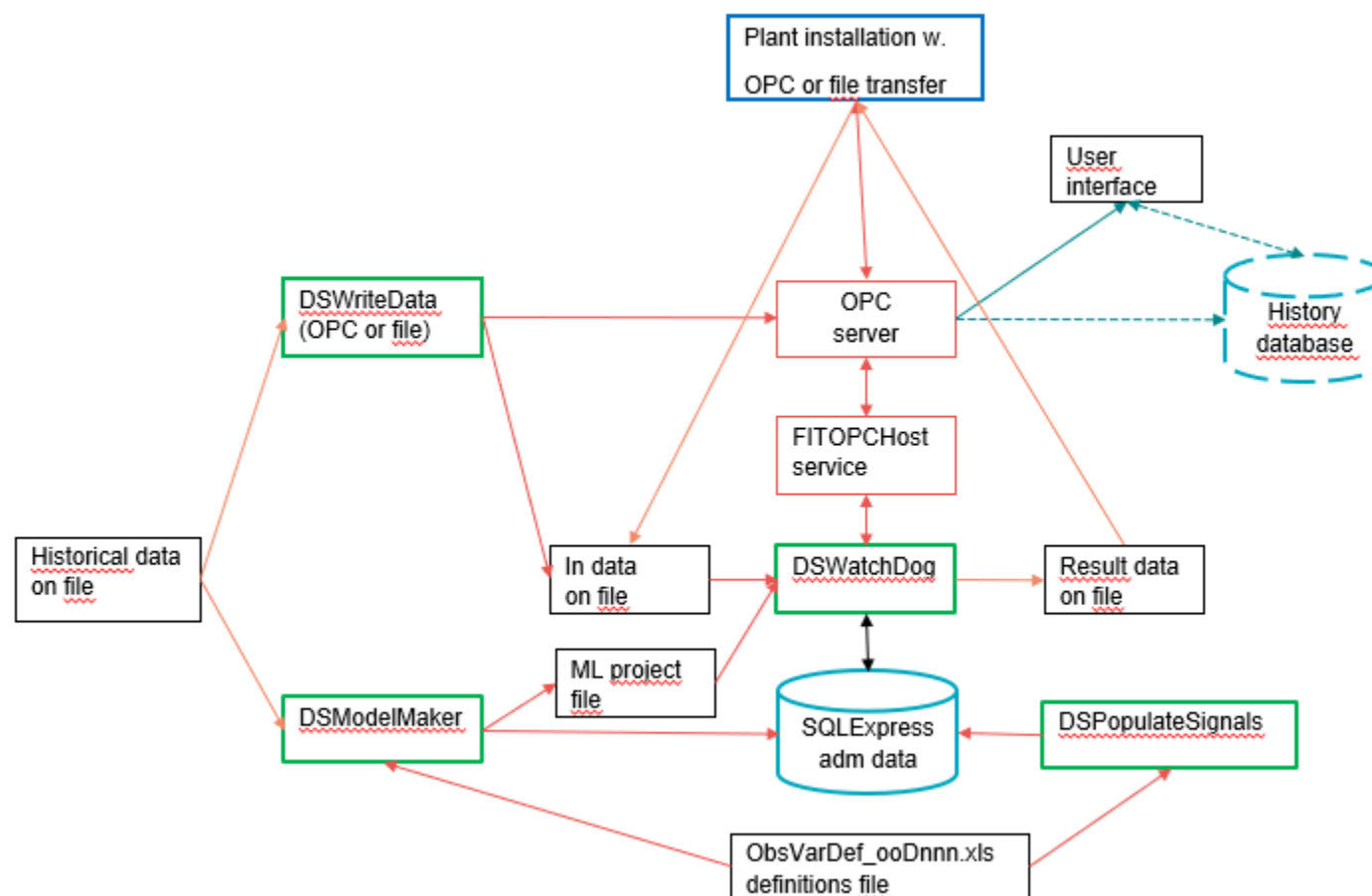
Dataöverföringen med dammanläggningen kan ske med OPC eller textfiler:

- Vid OPC sker datautbytet av både in- och utdata via en OPC-server.
- Vid filkommunikation sker utbytet med textfiler, en fil per ny observation av alla signaler. In-data läggs på en in-katalog, och utdata från DSWatchDog läggs in en ut-katalog.

En databas är inlagd i SQL Express för att hantera nödvändiga exekveringsdata såsom signal- och modellgränsvärden. Databasen har 4 tabeller: ModellLimits, Models, NodeServers och Signals.

För att enkelt kunna flytta applikationen mellan olika installationer är alla installationsspecifika egenskaper, såsom olika signaler och modeller, samlade i en Excelfil, ObsVarDef.

Varje komponent beskriva mer detaljerat i texten nedan.



Figur 51. Blockschema över programlösningens komponenter. De program som utvecklats i projektet är grönmärkade

10.4 DEFINITIONSFILEN OBSVARDEF

Data till tabellen Signals och för modellbyggandet ligger i en Excelfil, ObsVarDef.xlsx. Filen har tre blad:

- VarDef - en lista med en rad per signal: dess olika namn, gränsvärden och vilka modeller de ingår i. För varje modell, som primärt utvecklats i Simca, finns det en kolumn där man markerar vilka variabler som skall vara X och vilka som skall vara Y.
- ObsDef - anger vilka observationer i träningsdata som ingår i respektive modell.
- Models - listar alla definierade modeller, typ (PCA eller PLS), samt om de är aktiva eller inte.

10.5 PROGRAM

Alla projektets program är skrivna i C# i Visual Studio 2019. Funktionsbibliotek (DLLer) från "Products," (2022) användes för att det har använts i tidigare projekt och kod kunde delvis återanvändas. All programkod är dokumenterad inline. Programmet Doxygen, Dimitri, (2022), används för att generera programdokumentationen i html-format.

Utvecklingen sker på en virtuell dator hos IVL. Resulterande binärfiler kopieras till de datorer där projektet skall installeras. Programparametrar som är specifika för respektive måldator specificeras i konfigureringsfiler. På så sätt kan samma program exekveras på alla måldatorer utan omkompilering.

10.5.1 DSPopulateSignals

Programmet används för att populera tabellen signals från projektets ObsVarDef_ooDnnn.xlsx-fil (oo är dammägare {VF, SK}, Dnnn är dammnummer).

Projektnamnet och namnet på ObsVarDef-filen specificeras i programmets konfigureringsfil, Figur 52.

Name	Type	Scope	Value
projectName	string	User	VD002
fullOVDName	string	User	D:\Eget\Projects\2021\IVL\Dammsäk3\Data\VF\ObsVarDef_VFD002.xlsx

Figur 52. Exempel på konfigureringsfil för DSPopulateSignals.

10.5.2 DSModelMaker

Det här programmet används för att skapa implementationens ML-projektfil för användning för exekvering av DSWatchDog. Filen innehåller alla modeller som specificerats i ObsVarDef-filen.

Inställningarna specificeras i dess konfigureringsfil, DSModelMaker.config, Figur 53. DataFileName anger den datafil med preprocessade data som skall användas vid modellgenereringen.

Name	Type	Scope	Value
DataRootDir	string	▼ User	▼ C:\Eget\Projects\2021\Dammsäk3\Data\VF\
OVDFileName	string	▼ User	▼ ObsVarDef_VFD002.xlsx
DataFileName	string	▼ User	▼ DS_VF_PP.txt
ProjectName	string	▼ User	▼ VD002

Figur 53. Programinställningar för DSModelMaker.

DSModelMaker är ett konsolprogram och exekveras bara då nya data föreligger eller då andra modeller skall genereras, Figur 54. Modelleringsresultaten läggs förutom i ML-projektfilen också i tabellen Models, som håller en lista över modellresultaten och huruvida varje modell är aktiv eller ej.

```

Done reading dataset
Done reading model definitions
Now modeling for Mdl01
Now modeling for Mdl02
Now modeling for Mdl03
Now modeling for Mdl04
Now modeling for Mdl05
Now modeling for Mdl06
Now modeling for Mdl07
Now modeling for Mdl08
Now modeling for Mdl09
Now modeling for Mdl10
Now modeling for Mdl11
DSModelMaker: 21-12-22 13:19:05 Project VD002 saved with 12 models.
Press <Enter> to exit

```

Figur 54. Skärmdump från en körning av DSModelMaker.

10.5.3 DSWriteData

DSWriteData används för att spela upp historiska data för DSWatchDog. Det används för programutveckling och testning. Indata är rådata så som de anlänt från dammägarna.

Programinställningarna specificeras i DSWriteData.config, Figur 55. Inställningen IsWriteToFile anger, om sann, att data skall skrivas som csv-filer. Är värdet falskt skrivs data till OPC-server på specificerad opcNode. Parametern updateRate anger tidsintervallet i [ms] mellan varje observation och gör det möjligt att köra DSWatchDog kort intervall mellan exekveringarna.

Name	Type	Scope	Value
updateRate	int	▼ User	▼ 5000
dataFileName	string	▼ User	▼ C:\Eget\Projects\2021\Dammsäk3\Data\VF\DS_VF_raw20_21.txt
outFileDir	string	▼ User	▼ C:\Eget\Projects\2021\Dammsäk3\Data\VF\in\
isWriteToFile	bool	▼ User	▼ True
opcNode	string	▼ User	▼ 192.168.11.67
opcDatetimeTag	string	▼ User	▼ V.D002.DateTime

Figur 55. Programinställningar för DSWriteData.

Programmet är ett konsolprogram som skriver en ny observation varje updateRate [ms], Figur 56.

```

Mounting dataset
Writing data to Data files in directory D:\Eget\Projects\2021\IVL\Dammsäk3\Data\VF\in\
Press <Enter> to abort writing data
time is 2021-10-26 02:39:48
time is 2021-10-26 02:39:53
time is 2021-10-26 02:39:58
time is 2021-10-26 02:40:03
time is 2021-10-26 02:40:08
time is 2021-10-26 02:40:13

```

Figur 56. Skärmdump från körning med DSWriteData, tidsangivelserna avser tidpunkten då data skrivs, inte tiden som var då observationens data registrerades.

10.5.4 DSWatchDog

DSWatchDog är ett konsolprogram där det mesta sker kring signalövervakning och modellexekvering sker. Programmets metod ExecuteSignals() är programmets huvudmetod: här sker signalvalideringen i två steg, först univariat och sedan multivariat. Om ett signalvärde förkastas i första steget, ersätts det med bortfall, NaN, inför datas användning i det multivariata steget (det univariata steget är en förbehandling inför det multivariata).

Programmet kommunicerar via filöverföring eller OPC beroende på dess konfigurerings. Programmet utför all bearbetning via ett klassbibliotek, SignalMonitor, via metदानrop, Figur 57.

```

#region public methods
/// <summary> The method creates a list of all signals in the Signals table, rel ...
1 reference
public int MountSignals(string projectName, string dbServerName, bool isUseConsole)...

/// <summary> CuSum setup
1 reference
public int SetupCuSum(FITModelExecList fitModelExec)...

/// <summary> Stes up model limit, reads from DB
1 reference
public int SetupModelLimits(FITModelExecList fitModelExec)...

/// <summary> Setup of OPC connections \todo Init valueLG to NaN on OPC server
1 reference
public int SetupOPC()...

///// <summary>
/// Called to invoke the monitoring delegate ...
1 reference
public int StartExecuteSignals(string projectName, int updateRate, string fileRepository,
    bool isReadFromFile, bool isWriteToOPC, bool isWriteToFile,
    SrvRefFITOPC.FITOPCSvcLibClient[] fitOPCVector, FITModelExecList fitModelExec)...

/// <summary> Used to stop the monitoring delegate
1 reference
public int StopExecuteSignals()...

#endregion

```

Figur 57. Publika metoder i SignalMonitor vilka görs från DSWatchDog. De fyra översta görs direkt då programmet startar. Metoden StartExecuteSignals() anropar SignalMonitors privata metod ExecuteSignals()

SignalMonitor har också 3 privata metoder. ExecuteSignals() är hjärtat i applikationen och redovisas nedan, GetNewSignals() och PutNewSignals() är metoderna som hanterar all kommunikation, både OPC och med filer, Figur 58.

```

#region private Methods

/// <summary> Signal monitor executer, updates data every updateRate [ms] as lon ...
1 reference
private int ExecuteSignals(SrvRefFITOPC.FITOPCSvcLibClient[] fitOPCVector)[...]

/// <summary> Gets new data to be monitored from file or OPC when read from file ...
1 reference
private int GetNewSignals(bool isReadFromFile)[...]

/// <summary> Puts output data to OPC and/or text file, once per iteration when ...
1 reference
private int PutNewSignals(bool isWriteToOPC, bool isWriteToFile, bool isReadFromFile)[...]

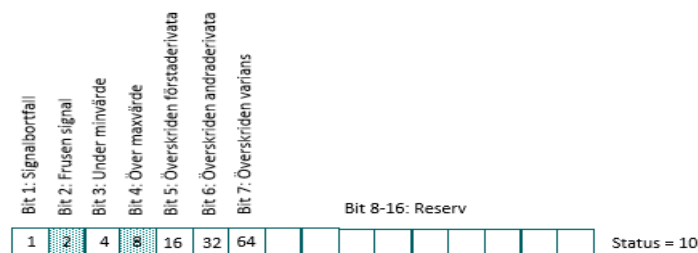
#endregion

```

Figur 58. ExecuteSignals()

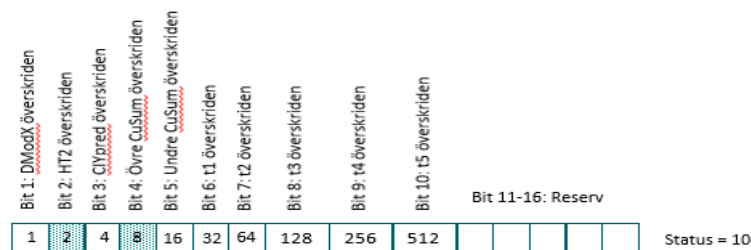
Metoden ExecuteSignals() anropas internt från SignalMonitor. Metoden exekverats repetitivt i egen prioriterad tråd med intervallet updateRate [ms] tills den avslutas via StopExecuteSignals().

Metodens exekvering börjar via GetNewSignals() som läser in en observation från fil eller OPC. Nästa steg är den univariata valideringen av observationens enskilda värden. Det är f.n. 7 olika tester som görs. Om något kriterium är överskridet ersätts datavärdet med bortfall (NaN) inför de multivariata beräkningarna. Resultaten samlas bit-vis i ett 16-bitars statusord, Figur 59.



Figur 59. Det univariata statusordet, för närvarande används bit 1–7. På bilden markeras signalen som frusen och större än accepterat maxvärde.

Efter att observationens alla värden är validerade, kanske med några värden utbytta till NaN, följer den multivariata valideringen. De aktiva modellerna exekveras och resultatet valideras mot dess utfallsgränser, Figur 60.



Figur 60. Det multivariata statusordet, bit 1–10 används.

Metodexekveringen avslutas med att resultaten exporteras till filer och/eller OPC via PutNewSignals(). Därefter lägger sig tråden i sovläge i updateRate [ms] tills den vaknar och kör ett nytt varv.

Programinställningarna för DSWatchDog görs i filen DSWatchDog.exe.config, Figur 61.

Name	Type	Scope	Value
updateRate	int	User	5000
isReadFromFile	bool	User	True
isWriteToOPC	bool	User	False
isWriteToFile	bool	User	True
fileRepository	string	User	C:\Eget\Projects\2021\Dammsäk3\Data\VF\
dbServerName	string	User	localhost\SQLEXPRESS
isUseConsole	bool	User	True
isReadFromOPC	bool	User	False
projectName	string	User	VD002
fitProjectFileNa...	string	User	C:\Eget\Projects\2021\Dammsäk3\Programs\DSModelMaker\DSModelMaker\bin\Debug\VD002.fit
fitLogFileName	string	User	C:\Eget\Projects\2021\Dammsäk3\Programs\DSWatchDog\VD002Exec.log
signLevel	float	User	0.99

Figur 61. Programinställningar för DSWatchDog

En mer utförlig beskrivning av inställningarna i Figur 61 är:

- *updateRate* [ms] anger hur ofta programmet skall hämta nya data. För utvecklingsversionen är den ofta 5000 [ms]. För on-line-versionen är *updateRate* = 300000 [ms] (5 min).
- *isReadFromFile* är sann om data skall läsas från fil, annars läses data från OPC-server
- *isWriteToOPC* är sann om resultaten skall skrivas till OPC-servern
- *isWriteToFile* är sann om resultaten skall skrivas till fil
- *fileRepository* anger sökväg till huvudkatalogen för dataöverföring
- *dbServerName* anger den databasserver där tabellen Signals finns
- *isUseConsole* är falsk om ingen utskrift skall ske till programfönstret.
- *projectName* anger projektets namn, vilket möjliggör att data för flera projekt/dammar kan ligga i samma tabell
- *fitProjectFileName* anger namnet på projektets ML-fil
- *fitLogFileName* är den fil till vilken programmeddelanden loggas
- *signLevel* är den signifikansnivå som används för statistiska larmgränser, för närvarande 99%, bör nog ökas till 99.7%.

Programmet kör i ett DOS-fönster där resultaten från varje exekvering av *ExecuteSignals()* skrivs, Figur 62.

```

serverName: localhost\SQLEXPRESS dbName: FITExec
MountSignals: mounted 90 signals
There are 4 models mounted.
Press <Enter> to stop model execution
SignalMonitor.ExecuteSignals: Limits exceeded for V.D002.S02.T03 status = 1
SignalMonitor.ExecuteSignals: Limits exceeded for V.D002.S12.T23 status = 1
SignalMonitor.ExecuteSignals: Limits exceeded for V.D002.S12.T24 status = 1
SignalMonitor.ExecuteSignals: Limits exceeded for V.D002.S13.T25 status = 1
SignalMonitor.ExecuteSignals: Limits exceeded for V.D002.S16.T31 status = 1
SignalMonitor.ExecuteSignals: Limits exceeded for V.D002.S21.T41 status = 1
SignalMonitor.ExecuteSignals: Limits exceeded for V.D002.S22.T43 status = 1
SignalMonitor.ExecuteSignals: Limits exceeded for V.D002.S23.T45 status = 1
SignalMonitor.ExecuteSignals: Limits exceeded for V.D002.S24.T47 status = 1
SignalMonitor.ExecuteSignals: Limits exceeded for V.D002.S25.T49 status = 1
SignalMonitor.ExecuteSignals: Limits exceeded for V.D002.S26.T51 status = 1
SignalMonitor.ExecuteSignals: Limits exceeded for V.D002.S27.T53 status = 1
SignalMonitor.ExecuteSignals: Limits exceeded for V.D002.S28.T55 status = 1
SignalMonitor.ExecuteSignals: Mdl01 DModX = 1.27 t1 = -0.10 t2 = 4.62 t3 = -1.98 HT2 = 2.26
SignalMonitor.ExecuteSignals: Mdl02 T04.pred = 4.0644 DModX = 0.96 t1 = 1.22 t2 = 4.55 t3 = -1.01 t4 = -0.89 HT2 = 3.40
SignalMonitor.ExecuteSignals: Mdl03 T34.pred = 1.7986 DModX = 0.80 t1 = 4.96 t2 = 3.14 t3 = -0.95 t4 = -1.97 HT2 = 5.88
SignalMonitor.ExecuteSignals: Mdl04 T60.pred = 1.9374 DModX = 1.25 t1 = -0.87 t2 = -3.18 t3 = 1.49 HT2 = 2.65
211026 02:45:20 New data written, 2015-01-01 00:00:00 sumExLimits = 13
SignalMonitor.ExecuteSignals: Limits exceeded for V.D002.S02.T03 status = 1
SignalMonitor.ExecuteSignals: Limits exceeded for V.D002.S12.T23 status = 1
SignalMonitor.ExecuteSignals: Limits exceeded for V.D002.S12.T24 status = 1
SignalMonitor.ExecuteSignals: Limits exceeded for V.D002.S13.T25 status = 1
SignalMonitor.ExecuteSignals: Limits exceeded for V.D002.S16.T31 status = 1
SignalMonitor.ExecuteSignals: Limits exceeded for V.D002.S21.T41 status = 1
SignalMonitor.ExecuteSignals: Limits exceeded for V.D002.S22.T43 status = 1
SignalMonitor.ExecuteSignals: Limits exceeded for V.D002.S23.T45 status = 1
SignalMonitor.ExecuteSignals: Limits exceeded for V.D002.S24.T47 status = 1
SignalMonitor.ExecuteSignals: Limits exceeded for V.D002.S25.T49 status = 1
SignalMonitor.ExecuteSignals: Limits exceeded for V.D002.S26.T51 status = 1
SignalMonitor.ExecuteSignals: Limits exceeded for V.D002.S27.T53 status = 1
SignalMonitor.ExecuteSignals: Limits exceeded for V.D002.S28.T55 status = 1
SignalMonitor.ExecuteSignals: Mdl01 DModX = 1.27 t1 = -0.10 t2 = 4.62 t3 = -1.98 HT2 = 2.26
SignalMonitor.ExecuteSignals: Mdl02 T04.pred = 4.0643 DModX = 0.96 t1 = 1.22 t2 = 4.55 t3 = -1.01 t4 = -0.89 HT2 = 3.40
SignalMonitor.ExecuteSignals: Mdl03 T34.pred = 1.7986 DModX = 0.80 t1 = 4.96 t2 = 3.14 t3 = -0.95 t4 = -1.97 HT2 = 5.88
SignalMonitor.ExecuteSignals: Mdl04 T60.pred = 1.9372 DModX = 1.25 t1 = -0.87 t2 = -3.18 t3 = 1.49 HT2 = 2.66
211026 02:45:25 New data written, 2015-01-01 01:00:00 sumExLimits = 13

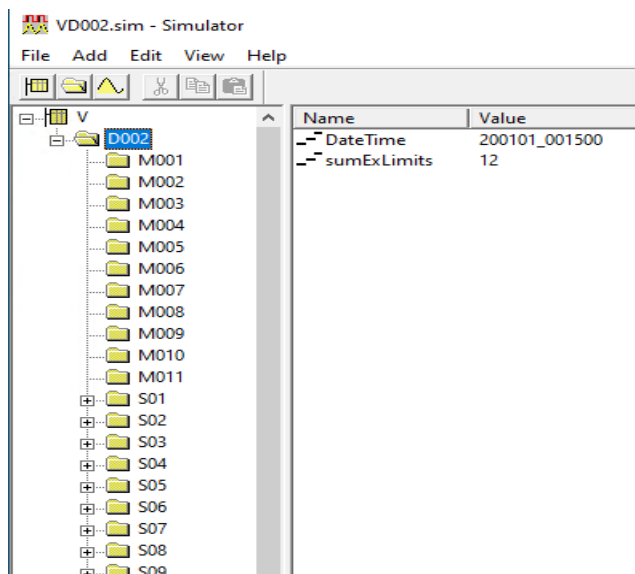
```

Figur 62. Skärmdump från DSWatchDog. Bilden visar hur programmet startat samt utskrifterna från en exekveringsrunda i ExecuteSignals(), först de univariata resultaten, därefter de multivariata för 4 aktiva modeller.

10.6 DATAKOMMUNIKATION

10.6.1 OPC-kommunikation

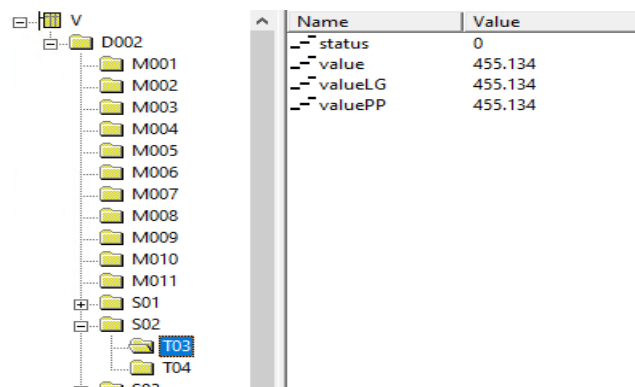
Vid OPC-kommunikation installeras en OPC-server för datautbytet. Vi har använt en publik, fri, OPC-server, ICONICS.Simulator, från Iconics, men man kan använda en godtycklig server. Det går också att kommunicera med olika servrar för olika data. Signalerna/taggarna är organiserade hierarkiskt på servern i noder: anläggningar (devices), grupper (groups), undergrupper (subgroups) och signaler (tags). Figur 63. Toppnoden, anläggning, förnärvande. V eller S, anger dammägarorganisation. Nästa nod i strukturen (grupp) anger damnumret, f.n. D002 och S001. Varje damgrupp har 2 signaler/taggar: DateTime och SumExLimits.



Figur 63. Den översta delen av den hierakiska OPC-strukturen för dammägare V och damm D002. Under D002 ligger modellnoder (M001 – M011) och mätenheter (S01 – Snn).

Den tredje nivåns noder, undergrupper, anger mätenheter och modeller.

Varje mätenhet har en undergrupp med signaler (T01 – Tnn). Varje signal har 4 taggar (status, rapporterat värde (value), senast godkända värde (valueLG) samt preprocessat värde (valuePP), Figur 64.



Figur 64. I bilden syns signalen T03, som till hör mätenheten S02. T03 har i sin tur 4 taggar: status, value, valueLG och valuePP.

Varje modellnod har i projektet 11 taggar: CuSumH, CuSumL, DModXpred, HT2pred, status, t1pred, t2pred, t3pred, t4pred, t5pred och ypred, Figur 65.

Name	Value
CuSumH	0
CuSumL	-0.384788
DModXpred	0.945957
HT2pred	4.63921
status	272
t1pred	1.04845
t2pred	3.79918
t3pred	-0.0880364
t4pred	-0.996252
t5pred	-1.#IND
ypred	4.05473

Figur 65. Modellen M002 tillhör damm D002 hos dammgärgare V. Modellen har 11 taggar som visar de senaste resultaten från modellens exekvering.

10.6.2 Filkommunikation

Vid filkommunikation överförs inkommande signaldata och resultatdata som csv-filer, observationsvis.

Inkommande filer:

Första raden har observationens tidsstämpel (DateTime), därefter kommer en rad per signal med signalnamn och mätvärde, Figur 66.

	Value
2020-01-01 00:30:00	
V.D002.S01.T01.value;	447.7008
V.D002.S01.T02.value;	35.90194
V.D002.S02.T03.value;	455.1333
V.D002.S02.T04.value;	3.608266
V.D002.S03.T05.value;	456.6124
V.D002.S03.T06.value;	4.190567
V.D002.S04.T07.value;	461.0256
V.D002.S04.T08.value;	2.420544
V.D002.S05.T09.value;	458.6048
V.D002.S05.T10.value;	10.26953
V.D002.S06.T11.value;	458.3039
V.D002.S06.T12.value;	5.897797
V.D002.S07.T13.value;	458.8557
V.D002.S07.T14.value;	4.348799

Figur 66. Exempel på dammgärgarens utfil.

Utgående filer:

Det är två varianter av utfiler. Båda är csv-filer (text). Den ena är avsedd att för dammgärgaren och den andra för IVL, för diagnostik.

Den som är avsedd för dammgärgaren är litet kortare än den andra. Första raden består DateTime och antalet signaler som har något överskridet gränsvärde. Därefter kommer rader med signalnamn och tillhörande signalstatus. Slutligen kommer modellresultaten, 5+A rader per modell med: modellstatus (Status), *DModXpred, upp till 5 tnpred (beroende på antalet komponenter för modellen, A), HT2pred, Ypred (signalnamn.pred) , CuSumH, CuSumL, Figur 67.

```

VD002_200101_000000.txt
1 2020-01-01 00:00:00;sumExLimits;12
2 V.D002.S01.T01.status;0
3 V.D002.S01.T02.status;8
4 V.D002.S02.T03.status;0
5 V.D002.S02.T04.status;0
6 V.D002.S03.T05.status;0
7 V.D002.S03.T06.status;0
8 V.D002.S04.T07.status;0
9 V.D002.S04.T08.status;0
10 V.D002.S05.T09.status;0
11 V.D002.S05.T10.status;0
12 V.D002.S06.T11.status;0
13 V.D002.S06.T12.status;8
14 V.D002.S07.T13.status;0
15 V.D002.S07.T14.status;0
...
162 V.D002.M009.T70.CuSumL;0.000
163 V.D002.M010.Status;112
164 V.D002.M010.DModXpred;0.455
165 V.D002.M010.t1pred;1.861
166 V.D002.M010.t2pred;2.470
167 V.D002.M010.HT2pred;3.210
168 V.D002.M010.T04.pred;4.862
169 V.D002.M010.T04.CuSumH;0.000
170 V.D002.M010.T04.CuSumL;-1.072
171 V.D002.M011.Status;49
172 V.D002.M011.DModXpred;1.632
173 V.D002.M011.t1pred;0.053
174 V.D002.M011.HT2pred;0.001
175 V.D002.M011.T70.pred;14.506
176 V.D002.M011.T70.CuSumH;0.000
177 V.D002.M011.T70.CuSumL;0.000

```

Figur 67. Exempel på dammägarens utfil.

Den andra filvarianten är mer detaljerad avseende signalerna och är avsedd för IVL för diagnostiska ändamål. Första raden består av DateTime och antalet signaler som har något överskridet gränsvärde. Därefter kommer fyra rader per signal med signalnamn plus endera signalvärde (value), signalstatus (status), preprocessat värde (valuePP) och senast godkända värde (valueLG). Slutligen kommer modellresultaten på samma sätt som ovan, Figur 68.

```

IVL_VD002_200101_000000.txt
1 2020-01-01 00:00:00;sumExLimits;12
2 V.D002.S01.T01.value;447.7029
3 V.D002.S01.T01.status;0
4 V.D002.S01.T01.valuePP;447.7029
5 V.D002.S01.T01.valueLG;447.7029
6 V.D002.S01.T02.value;35.90159
7 V.D002.S01.T02.status;8
8 V.D002.S01.T02.valuePP;NaN
9 V.D002.S01.T02.valueLG;NaN
10 V.D002.S02.T03.value;455.1353
11 V.D002.S02.T03.status;0
12 V.D002.S02.T03.valuePP;455.1353
13 V.D002.S02.T03.valueLG;455.1353
14 V.D002.S02.T04.value;3.608219
...
425 V.D002.M009.T70.pred;11.423
426 V.D002.M010.Status;112
427 V.D002.M010.DModXpred;0.455
428 V.D002.M010.t1pred;1.861
429 V.D002.M010.t2pred;2.470
430 V.D002.M010.HT2pred;3.210
431 V.D002.M010.T04.pred;4.862
432 V.D002.M011.Status;49
433 V.D002.M011.DModXpred;1.632
434 V.D002.M011.t1pred;0.053
435 V.D002.M011.HT2pred;0.001
436 V.D002.M011.T70.pred;14.506

```

Figur 68. Exempel på dammägarens utfil.

10.7 DATABAS

Data för signaler och modeller hanteras i en SQL-databas (MS SQL Express). Detta för att implementationen av mjukvaran ska vara modulär och lätt överförbar till en ny anläggning. Hårdkodade värden i programkoden är ofta svåra att överblicka vilket kan innebära att ändringar som borde gjorts inte alltid görs. Detta är dessutom generellt dålig praxis och tenderar ge en högre buggbenägenhet. Databasen har 4 tabeller: ModellLimits, Models, NodeServers och signals, Figur 69.

ModellLimits	Models	NodeServers	Signals
PName	[Index]	Id	Varlx
MName	PName	ServerName	ProjectName
IName	MName	NodeName	VariableName
BinNo	MType		OPCTagName
LoLimit	IsActive		NodeServerId
UpLimit	Comp		MinLimit
	R2X		MaxLimit
	R2Y		MaxConstant
	Q2		MedianMinus
	CreationDate		DBTagName
	SDYres		Monitored
			MaxAbs1stDer
			MaxAbs2ndDer
			MaxMovVar
			RangeMin
			RangeMax
			MinWarn
			MaxWarn
			IVLminLim

Figur 69. Projektets databas består av 4 tabeller.

10.7.1 ModellLimits

Tabellen ModellLimits används för modellparametrarnas gränsvärden, Figur 70.

PName	MName	IName	BinNo	LoLimit	UpLimit
VD002	Mdl01	t1	10	4.67894173	8.032213
VD002	Mdl01	t1	11	5.036745	8.333357
VD002	Mdl01	t1	12	5.364814	8.46351051
VD002	Mdl01	t1	13	5.38093567	8.65553
VD002	Mdl01	t1	14	5.4071207	8.43708
VD002	Mdl01	t1	15	4.990913	8.07052
VD002	Mdl01	t1	16	4.615162	8.038365
VD002	Mdl01	t1	17	4.73182869	7.51887226
VD002	Mdl01	t1	18	4.65429354	7.52894354
VD002	Mdl01	t1	19	3.93957686	7.38122
VD002	Mdl01	t1	20	2.602622	7.240633
VD002	Mdl01	t1	21	0.9858823	7.033077
VD002	Mdl01	t1	22	-0.676514149	6.2003417
VD002	Mdl01	t1	23	-2.452134	4.253874
VD002	Mdl01	t1	24	-4.96580231	2.142414

Figur 70. Exempel på tabellen ModellLimits.

10.7.2 Models

Tabellen Models innehåller modelleringsresultaten från DSModelMaker för projektets alla modeller, liksom en indikator, IsActive, som kan användas för att aktivera/stänga av exekveringen av respektive modell, Figur 71.

	PName	MName	MType	Comp	R2X	R2Y	Q2	CreationDate	SDYres	IsActive
1	VD002	Mdl01	PCA	3	0.7183943	NULL	NULL	2021-12-15 13:44:33.000	NULL	1
2	VD002	Mdl02	PLS	4	0.914058	0.9855031	0.9769129	2021-12-15 13:44:46.000	0.1968146	1
3	VD002	Mdl03	PLS	5	0.9459207	0.8745148	0.6492641	2021-12-15 13:44:46.000	0.07846433	1
4	VD002	Mdl04	PLS	4	0.8712435	0.9747629	0.9411257	2021-12-15 13:44:47.000	0.3420514	1
5	VD002	Mdl05	PLS	3	0.7670978	0.966	0.9447652	2021-12-15 13:44:47.000	0.6833795	1
6	VD002	Mdl06	PLS	3	0.8118477	0.9316496	0.9000871	2021-12-15 13:44:48.000	0.2995642	1
7	VD002	Mdl07	PLS	1	0.2930332	0.7299023	0.6657086	2021-12-15 13:44:48.000	0.1286348	1
8	VD002	Mdl08	PLS	4	0.8411904	0.9375236	0.8711108	2021-12-15 13:44:48.000	0.03358244	1
9	VD002	Mdl09	PLS	2	0.710816	0.9832014	0.9783506	2021-12-15 13:44:48.000	2.094327	1
10	VD002	Mdl10	PLS	2	0.8461459	0.8944732	0.8759262	2021-12-15 13:44:49.000	0.388687	1
11	VD002	Mdl11	PLS	1	0.4844239	0.8513403	0.8450428	2021-12-15 13:44:49.000	9.903655	1

Figur 71. Exempel på modelltabell.

10.7.3 NodeServers

NodeSvertabellen används för att knyta signalernas OPC-namn till rätt dator (NodeName = IP-adress) och OPC-Server (ServerName). Detta sker genom att kolumnen NodeServerId i tabellen Signals länkar till Id i den här tabellen, Figur 72.

Id	ServerName	NodeName
1	ICONICS.Simulator.1	192.168.11.67

Figur 72. Exempel på tabellen NodeServers. Här ligger alla signaler på samma OPC.server, därför har den bara en rad.

10.7.4 Signals

Tabellen Signals innehåller signaldefinitionerna: namn och de univariata parametergränserna, Figur 73. Dessa data importeras från filen ObsVarDef_ooDnnn.xls med programmet DSPopulateSignals. Kolumnen Signals.NodeServerId länkar till NodeServers.Id, på så sätt kan alla signaler ligga på godtyckliga datorer och OPC-servrar.

Column Name	Data Type	Allow Nulls
Varix	int	☐
ProjectName	nvarchar(255)	☒
VariableName	nvarchar(255)	☒
OPCTagName	nvarchar(255)	☒
NodeServerId	int	☒
MinLimit	real	☒
MaxLimit	real	☒
MaxConstant	smallint	☒
MedianMinus	real	☒
DBTagName	nvarchar(255)	☒
Monitored	bit	☒
MaxAbs1stDer	real	☒
MaxAbs2ndDer	real	☒
MaxMovVar	real	☒
RangeMin	real	☒
RangeMax	real	☒
MinWarn	real	☒
MaxWarn	real	☒
IVLminLim	real	☒

Variabelindex
Projektamn - tex VD002
Variabelnamn – anonymiserat
Signalens OPCnamn
Den dator där signalens OPC server finns
Undre signalgränsvärde
Övre signalgränsvärde
Max antal samma värden i följd
Antal värden bakåt i tiden för medianberäkning
Signalens namn i historikdatabasen
Monitored är sann om signalen skall övervakas
Gränsvärde för signalens 1a-derivata
Gränsvärde för signalens 2a-derivata
Gränsvärde för glidande varians
Nedre diagramgräns vid plottnig

Figur 73. Kolumndefinitioner med kommentarer för tabellen Signals

En delmängd av Signals visas i Figur 74.

VarId	ProjectName	VariableName	OPCTagName	NodeServerId	MinLimit	MaxLimit	MaxConstant	MedianMinus	DBTagName	Monitored	MaxAbs1stDer	MaxAbs2ndDer	MaxMovVar	RangeMin	RangeMax	MinWarn	MaxWarn	IVLimit
1	VD002	V.D002.S01.T01	V.D002.S01.T01	1	448.8	452	96	NULL	V.D002.S01.T01	1	1.3	2	0.05	448	452	NULL	NULL	NULL
2	VD002	V.D002.S01.T02	V.D002.S01.T02	1	2	6	96	12	V.D002.S01.T02	1	2	3	0.05	1	7	NULL	NULL	NULL
3	VD002	V.D002.S02.T03	V.D002.S02.T03	1	453.3	457	96	NULL	V.D002.S02.T03	1	2	3	0.05	452	457	NULL	NULL	NULL
4	VD002	V.D002.S02.T04	V.D002.S02.T04	1	0	10	96	6	V.D002.S02.T04	1	3.5	5	1	1	7	NULL	NULL	NULL
5	VD002	V.D002.S03.T05	V.D002.S03.T05	1	456.4	459.3	96	NULL	V.D002.S03.T05	1	2	3	10	455.5	460	NULL	NULL	NULL
6	VD002	V.D002.S03.T06	V.D002.S03.T06	1	0	10	96	1	V.D002.S03.T06	1	2	3	2	1	18.5	NULL	NULL	NULL
7	VD002	V.D002.S04.T07	V.D002.S04.T07	1	460	464	96	6	V.D002.S04.T07	1	2	3	0.05	459	465	NULL	NULL	NULL
8	VD002	V.D002.S04.T08	V.D002.S04.T08	1	-20	50	96	NULL	V.D002.S04.T08	1	0.7	1	2	2	5	NULL	NULL	NULL
9	VD002	V.D002.S05.T09	V.D002.S05.T09	1	456	462	96	6	V.D002.S05.T09	1	2	3	0.5	456.5	462	NULL	NULL	NULL
10	VD002	V.D002.S05.T10	V.D002.S05.T10	1	7.2	16	96	NULL	V.D002.S05.T10	1	2	3	0.5	4	13	NULL	NULL	NULL
11	VD002	V.D002.S06.T11	V.D002.S06.T11	1	456	461	96	6	V.D002.S06.T11	1	0.7	1	0.5	456.5	461	NULL	NULL	NULL
12	VD002	V.D002.S06.T12	V.D002.S06.T12	1	3.5	5	96	6	V.D002.S06.T12	1	1.3	2	0.5	-2	10	NULL	NULL	NULL
13	VD002	V.D002.S07.T13	V.D002.S07.T13	1	457.2	461	96	NULL	V.D002.S07.T13	1	1.3	2	0.5	457	461	NULL	NULL	NULL
14	VD002	V.D002.S07.T14	V.D002.S07.T14	1	2	7	96	6	V.D002.S07.T14	1	2	3	0.5	1.5	6.5	NULL	NULL	NULL
15	VD002	V.D002.S08.T15	V.D002.S08.T15	1	420.7	422.5	96	NULL	V.D002.S08.T15	1	0.7	1	0.5	420	422.5	NULL	NULL	NULL
16	VD002	V.D002.S08.T16	V.D002.S08.T16	1	0	6	96	18	V.D002.S08.T16	1	2	3	0.5	-0.5	5.5	NULL	NULL	NULL
17	VD002	V.D002.S09.T17	V.D002.S09.T17	1	388	391	96	NULL	V.D002.S09.T17	1	1.3	2	0.5	388	391	NULL	NULL	NULL
18	UD001	U.D001.S06.T18	U.D001.S06.T18	1	1.6	6	64	15	U.D001.S06.T18	1	5	9	0.06	5	14.6	NULL	NULL	NULL

Figur 74. Delmängd av tabellen Signals.

11 Workshops

Under projektets gång har ett antal arbetsmöten genomförts. De flesta har varit digitala då rådande omständigheter inte möjliggjort fysiska möten i stor utsträckning. Ett fysiskt möte har dock kunnat genomföras.

- Tidigt under projektet genomfördes en digital workshop tillsammans med Vattenfall för att se över och diskutera visualisering av multivariata modeller.
- Under hösten 2021 genomfördes en serie av två digitala arbetsmöten rörande olika typer av tillståndsavvikelser hos en damm samt simulering av dessa. Första tillfället genomfördes med Fortum och den andra med hela referensgruppen.
- Den 3 november 2021 genomfördes ett fysiskt möte tillsammans med referensgruppen. Arbetsgruppen träffades vid en dammanläggning i Västerbotten. Under förmiddagen genomfördes ett längre referensgruppsmöte där diskussioner kring, projektets framsteg, en ny ansökan, resultat kring genombrottssimulering samt vad som händer inom branschen rörande mätteknik. Under eftermiddagen genomfördes ett studiebesök till dammanläggningen Bjurfors Nedre.
- Ett kortare digitalt möte genomfördes den 20 november 2021. Temat för detta tillfälle var visualisering av projektets resultat i befintliga system och diskussionerna fokuserade på möjligheter att förvalta och visualisera de implementerade lösningarna.

12 Slutsatser

- Univariata metoder och multivariata exempelmodeller har implementerats i projektets två fallstudier. Övervakningen är i drift och dess resultat kommer behöva itereras för att förbättra övervakningens prestanda.
- Multivariata modeller har utvecklats för två fallstudier. De implementerade modellerna är PCA och PLS modeller. PLS-modellerna visar generellt sätt en hög prediktionsförmåga och dess resultat anses vara tillförlitliga.
- Andra typer av regressionsmodeller, såsom MLR och HST, har utvecklats och dess resultat har jämförts med motsvarande PLS modeller. Resultaten visar att PLS, i dessa fall, resulterar i den bäst presterande modellen vid prediktion av en givarsignal i dammkroppar. MLR resulterar i överanpassade modeller vilket minskar dess tillförlitlighet avsevärt i system där antalet givarsignaler är stort.
- För de PLS modeller som skapats inom projektet har endast en givarsignal utgjort Y. Vid PLS-modellering finns möjlighet att inkludera fler givare i Y vilket troligtvis skulle vara användbart för grupper av givarsignaler i fallstudierna. En signal består av en struktur och brus. Brus yttrar sig som en stokastisk variation mellan mätvärden och är förekommande både i X och Y. Om två givarsignaler med samma struktur (hög korrelation) men olika brus (vilket oftast är fallet) skulle utgöra Y kan strukturen förstärkas och bruset reduceras, vilket skulle kunna förbättra modellerna.
- Konfidensintervall för cykliska processer är utforskat område och ingen litteratur på ämnet har identifierats under projektet. Utveckling och forskning har skett under projektets gång och ett första utkast på konfidensintervall har implementerats. Mer forskning krävs på området och metoderna som använts behöver vidareutvecklas.
- För att resultaten från modeller och övervakning skall vara användbara behöver dammägare utveckla sin kompetens rörande signalövervakning och multivariat modellering. Modeller och övervakning behöver underhållas och eventuellt justeras efter hand för att systemet skall ha en fortsatt hög användbarhet.
- Resultaten från övervakningsmetoderna för avvikelser för enskilda givare och övergripande dammtillstånd som baseras på multivariata metoder pekar på att utformningen eller värdet på gränsvärdena behöver förändras.
- En mycket viktig slutsats, som dragits flera gånger innan detta projekt, är kraven på datakvalitet. För att smart, effektiv och automatiserad övervakning skall utvecklas behöver de data som både används vid utvecklande, liksom data som övervakas vara av god kvalitet. För att data skall komma till nytta i framtiden behöver den goda kvaliteten säkerställas redan idag. Data består i detta fall inte bara av signaler från givare utan även av dokumentation om vad som fysiskt gjorts med givarna. Detta handlar om kalibrering, underhåll och givarbyten. När dessa genomförs är en typ av information som lätt glöms bort med tiden. Tydliga och standardiserade loggböcker över dessa ingrepp och justeringar kommer i framtiden vara mycket hjälpfyllt då detta kan förklara beteenden i signalerna.

13 Termer och förkortningar

Beteckning	Förklaring
ANN	Artificella Neurala Nät/ Artificial Neural Network
EWMA	Exponentiellt viktad medelvärdesbildning/ Exponentially Weighted Moving Average En populär form av glidande medelvärdesbildning som kan implementeras med bara ett värde i minnet. Det fönster som appliceras på signalen har en exponentiell form.
HMI	Användargränssnitt/ Human-Machine Interface
HST	Hydrostatisk säsong tidsmodell/ Hydrostatic-season-time model
Median	Det mittersta värdet i en datamängd; det värde som efter sortering efter värde finns i mitten av dataserien
ML	Maskininlärning/Machine Learning
MLR	Multiple Linear Regression/ Multipel Linjär Regression
MSPC	Multivariat statistisk processkontroll/ Multivariate Statistical Process Control
OPC	Open Process Communication. En standard för industriell signalöverföring
Outlier	Avvikande värde
PCA	Principalkomponentanalys/ Principal Components Analysis
PLS	Partiell minsta kvadratmetod eller projektioner till latent strukturer/Partial Least Squares or Projection to Latent Structures
SCADA	Sensory Control and Data Acquisition
SPC	Statistisk processkontroll/Statistical Process Control
Spik	Momentant avvikande värde som snabbt återgår till normalläge
SQL	Structured Query Language

14 Referenser

- Ahn, S. J., Rauh, W., & Warnecke, H.-J. (2000). Least-squares orthogonal distances fitting of circle, sphere, ellipse, hyperbola, and parabola. *Pattern Recognition*, 34, 2283-2303.
- Bonelli, S., Tourment, R., & Felix, H. (2003, 9-11 Oct 2003). *Analysis of earthdam monitoring data*. Paper presented at the Selected Problems of Water Engineering, Krakow, Poland.
- Bulyaculov, R. R., Schogoleva, K. P., Yakovlev, I. N., & Roskostov, R. A. (2017). *Modelling and analysis of the median filter algorithm of suppression of impulse noise*. Rapport presenterad på 2017 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), Saint Petersburg, Russia.
- C-sharp. (2021). Hämtad från <https://sv.wikipedia.org/wiki/C-sharp>
- Chong, I.-G., & Jun, C.-H. (2005). Performance of some variable selection methods when multicollinearity is present. *Chemometrics and intelligent laboratory systems*, 78(1-2), 103-112.
- Dimitri, v. H. (2022). Doxygen: Doxygen. Hämtad från www.doxygen.nl
- Duda, R. O., Hart, P. E., & Stork, D. G. (2006). *Pattern Classification* (2nd ed.): Wiley Interscience.
- Ekonomifakta. (2020). Sveriges totala elproduktion fördelat på energislag, 2019, TWh. Hämtad från https://www.ekonomifakta.se/Fakta/Energi/Energibalans-i-Sverige/Elproduktion/?gclid=Cj0KCQiA7qP9BRCLARIsABDaZziz1w25ccMa5Uqt-Zf_Ij0AJpXpjMpNoJrzTqa0focf9QzZ0zGBX9IaAm4gEALw_wcB&graph=/12218/all/all/
- Eriksson, L. B., T. Johansson, E. T., J., & Vikström, C. (2003). *Multi- and Megavariate Data Analysis. Principles and Applications*: Umetrics.
- Gamse, S., & Oberguggenberger, M. (2017). Assessment of long-term coordinate time series using hydrostatic-season-time model for rock-fill embankment dam. *Structural Control and Health Monitoring*, 24(1), e1859. doi:10.1002/stc.1859
- GENESIS64. (2021). *HMI/SCADA | ICONICS Software Solutions*. Hämtad från <https://iconics.com/Products/GENESIS64>
- Haykin, S. S. (1994). *Neural Networks: a comprehensive foundation* (1st ed.). Indianapolis, USA.: Macmillan Publishing Co.
- Hyper Historian. (2021). *Product | ICONICS Software Solutions*. Hämtad från <https://iconics.com/Products/Hyper-Historian>
- Iconics. (2021). Hämtad från <https://en.wikipedia.org/wiki/Iconics>
- Jacobson, A., Torbjörn, J., Håkan, F., & Björk, A. (2021). *Detektion av givar- och kommunikationsfel vid dammövervakning* (Rapport 2021:735). Hämtad från <https://energiforsk.se/media/29553/detektion-av-givar-och-kommunikationsfel-vid-dammovervakning-energiforskrapport-2021-735.pdf>
- Johansson, T., Fridén, H., Sundberg, M., Samuelsson, O., & Björk, A. (2020). *Datadrivna metoder för att detektera avvikande mätvärden inom dammsäkerhet* (Rapport 2020:640). Hämtad från <https://energiforsk.se/media/27536/datadrivna-metoder-for-att-detektera-avvikande-matvarden-inom-dammsakerhet-energiforskrapport-2020-640.pdf>
- MATLAB. (2021). Hämtad från <https://sv.wikipedia.org/wiki/Matlab>
- Microsoft Visual Studio. (2021). Hämtad från https://en.wikipedia.org/wiki/Microsoft_Visual_Studio
- Miljöbalken (1998:808)*. (1998). Stockholm: Regeringskansliet.
- Miljödepartementet. (2015, 2015-04-02). Dammsäkerhet, Tydliga regler och effektiv tillsyn. Hämtad från <https://www.regeringen.se/rattsliga-dokument/statens-offentliga-utredningar/2012/06/sou-201246/>

- Miljöväder ska visualisera utsläpp och buller. (2020). Hämtad från <https://www.ivl.se/vart-erbjudande/forskning/luft/miljovader-ska-visualisera-utslapp-och-buller.html>
- Montgomery, D. C. (2013). *Statistical quality control : a modern introduction*. Hoboken, N.J.: Wiley.
- .NET Framework. (2020). Hämtad från https://sv.wikipedia.org/wiki/.NET_Framework
- Nilsson, Å. (2014). *DAMMSÄKERHET, Dam safety performance monitoring and data management – best practice*. Hämtad från Stockholm: <https://energiforskmedia.blob.core.windows.net/media/18414/dam-safety-performance-monitoring-and-data-management-kortversion-energiforskrapport-2014-9.pdf>
- OSIsoft. (2020). Hämtad från <https://en.wikipedia.org/wiki/OSIsoft>
- Piroddi, L., & Spinelli, W. (2003). *Long-range nonlinear prediction: a case study*. Paper presented at the 42nd IEEE Conference on Decision and Control.
- Principal Component Analysis. (2019). Hämtad från https://en.wikipedia.org/wiki/Principal_component_analysis
- Products. (2022). Hämtad från <http://www.frideninfotech.eu/Products.html>
- RIDAS - Energiföretagens riktlinjer för dammsäkerhet. (2019). (J. Bladh Ed.). Stockholm: Energiföretagen Sverige.
- Savitzky, A., & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.*, 36(8), 1627-1639.
- SIMCA - Master your data! (2021). Hämtad från <https://www.sartorius.com/en/products/process-analytical-technology/data-analytics-software/mvda-software/simca>
- SQL Server Express. (2020). Hämtad från https://en.wikipedia.org/wiki/SQL_Server_Express
- SvenskaKraftnät. (2020). *Dammsäkerhet, Tillämpliga regelverk, vägledning och stöd*. Hämtad från Sundbyberg:
- Varmuza, K., & Filzmoser, P. (2008). *Introduction to Multivariate Statistical Analysis in Chemometrics* (pp. 1-328). Hämtad från <http://kcl.eblib.com/patron/FullRecord.aspx?p=427098>
- What is OPC? (2021). Hämtad från <https://opcfoundation.org/about/what-is-opc/>

METODER FÖR DATADRIVEN DAMM-ÖVERVAKNING

Dammövervakning har alltid varit ett stort och viktigt område för svenska dammägare. Det är en utmaning att kontinuerligt övervaka en dammkonstruktion och snabbt kunna reagera på förändringar som visar på en avvikelse i dammen.

Samtidigt kan avvikelser i själva övervakningsutrustning förekomma, som skulle kunna uppfattas som en avvikelse i dammen. Vanligt förekommande övervakningsfel kan bero på fel i givare eller i mätvärdets överföring från givare till datacentral. Den typen av fel behöver detekteras, identifieras och skiljas från avvikelser som kan tyda på en förändring i dammen.

Den här rapporten presenterar ett antal exempel på olika typer av multivariata modeller baserade på etablerade metoder inom maskininlärning. Rapporten presenterar även tillvägagångssätt för att utveckla och utvärdera multivariata modeller. Ett antal utvecklade övervakningsmetoder beskrivs och resultat baserade på de framtagna modellerna förklaras.

Tillsammans med dammägare har multivariata modeller och övervakningsmetoder driftsatts i de två fallstudierna.

Resultaten kan ligga till grund en för fortsatt utveckling av kostnadseffektiv multivariat övervakning av dammar samt en kurs för dammägare som syftar till att öka nyttan med övervakning av givarsignaler.

Ett nytt steg i energiforskningen

Energiforsk är en forsknings- och kunskapsorganisation som samlar stora delar av svensk forskning och utveckling om energi. Målet är att öka effektivitet och nyttiggörande av resultat inför framtida utmaningar inom energiområdet. Vi verkar inom ett antal forskningsområden, och tar fram kunskap om resurseffektiv energi i ett helhetsperspektiv – från källan, via omvandling och överföring till användning av energin. www.energiforsk.se



Energiforsk