UPTEC STS07 007
Examensarbete 20 p
Februari 2007

Data Mining av processdata

Principalkomponentanalys av processdata från ett av Vattenfall AB:s vattenkraftverk

Björn Thorsélius

UPPSALA
UNIVERSITET

Teknisk- naturvetenskaplig fakultet
UTH-enheten

Besöksadress:
Ångströmlaboratoriet
Lägerhyddsvägen 1
Hus 4, Plan 0

Postadress:
Box 536
751 21 Uppsala

Telefon:
018 – 471 30 03

Telefax:
018 – 471 30 00

Hemsida:
http://www.teknat.uu.se/student

Abstract

Data Mining of Process Data

Björn Thorsélius

In most process industries, monitoring of the process is performed to ensure quality, safety and economic requirements imposed on the production. The monitoring process often involves data collection from a large amount of entities and an increased amount of data calls for techniques to handle large data sets. Data Mining methods are developed for this specific reason and can be applied to data from complex industrial production processes. Pattern recognition thereby becomes possible and hence the operator can gain knowledge about the process.

This thesis discusses how process data from a specific hydro power station, operated by Vattenfall AB, can be analysed using Data Mining methods. Particularly Principal Component Analysis (PCA) is studied and the thesis explains in detail how a PCA is applied to the specific process data. Considerations that must be taken prior to a PCA are also discussed and the thesis aims to explain how PCA can contribute to increase the process knowledge of the operators. A few other methods for analysing the process data are mentioned and a discussion about how the PCA can be refined is included.

The thesis clearly shows the potential of PCA when analysing process data from a hydro power station. A possible application of PCA in a real-time situation is demonstrated by a demonstration application. With use of PCA the demonstration application shows how operational modes not classified as normal can be detected before any of the conventional alarm limits have been breached.

Handledare: Erik Hjelmvik, Vattenfall Research and Development
Ämnesgranskare: Kjell Orsborn, Uppsala universitet
Examinator: Elísabet Andrésdóttir, Uppsala universitet
ISSN: 1650-8319, UPTEC STS07 007
Sponsor: Vattenfall AB Vattenkraft

Sammanfattning

Stora mängder process- och produktionsdata samlas in från Vattenfall AB Vattenkrafts anläggningar. Data i form av exempelvis temperaturer, flöden, vattennivåer och producerad effekt hämtas i realtid in till en databas. Databasen utgör potentiellt en stor källa av värdefull kunskap om anläggningsdriften. Idag görs dock inga analyser av processdata. Anledningen är att det för närvarande saknas lämpligt analysverktyg.

Arbetet utgör en utredning av hur processdata från vattenkraftsanläggningar kan analyseras på automatisk väg genom att använda avancerade verktyg inom området för Data Mining. Avsikten med arbetet är därmed att det ska fungera som underlag till ett, för Vattenfall AB Vattenkraft, nytt analysverktyg, vilket ska bidra till att göra kontinuerliga analyser av processdata från vattenkraftsanläggningarna möjliga.

Särskilt en metod, principalkomponentanalys (PCA), har studerats. PCA är en metod som, utan att försumma viktig information i data, reducerar antalet dimensioner hos data-mängden, så att variationer i data synliggörs. För att påvisa hur PCA kan appliceras på Vattenfall AB Vattenkrafts anläggningar, samt för att poängtera potentialen hos omfattande processdata-analyser, har en demonstrationsapplikation konstruerats. Applikationen visar hur PCA kan användas för att underlätta driftövervakning av vattenkraftverk. Med PCA kan samtliga variabler inkluderas i en analys och störningar i processen, som eventuellt baseras på små avvikelser hos ett flertal variabler samtidigt, kan därmed upptäckas. På så vis skulle de statiska larmgränserna kunna kompletteras av användandet av en mer dynamisk PCA-modell.

Examensarbetet visar att PCA är en metod med potential att analysera processdata från vattenkraftsanläggningar. Dessutom belyses metoder vilka kan förbättra resultaten från en PCA. Förbehandling av data kan exempelvis vara särskilt viktigt för att rättvisande resultat ska kunna erhållas från analysen.

Examensarbetet, som omfattar 20 poäng, har, på uppdrag av Vattenfall AB Vattenkraft, utförts i samarbete med Vattenfall Research and Development AB i Älvkarleby, som en del av civilingenjörsutbildningen System i Teknik och Samhälle vid Uppsala universitet.

Förkortningar

AI	artificiell intelligens
ANN	artificiella neurala nätverk
BSS	Blind Source Separation
DBSCAN	Density Based Spatial Clustering of Applications with Noise
DC	driftcentral
DM	Data Mining
GIGO	Garbage In, Garbage Out
IC	Independent Component
ICA	Independent Component Analysis
KDD	Knowledge Discovery in Databases
KNN	K Nearest Neighbour
MK	minsta kvadrat
OLE	Object Linking and Embedding
OPC	OLE for Process Control
PC	principalkomponent
PCA	principalkomponentanalys
PLC	Programmable Logic Controller
PLS	Partial Least Squares Projection to Latent Structures
PT	processtillstånd
PRESS	Predicted Residual Error Sum of Squares
RMSECV	Root-Mean-Square Error of Cross-Validation
TK	tillståndskontroll
VSN	Vattenfall Service Nord

Innehållsförteckning

Bilageförteckning

1 Inledning

”We are drowning in information, but starving for knowledge.”
John Naisbitt (1929-), författare och futurist

I takt med den fortlöpande utvecklingen av ny teknik har ökade möjligheter för realtidsövervakning av industriell produktion vuxit fram. Att sensorer och givare har blivit mindre, mer exakta, snabbare och billigare att införskaffa och installera har inneburit att företag i allt större utsträckning nu använder sig av tekniska lösningar för att övervaka sina produktionsprocesser. Med relativt enkla medel, såsom genom att mäta exempelvis olika temperaturer, hastigheter och tryck, kan driftansvarig från ett kontrollrum på så sätt kontrollera åtskilliga parametrar av vikt för processens fortgång.

Samtidigt har ökade möjligheter för lagring av data, i form av modern databasteknik och växande lagringskapacitet, lett till att stora mängder data kan sparas på disk varefter den på ett effektivt sätt kan hanteras. Med elektroniskt lagrade data kan analyser av trender eller avvikelser lättare utföras. Analyserna kan underlätta förståelsen för och ge en ökad kunskap om olika delar av produktionsprocessen, vilket kan vara av vikt vid exempelvis prognostisering eller för planerings- och underhållsarbete. I många fall handlar det dock om så stora datamängder att manuella analysmetoder inte är applicerbara. Häri ligger ett problem. Det är för att kunna hantera, analysera och extrahera relevant information ur den alltjämt växande datamängden, som metoder för Data Mining har vuxit fram.

Examensarbetet utfördes, på uppdrag av Vattenfall AB Vattenkraft, i samarbete med Vattenfall Research and Development AB i Älvkarleby, som en del av civilingenjörsutbildningen System i Teknik och Samhälle vid Uppsala universitet och utreder hur metoder inom området för Data Mining kan appliceras på en vattenkraftsanläggning.

1.1 Problembakgrund

I ett samarbete med Vattenfall Research and Development AB samlar Vattenfall AB Vattenkraft in och lagrar stora mängder process- och produktionsdata från många av sina anläggningar. Data i form av exempelvis temperaturer, flöden, vattennivåer och producerad effekt hämtas i realtid in till en databas. Databasen utgör potentiellt en stor källa av värdefull kunskap om anläggningsdriften. Idag görs dock inga analyser av processdata. Anledningen är att det för närvarande saknas lämpligt analysverktyg. Nästa steg i samarbetet är att visa huruvida det är möjligt att, på automatisk väg, finna relevanta samband mellan olika mätvärden i databasen och hur dessa samband sedan kan användas för att fortgående analysera anläggningarnas drift. Samtliga driftstopp av vattenkraftsanläggningar innebär nämligen enorma kostnader för Vattenfall AB. Genom att lära ny kunskap om hur driften av anläggningarna fungerar hyser Vattenfall AB hopp om att kunna öka driftsäkerheten i sina anläggningar och på så sätt minska risken för oönskade driftstopp. Att förhindra ett enda driftstopp skulle bespara företaget stora summor pengar. Med hjälp av lämpliga analysmetoder skulle utgifter genererade av oönskade driftstopp kunna undvikas.

1.2 Syfte

Arbetet utgör en utredning av hur processdata från vattenkraftsanläggningar kan analyseras på automatisk väg med hjälp av verktyg inom området för Data Mining. Avsikten med rapporten är därmed att fungera som underlag till ett, för Vattenfall AB Vattenkraft, nytt analysverktyg, vilket ska bidra till att göra kontinuerliga analyser av processdata från vattenkraftsanläggningarna möjliga.

> Syftet med arbetet var att identifiera metoder, inom området för Data Mining, som är lämpliga att använda för att, på automatisk väg, analysera processdata från en vattenkraftsanläggning. Syftet var vidare att implementera en metod för att påvisa hur den kan appliceras på Vattenfall AB Vattenkrafts anläggningar samt för att poängtera potentialen hos omfattande processdata-analyser.

1.3 Mål

Arbetet är ämnat som grund till eventuell framtida implementering av automatiska analysverktyg i Vattenfall AB Vattenkrafts system för att öka driftsäkerheten i Vattenfall AB:s vattenkraftsanläggningar. Målsättningen med arbetet var att skapa förståelse för hur kontinuerliga analyser av processdata från vattenkraftsanläggningarna kan utföras. Målet var vidare att insikter från arbetets resultat ska göra framtida analyser av pågående driftsituation möjliga. Genom att skapa ökad förståelse för anläggningsdriften, i form av lättolkade analysverktyg, var målsättningen att avvikande processbeteende i framtiden ska kunna detekteras tidigare än vad de befintliga analysverktygen medger. Förhoppningen var därför att arbetets resultat kan komma att ligga till grund för att Vattenfall AB Vattenkraft ska kunna skapa nya möjligheter för prognostisering och utvärdering av anläggningsdrift samt att driftsäkerheten hos deras anläggningar ska kunna höjas, vilket i förlängningen kan stärka Vattenfall AB:s position som ett av Europas ledande energiföretag.

1.4 Avgränsning

Arbetet inbegriper utveckling av en, för Vattenfall AB Vattenkraft, skräddarsydd metod för analysering av processdata. Däremot var avsikten inte att generera en färdig produkt utan arbetet skulle snarast leda fram till en prototyp som visar på nyttan hos metoden och som kan utgöra grund för vidare utveckling och implementering av analysverktyg som kan användas i Vattenfall AB Vattenkrafts verksamhet.

1.5 Erkännanden

Först och främst riktas ett varmt tack till Erik Hjelmvik, som har agerat handledare under examensarbetet. Erik har kontinuerligt bidragit med konstruktiv kritik och har varit ett värdefullt stöd under examensarbetets gång. Ett särskilt tack riktas också till underhållsingenjör Sören Ekström, som visat stort tålamod och lagt ner både tid och engagemang i examensarbetet. Med Sörens hjälp har det varit möjligt att göra arbetet mer anläggningsspecifikt. Även professor Bengt Carlsson får ett tack för värdefulla tips angående metodapplicering.

2 Bakgrund

Vattenfall AB är en av de fyra största elproducenterna i Europa och driver sammanlagt 102 vattenkraftverk av skiftande storlek runt om i Sverige. Vattenkraften utgör, näst efter kärnkraften, den största energikällan i Vattenfall AB:s elproduktion. Inom Vattenfall AB Vattenkraft sker underhåll samt det fortlöpande tillsynsarbetet av vattenkraftanläggningarna genom så kallad rondning. Vid rondning utför lokal driftpersonal rutinmässig tillståndskontroll (TK) på plats i anläggningarna. Driftpersonalen har då möjlighet att upptäcka till exempel eventuella läckage eller andra fysiska störningar samtidigt som den ges tillfälle att läsa av mätvärden på givare som finns placerade runt om i anläggningarna. De avlästa mätvärdena förs in i det så kallade TK-systemet [1] via en handdator. Vattenkraftsanläggningarna är dessutom utrustade med en så kallad Programmable Logic Controller (PLC), som i vissa fall automatiskt läser av anläggningarnas samtliga givare varje sekund. Processdata skickas sedan i realtid, via satellit, till en driftcentral (DC), stationerad i antingen Storuman, Bispgården eller Voullerim, som dygnet runt sköter styrning och driftövervakning av vattenkraftverken. På så vis sker kontinuerligt en automatiserad driftövervakning. Vid behov kan tjänstgörande driftledare justera driften av ett vattenkraftverk eller larma ansvarig driftgrupp. Vattenfall Service Nord (VSN) är uppdelad i ett antal driftgrupper som utför det lokala underhållsarbetet samt lokal TK av vattenkraftverken via rondning.

2.1 Conwide

Det övervakningssystem som Vattenfall AB Vattenkraft använder sig av benämns Conwide System III och är utvecklat av Conwide AB. Conwide AB har specialiserat sig på att utveckla datorbaserade system för TK och besiktningsarbete inom kraftbranschen. Enligt Conwide AB fungerar Conwide System III som ett ”insamlings- och analysverktyg för underhållsoptimering av industriella anläggningar” (Conwide, 2006-11-16). TK-systemet är framtaget för att kunna samla in data från och behandla resultat av inspektioner och besiktningar utförda på utrustning som används i produktionen. I Conwide-systemet kan lokalt stationerad driftpersonal samt operatörer vid aktuell DC övervaka de mätvärden som samlats in från respektive givare. För att upptäcka avvikelser och störningar i anläggningarna, innan allvarliga konsekvenser uppstår, ger övervakningssystemet larm då olika gränsvärden över- respektive underskrids. Vidare finns enklare trendanalyser att tillgå i Conwide System III. Trendanalyserna medger grafisk representation av mätvärden under önskad tidsperiod för upp till fyra mätpunkter samtidigt.

2.1.1 Processdatahantering i Conwide System III

Insamlade processdata lagras elektroniskt på en, av Conwide AB tillhandahållen, central databas. Vid de anläggningar där mätvärdesavläsningen endast utförs vid rondning blir, av förståeliga skäl, rådande analysmöjligheter lidande av den grova tidsupplösningen, eftersom rondning endast sker ungefär en gång per vecka. I de fall där PLC:n används för automatisk mätvärdesavläsning är upplösningen betydligt bättre. Här läses istället

[1] Se avsnitt 2.1.

givarna av kontinuerligt och lagras tillfälligt i den så kallade anpassningsenheten[2] med sekundupplösning. Högupplöst data kan emellertid endast lagras i anpassningsenheten under en begränsad tidsperiod om cirka 28 dygn innan de hämtas till den centrala databasen. Den centrala databasen är dock omodern och av utrymmesskäl lagras där endast ett värde[3] per dygn och givare.

2.2 Tillståndskontrolldatabas

Vattenfall Utveckling AB[4] drev år 2005 ett projekt för att se om data kunde lagras på ett mindre utrymmeskrävande sätt. Anledningen till projektet var att utvärdera moderna metoder för insamling och lagring av mätdata i en central databas på sekundnivå. Resultatet av projektet blev en så kallad TK-databas. Den är numera kopplad till två av Vattenfall AB:s vattenkraftverk. Med TK-databasen medges lagring av högupplöst[5] data över längre tid; åtminstone ett flertal år. Sedan mars månad år 2006 har TK-databasen varit i drift och processdata från nyss nämnda kraftverk finns därmed från dess fram till idag. Utöver lagring av processdata används inte TK-databasen för något annat ändamål eftersom den fortfarande bara utgör en testplattform. TK-databasen tillkom enbart för att utvärdera möjligheterna för lagring av processdata och tanken var initialt att den skulle tas ur bruk efter testperioden. I samband med projektet 2005 skapades även programmet ExportTKData vilket genom SQL-förfrågningar kan hämta efterfrågad data ur TK-databasen. Det är ExportTKData som använts inom examens-arbetet för att tillgå processdata.

2.3 Analyser som görs idag

De enkla trendanalyser som medges i Conwide System III används idag sällan och i liten utsträckning. Främst är det driftgrupperna som, för att kontrollera anläggnings-driften över tid, använder sig av trend-verktyget i Conwide-systemet. Utöver trend-analyserna görs inga analyser av de insamlade processdata. Kunskap om samband inom anläggningarnas drift baseras därför mer på erfarenhetsmässiga lärdomar som drift-personalen tillskansat sig genom att arbeta i anläggningen under en längre tid. Förståelsen för hur delar av anläggningarna påverkar varandra blir därför förhållandevis grovhuggen, då processen är uttalat dynamisk med exempelvis inneboende tidsförskjutningar. Vid processförändringar påverkas nämligen inte anläggningens samtliga delar samtidigt. Exempelvis varierar uttagen effekt i regel förhållandevis snabbt mellan olika värden, medan det tar längre tid för en oljetemperatur att ändras.

2.4 Larmnivåer

För många av de mätvärden som samlas in från vattenkraftsanläggningarna finns det övre respektive undre larmnivåer. Nivåerna bestäms initialt av produktleverantören men kan justeras av driftpersonalen. Utifrån yttre omständigheter, såsom årstidsväxlingar, kan därmed larmnivåerna anpassas för att fungera så väl som möjligt för stunden. Två typer av larm används; mjuka respektive hårda larm. Ett mjukt larm är en mildare grad

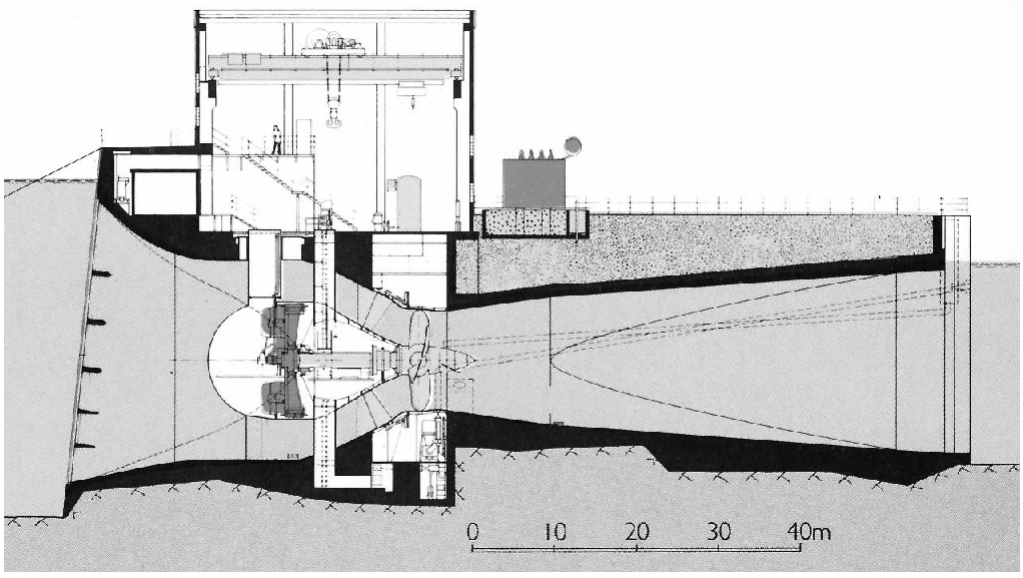
[2] Mer om anpassningsenheten kan läsas i avsnitt 2.5.1.
[3] dygnsmedelvärde
[4] numera Vattenfall Research and Development AB
[5] på sekundnivå

av larmnivå och leder endast till att personalen på driftcentralen underrättas om larmnivåöverskridelsen. DC kan då, utifrån aktuell situation, avgöra om driftpersonal behöver åtgärda något på plats vid anläggningen eller ej. Den andra typen av larm, det hårda larmet, är en starkare grad av larm och innebär att systemet slår ifrån automatiskt för att undvika ett eventuellt haveri.

Flera brister kan påpekas med det befintliga larmsystemet. Som larmen är utformade blir systemet påtagligt statiskt. Hänsyn tas inte till någon samverkan mellan variabler. Särskilt start respektive stopp av anläggningar kan innebära kritiska situationer. Olika variablers mätvärden varierar då kraftigt sinsemellan och stora påfrestningar och förändringar sker i anläggningen. Temperaturökningar samtidigt som uttagen effekt sjunker skulle exempelvis kunna undgå detektion, trots att det skulle kunna vara en indikation på ett olämpligt processtillstånd (PT). Ett antal tidigare examensarbeten[6] gjorda för Vattenfall AB Vattenkraft behandlar dynamiska larm mer ingående.

2.5 Det aktuella vattenkraftverket

Examensarbetet är baserat på processdata från en specifik vattenkraftsanläggning. Den aktuella anläggningen är uppkopplad mot TK-databasen vilket innebär att högupplöst processdata därför finns från en längre tidsperiod än för Vattenfall AB Vattenkrafts övriga vattenkraftverk. Samtidigt är datainsamlingen vid det aktuella vattenkraftverket särskilt omfattande på grund av att det i samband med ett tidigare utvecklingsprojekt installerades ett antal nya givare i anläggningen.



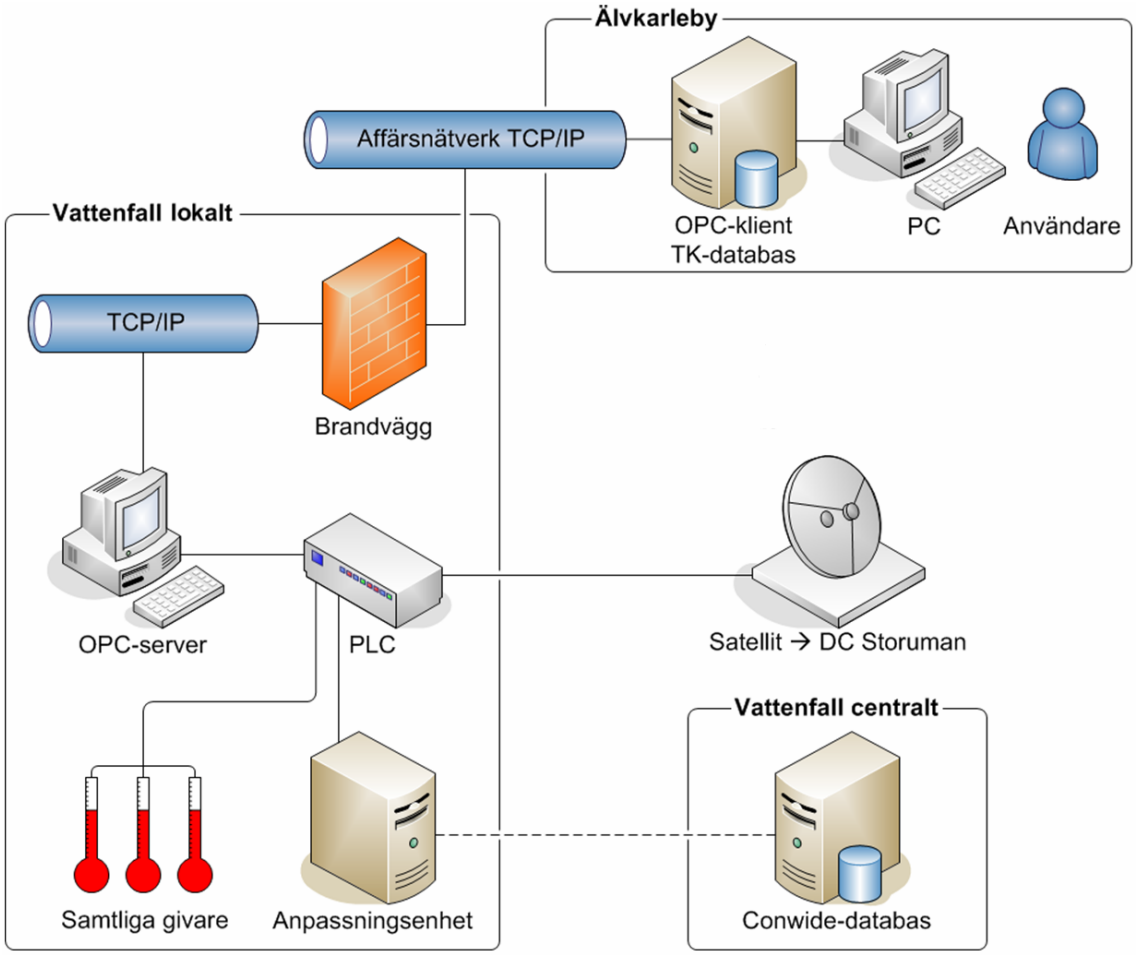
Figur 2.1 Schematisk bild över vattenkraftverket.

Vattenkraftverket består av två generatorer, G1 och G2, vilka båda är så kallade bulbturbiner[7], som lämpar sig väl då fallhöjden är låg och flödet stort. Ovan, i figur 2.1,

[6] Se till exempel Masman & Blom (2000) samt Glemme (2000).
[7] Rörkaplanturbin är ett annat namn på bulbturbin.

visas en schematisk bild över vattenkraftverket. Med en årlig elproduktion på 100 GWh rör det sig om ett av de mindre vattenkraftverken i Sverige.

Runt om i kraftverket finns givare avsedda att mäta tillståndet hos olika delar av anläggningen. Kraftverkets båda generatorer har 76 givare vardera. Dessutom finns ytterligare 37 givare i andra delar av kraftverket. Bilaga 1 utgör en förteckning över samtliga givare vid det aktuella vattenkraftverket. Givarna mäter bland annat olika temperaturer, tryck, effekter och nivåer och PLC:n registrerar automatiskt ett värde från samtliga givare varje sekund. I figur 2.2 visas en schematisk bild över hur processdata hanteras vid kraftverket. Figuren kan vara till hjälp för att lättare förstå det resonemang som följer i avsnitt 2.5.1.



Figur 2.2 Schematisk bild över hanteringen av processdata från det aktuella vatten-kraftverket.

2.5.1 Processdatahantering vid vattenkraftverket

PLC:n i det aktuella vattenkraftverket kan ses som en processdatakoordinator. Den insamlar, hanterar och distribuerar processdata. Processdata hämtas från PLC:n dels till DC i Storuman via satellit, dels till anpassningsenheten och dels till en OPC-server[8].

[8] OLE for Process Control, där OLE står för Object Linking and Embedding, är en standard för kommunikation mellan datorer.

Anpassningsenheten är den del av Conwide-systemet som lagrar processdata i upp till 28 dygn. Varje dygn hämtas dessutom processdata, i form av dygnmedelvärden från de respektive givarna, från anpassningsenheten till Conwide-systemets databas[9]. En OPC-klient stationerad i Älvkarleby hämtar i sin tur processdata[10] från OPC-servern. OPC-klienten lagrar all processdata i TK-databasen, vilken operatörer kan tillgå via en PC för vidare hantering av processdata.

Varje sekund registrerar PLC:n mätvärden från respektive givare runt om i vattenkrafts-anläggningen. Samtliga mätvärden tillsammans utgör den processdata som varje dygn dels lagras i TK-databasen, dels lagras inom Conwide System III. Att kopplingen mellan anpassningsenhet och Conwide-databas, i större utsträckning än övriga kopplingar, är förenklad indikeras i figur 2.2 av den streckade linjen.

2.6 Notation

Ur TK-databasen hämtas data, med hjälp av ExportTKData, i matrisform med värden från respektive givare presenterade kolumnvis. Varje rad representerar således en tid-punkt, observation, och den totala processdatamatrisen från en generator har följaktligen storleken $[m \times 76]$ där m anger antalet observationer.

I rapporten beskrivs processvariablerna i regel på vektor- eller matrisform. För att enkelt kunna särskilja skalärer, vektorer och matriser följer här en beskrivning av den notation som används genomgående i rapporten. Kursiv gemen (exempelvis x) betecknar skalär. Vektorer symboliseras av fetstilt gemen (exempelvis $\mathbf{x}$) och matriser betecknas av fetstilt versal (exempelvis $\mathbf{X}$). In- och utdata betecknas med x, $\mathbf{x}$ eller $\mathbf{X}$ respektive y, $\mathbf{y}$ eller $\mathbf{Y}$. $x_i(k)$ eller $y_i(k)$ avser värdet på den i:te in- eller utdatavariabeln vid tidpunkt k.

[9] Detaljer kring kopplingen mellan anpassningsenheten och Conwide-systemets databas är inte av vikt för det här arbetet och har därför förenklats kraftigt i figur 2.
[10] För att minimera nättrafiken hämtas endast förändrade mätvärden.

3 Data Mining

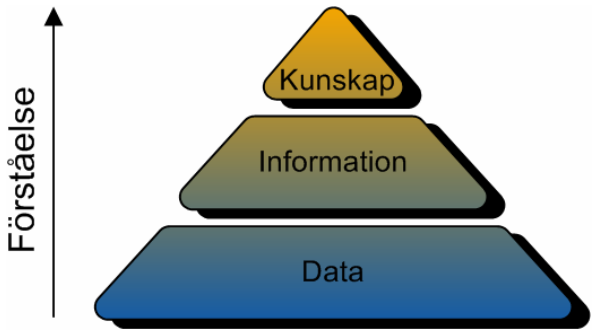
Det finns ingen entydig tolkning av begreppet Data Mining[11] (DM). Uttrycket har sitt ursprung i skilda ämnesområden, såsom matematisk statistik och artificiell intelligens (AI) och beroende på betraktarens teoretiska bakgrund tolkas DM-begreppet på olika vis. En statistiker och en person med datavetenskaplig bakgrund kan båda vara överens om att de arbetar med skilda saker. Likväl skulle de båda två, samtidigt kunna hävda att de arbetar med DM-relaterade metoder. Flera olika definitioner av DM förekommer nämligen i litteraturen. Den definition som valts för examensarbetet är hämtad från Chen & Chen (2006) och begränsar inte DM till att baseras på enbart ett av ovanstående ämnesområden;

> ”Efficient extraction of interesting (valid, non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases.”

Ämnesområdet DM är inget nytt fenomen, men har i samband med ny och mer kraftfull teknik kommit att användas på sätt som tidigare inte varit möjliga. I litteraturen florerar andra benämningar på metoder som i det här arbetet faller inom begreppet DM. Några andra vanligt förekommande termer är Knowledge Discovery in Databases (KDD)[12], knowledge extraction, data archeology, pattern analysis och information harvesting (Sagonas, 2005).

3.1 Data – information – kunskap

För att undvika begreppsförvirring, som annars kan uppstå vid diskussioner kring till synes närbesläktade koncept, såsom data, information och kunskap, följer här en kortare redogörelse över den tolkning som använts genomgående i examensarbetet.



Figur 3.1 Förhållandet mellan data, information och kunskap.

Det som skiljer begreppen åt är graden av förståelse. Data inbegriper ingen förståelse över huvud taget medan kunskap baseras på en mycket god förståelse. För ytterligare

[11] sv. informationsutvinning

[12] KDD avser i regel hela processen att omvandla data till kunskap och bland annat Fayyad et al. (1996) och Sagonas (2005) menar att DM endast är ett av stegen i omvandlingsprocessen. DM och KDD är sålunda inte en och samma sak. Här är dock avsikten endast att lyfta fram olika begrepp. Grundligare genomgång av betydelserna görs därför inte. Se exempelvis Fayyad et al. (1996) för en mer utförlig beskrivning av KDD.

klarhet ges i figur 3.1 en beskrivande bild över relationen mellan data, information och kunskap. Bilden visar att information, i det här sammanhanget, påträffas mellan data och kunskap, vilket indikerar att graden av förståelse förenad med information varken är obefintlig eller väldigt hög.

En korrekt tolkning av bilden är att data är grunden för all kunskap. Kunskap baseras alltid på information som i sin tur baseras på data. Nedan följer specifika exempel på hur data, information respektive kunskap, i fallet med drift av ett vattenkraftverk, skiljs åt, under ovan beskrivna tolkning.

Data: Data utgörs helt enkelt av de mätvärden som registreras av givarna i anläggningen. Enbart mätvärden säger i princip ingenting om driften av vattenkraftverket. Graden av förståelse är obefintlig.

Information: Förståelse för hur vattenkraftverket är uppbyggt innebär att mätvärden kan sättas i relation till varandra. Med ökad förståelse kan också en serie mätvärden kopplas samman och skapa trendkurvor över tid för respektive variabelvärde. Den ökade förståelsen ger information om kraftverksdriften. Graden av förståelse är dock fortfarande begränsad.

Kunskap: Vidare insikt i samband mellan variabler ger förståelse för hur variabelvärden påverkar varandra. Förståelse om varför variationer i variabelvärden uppstår och hur de påverkar kraftverksdriften är kunskap om driftsprocessen. Graden av förståelse är mycket hög.

För att öka kunskapen, i det här fallet om drift av vattenkraftverk, behövs således en god grund av data. Information kan i regel utvinnas ur data med förhållandevis enkla medel. Enligt ovan handlar det i det här fallet om att relatera data till exempelvis plats och tidpunkt och på så vis sätta den i ett större sammanhang. Att sedan erhålla kunskap ur den utvunna informationen är avsevärt mer mödosamt. Människan klarar inte alltid av det på egen hand utan behöver verktyg för att skapa sig den grad av förståelse som krävs för att informationen ska kunna betraktas som kunskap. Det är här examensarbetet har sin fokus. En stor mängd processdata är redan insamlad från det aktuella vattenkraftverket och tillräcklig förståelse finns för att all processdata ska ge information om driften. Viss kunskap om driften finns naturligtvis också, men Vattenfall AB Vattenkraft vill nu förskansa sig ytterligare förståelse för att öka kunskapsnivån. Tanken är att i examensarbetet utreda vilka möjligheter som finns för att öka driftförståelsen med hjälp av metoder inom området för DM.

Trots att det naturligtvis är kunskap, med hög grad av förståelse, som eftersträvas, kan inte vikten av data negligeras eftersom all kunskap i grund och botten baseras på just data. Garbage In, Garbage Out (GIGO) är en aforism som används inom datavetenskapen och som passar väl in i sammanhanget. Endast korrekt data kan leda till god kunskap, på samma sätt som inkorrekt data garanterat inte leder till värdefull kunskap. Det är därför av stor vikt att data är av god kvalitet. Det är samtidigt viktigt att medvetenheten om eventuella brister i data är hög, för att en lämplig bedömning av kunskapens tillförlitlighet ska kunna göras.

3.2 Användningsområden

Definitionen av DM given ovan inskränker inte användning av DM-tekniker till något specifikt affärsområde. Metoder för DM är egentligen enbart beroende av tillgången på data snarare än typen av data. DM kan därför användas inom vitt skilda branscher och på vitt skilda typer av data. Det finns klassiska DM-metoder som, beroende på hur tillgänglig data är utformad, lämpar sig olika väl till att användas för lika klassiska DM-problem. För en insikt i hur DM kan användas på konkreta problemformuleringar ges här en sammanställning av ett par, mer eller mindre klassiska, uppgifter som, enligt bland andra Berry & Linoff (1997:5) och Sagonas (2005), med fördel löses med hjälp av DM-metoder.

3.2.1 Klassificering

Givet fördefinierade grupperingar av data kan metoder för klassificering tilldela nya observationer en klasstillhörighet. Metoderna baseras därmed på så kallad övervakad inlärning (Bishop, 1995:10) vilket innebär att tillgång till träningsdata finns. Metoderna kan först tränas på en del av befintlig data för att sedan testas mot resterande del av befintlig data. Innan metoderna används till att klassificera ny data kan de därmed jämföras utifrån hur de presterar och lämplig metod kan väljas. En typisk klassificeringsuppgift är exempelvis att klassificera olika ämnesprover med avseende på ett antal olika egenskaper. En av de mest omnämnda klassificeringsmetoderna kallas K Nearest Neighbour (KNN). Vid användning av KNN räknas först avståndet från varje okänt prov till varje redan klassificerat prov ut. Varje okänt prov tilldelas sedan klass utifrån vilka klasser de närmaste grannarna tillhör där K avgör antalet grannar som betraktas. Med hjälp av träningsdata kan ett lämpligt värde på K väljas så att klassificeringen blir så korrekt som möjligt i det specifika fallet.

3.2.2 Länkanalys

I en länkanalys är avsikten att finna samband[13] i data. Association Rules[14] och Sequence Mining[15] är två metoder med konkreta användningsområden. Associations-regelanalys används ofta (Sagonas, 2005) till så kallade market basket problem[16]. Data består då av ett antal transaktioner där varje transaktion är en samling av varor inhandlade av en kund vid ett köptillfälle. En associationsregelanalys leder fram till ett antal associationsregler[17] som, i fallet med kundvagnsproblemet, visar samband i kundernas köpvanor. Kunskap från analysen kan sedan, bland annat, användas till att effektivisera marknadsföring. I en sekvensbaserad analys kan data från exempelvis en möbelbutik utmynna i kunders inköpsmönster. Analysen kan indikera typer av varor, i det här fallet möbler, som kunder frekvent köper sekvensvis. Det skulle exempelvis gå att visa att många av de kunder som köper ett soffbord, vid ett tidigare tillfälle inhandlat en soffa. Kunskapen kan användas för riktad marknadsföring.

[13] Till exempel samband på formen “om A så B”.
[14] sv. associationsregelanalys
[15] sv. sekvensbaserad analys
[16] sv. kundvagnsproblemet
[17] {A} → {B} och {C, E} → {A, B, D} är två exempel på associationsregler.

3.2.3 Klustring

För att gruppera poster i data utifrån hur väl de liknar varandra finns klustringsmetoder. Med många variabler och poster i data är det svårt att manuellt göra en liknande indelning och då kan en klustringsmetod vara till stor hjälp. Klustring är ett exempel på oövervakad inlärning (Bishop, 1995:10) vilket innebär att det inte finns något fördefinierat mål att jämföra klustringsresultatet med. Användaren får själv avgöra vilken metod, och vilka parametrar, som lämpar sig bäst för aktuell data. Klustringsmetoder hittar grupper av objekt så att objekten liknar varandra inom grupperna samt skiljer sig åt mellan grupperna (Sagonas, 2005). Klustring kan liknas vid klassificering med den skillnaden att det vid klustring inte finns någon kunskap om antalet grupper eller skillnaden grupper emellan. Klustringsmetoder kan med fördel appliceras på liknande typer av data som klassificeringsmetoder. Det finns flera olika typer av klustringsmetoder (Ibid.). Centrumbaserade klustringsmetoder bygger grupper av data så att en grupp är en samling av dataobjekt så att ett objekt inom gruppen ligger närmare gruppens centrum än någon annan grupps centrum. En grupps centrum utgörs vanligtvis av en centroid[18] (Ibid.) och metoden kallas då *k*-means där *k* anger antalet kluster. En annan vanlig typ av klustringsmetoder baserar kluster med avseende på densitet. En vanligt förekommande densitetsbaserad klustringsmetod kallas Density Based Spatial Clustering of Applications with Noise (DBSCAN). Kluster definieras då istället som täta regioner av punkter som separeras från andra täta punktsamlingar av glesa områden (Ibid.).

3.3 Val av metod

DM är inte magi och förståelse för vilket problem som önskas lösas behövs för att veta vilken DM-metod som är lämplig att applicera på en specifik uppsättning data (Berson et al., 2000). Samtidigt är en DM-metod i regel utformad efter en viss typ av data och vilka problem som över huvud taget går att lösa är därför beroende på utformningen av tillgänglig data.

Litteratur om DM berör vanligtvis ovanstående problemställningar utifrån ett data-vetenskapligt perspektiv. Metoderna som presenteras för att angripa uppgifterna baseras därför i regel främst på datavetenskapligt kunnande där kodning av datorprogram är den centrala delen av arbetet. I enlighet med vad som nämnts tidigare kan istället ett statistiskt perspektiv appliceras på problemen (Giudici, 2003:129). Det är framförallt det statistiska perspektivet som varit utgångspunkten för examensarbetet. De metoder som, inom arbetets ramar, studerats mer ingående finns frekvent presenterade i litteratur[19] som berör multivariat analys av data. Med anledning av de stora datamängderna har dock datavetenskapliga metoder varit nödvändiga för att lösa de mer statistiska problemen. Se kapitel 5 för en mer detaljerad beskrivning av använd programvara.

[18] "Medelpunkten av ett objekt definierad som medelvärdet av koordinaterna för alla de punkter som ingår i det." (Nationalencyklopedin, sökord "centroid", 2006-11-16)
[19] Se till exempel Hastie et al. (2001) och Eriksson et al. (2006).

4 Principalkomponentanalys (PCA)

En, av många, vanligt förekommande anledningar till att analyser av data inte alltid är enkla att utföra i praktiken är att antalet variabler är stort. Med fler än fem variabler brukar data betecknas multivariat (Eriksson et al., 2006:22). När antalet variabler i data växer, ökar samtidigt komplexiteten. Ett problem som uppstår i samband med ökande komplexitet är att det blir allt svårare att få en övergripande bild av variationer hos data. Då varje variabel spänner upp en ny dimension blir det, när antalet dimensioner överstiger tre, en orimlig uppgift att försöka överblicka variationerna. För att kunna säga något om variationerna är det därför nödvändigt att använda någon form av hjälpmedel. I följande kapitel beskrivs en dimensionsreducerande metod. Med hjälp av metoden blir det betydligt lättare att analysera händelser i processen baserat på registrerad data och metoden är därför lämplig att använda vid driftövervakning. Det är vid användande av den beskrivna metoden som tyngdpunkten för examensarbetet ligger.

4.1 En statistisk modell

Idag finns många olika tekniker inom ramen för DM. En vanligt använd teknik, med rötterna inom statistiken, är principalkomponentanalys (PCA). Vid PCA görs mång-variat data mer överskådlig genom att projicera den på ett mindre antal så kallade principalkomponenter (PC:er). I praktiken innebär det att processituationen på ett enklare sätt kan studeras, varpå avvikande PT kan detekteras.

Som tidigare nämnts representerar varje rad i processdatamatrisen en unik tidpunkt. För varje tidpunkt finns värden från givarna registrerade. Varje tidpunkt kan på så vis sägas inrymma information om rådande tillstånd hos respektive variabel. Fortsättningsvis kommer varje tidpunkt att ses representera ett specifikt PT. Genom att jämföra PT med vad som kan betraktas som normaltillstånd är tanken att anomalier i driften ska kunna upptäckas. För människan, som har svårt att föreställa sig fler dimensioner än tre, blir den här jämförelsen inte lätt att utföra eftersom antalet givare i vattenkraftverket vida övertiger tre. Att enbart studera värden från tre givare och utesluta resterande givarvärden är inte en tilltalande lösning på problemet eftersom det skulle innebära att potentiellt viktig information utesluts ur analysen. Eriksson et al. (2006:25) poängterar vikten av att inte utesluta data eller variabler eftersom det främst är i korrelationsmönster mellan variabler, snarare än bland individuella variabelsignaler, som det går att finna intressant information. Istället kan PCA vara ett kraftfullt verktyg för att underlätta jämförelser av olika PT, där hänsyn är tagen till samtliga av de tillgängliga data.

PCA utformades initialt av Cauchy[20] (Eriksson et al., 2006:39), men presenterades inom statistiken först av Pearson[21] som beskrev PCA i termer av linjer och plan med bäst passning till system av punkter i rymden (Jackson, 1991:13). Idag utgör PCA basen för multivariat data-analys (Rosén, 1998:63). Den centrala idén med PCA är att reducera antalet dimensioner hos en datamängd, bestående av ett stort antal variabler som står i

[20] Augustin Louis Cauchy (1789-1857), fransk matematiker (Råde & Westergren, 1998:519).
[21] Karl Pearson (1857-1936), brittisk statistiker och professor i bland annat tillämpad matematik och mekanik (Nationalencyklopedin, sökord ”Pearson, Karl”, 2006-11-16).

ett inbördes förhållande till varandra, samtidigt som så mycket som möjligt av variationen i datamängden behålls (Jolliffe, 1986:1). Vid PCA transformeras data till en uppsättning nya, okorrelerade variabler, PC:er, ordnade så att de första PC:erna beskriver den mesta av variationen bland samtliga av de ursprungliga variablerna (Ibid.). Multivariat data kan då representeras av ett lågdimensionellt plan så att det blir möjligt att överblicka data (Eriksson et al., 2006:39). Överblicken kan avslöja eventuella grupper av observationer, trender, outliers[22] samt samband mellan observationer och variabler och mellan variablerna själva (Ibid.). PCA är framförallt en deskriptiv metod (Jolliffe, 1986:205).

4.2 Matematisk tolkning av PCA

I följande avsnitt förklaras matematiken bakom PCA. För ökad tydlighet kommer metoden att beskrivas grafiskt i ett senare avsnitt.

Om **X** är en enhetsskalad och medelcentrerad[23] processdatamatris med *m* rader och *n* kolumner, där variabelvärden utgör kolumner och observationer utgör rader, kan **X** uttryckas som summan av *r* $\mathbf{t}_i$ och $\mathbf{p}_i$ vektorpar, där *r* anger rangen av **X**;

$$\mathbf{X} = \mathbf{t_1}\mathbf{p_1}^T + \mathbf{t_2}\mathbf{p_2}^T + \ldots + \mathbf{t}_a\mathbf{p}_a^T + \ldots + \mathbf{t}_r\mathbf{p}_r^T \tag{1}$$

r måste här vara mindre eller lika med den mindre dimensionen av **X**; $r \leq \min\{m,n\}$ (Wise et al., 2006:99). Normalt sett trunkeras PCA-modellen efter *a* PC:er och återstående variationsriktningar förenas då i residualmatrisen **E** (Wise et al., 2006:99);

$$\mathbf{X} = \mathbf{t_1}\mathbf{p_1}^T + \mathbf{t_2}\mathbf{p_2}^T + \ldots + \mathbf{t}_a\mathbf{p}_a^T + \mathbf{E} \tag{2}$$

eller

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{3}$$

$a = r$ ger $\mathbf{E} = 0$ eftersom alla variationsriktningar då beskrivs (Rosén, 1998:64). $\mathbf{t}_i$ - och $\mathbf{p}_i$ -vektorparen sorteras efter den mängd varians de fångar i **X** (Ibid.). **T**:s vektorer $(\mathbf{t}_i)$ kallas för scores och innehåller information om hur observationerna förhåller sig till varandra medan **P**:s vektorer $(\mathbf{p}_i)$ kallas för loadings och innehåller information om variablernas inbördes relationer (Ibid.).

Matematiskt sett bygger PCA på en egenvärdesuppdelning av variablernas kovariansmatris. För den givna processdatamatrisen **X**, med *m* rader och *n* kolumner, definieras kovariansmatrisen som

$$\mathrm{cov}(\mathbf{X}) = \frac{\mathbf{X}^T\mathbf{X}}{m-1} \tag{4}$$

[22] En outlier är en enskild observation långt från övriga observationer. Det stora avståndet beror i regel på någon form av icke naturlig avvikelse, varför outliers bör hanteras särskilt. Mer om hantering av outliers följer i avsnitt 5.1.4.1.

[23] Se avsnitt 4.4.1 för en utveckling av begreppen enhetsskalning och medelcentrering.

under förutsättning att kolumnerna i **X** är medelcentrerade (Wise et al., 2006:100.). En kolumnvektor $\mathbf{p}_i$ av **P** är den *i*:te egenvektorn av kovariansmatrisen $\text{cov}(\mathbf{X})$ så att, för varje $\mathbf{p}_i$,

$$\text{cov}(\mathbf{X})\mathbf{p}_i = \lambda_i \mathbf{p}_i \tag{5}$$

där λ_i är egenvärdet associerat med egenvektorn $\mathbf{p}_i$ (Ibid.). Noterbart är att

$$\mathbf{X}\mathbf{p}_i = \mathbf{t}_i \tag{6}$$

för **X** och varje par av $\mathbf{t}_i$ och $\mathbf{p}_i$. Med andra ord är score-vektorn $\mathbf{t}_i$ den linjärkombination av de ursprungliga **X**-variablerna som definieras av $\mathbf{p}_i$. Det innebär att $\mathbf{t}_i$ är projektionerna av **X** på $\mathbf{p}_i$ (Ibid.). $\mathbf{t}_i$- och $\mathbf{p}_i$-vektorparen sorteras i avtagande ordning med avseende på det egenvärde, λ_i, som associeras med kolumnvektorn $\mathbf{p}_i$. λ_i är ett mått på variansen som förklaras av $\mathbf{t}_i$- och $\mathbf{p}_i$-vektorparen. I den här kontexten kan varians ses som information (Ibid.). Det första $\mathbf{t}_i$- och $\mathbf{p}_i$-vektorparet fångar den största mängd information som är möjlig ur **X**. Efterföljande vektorpar fångar i sin tur den största mängd information ur variationerna som återstår i **X** när $\mathbf{t}_i\mathbf{p}_i^T$ är subtraherad (Rosén, 1998:65). Utan betydande informationsförlust kan därför processdata i **X** beskrivas av betydligt färre PC:er än antalet ursprungliga variabler. Score-vektorerna består av de ursprungliga observationerna projicerade på det nya koordinatsystem, den PC-rymd, som definieras av PC:erna. För att projicera nya observationer på PC-modellen räknas scores ut enligt ekvation 7 (Rosén, 1998:65).

$$\hat{\mathbf{T}} = \mathbf{X}_{ny}\mathbf{P} \tag{7}$$

Värt att notera är att testdata, naturligtvis, måste medelcentreras och enhetsskalas innan beräkning av scores är möjlig.

4.2.1 Statistisk passning till modellen

Det finns olika sätt att validera den PCA-modell som konstruerats. Två av de vanligaste metoderna förklaras i nedanstående avsnitt. Med hjälp av metoderna är det möjligt att grafiskt representera residualvektorerna, vilket ska visa sig värdefullt för att bland annat detektera outliers.

4.2.1.1 Squared prediction error (SPE)

Squared prediction error (SPE) är det euklidiska avståndet från en observation till det hyperplan som utgörs av PC:erna. En valid modell ska ge små avstånd mellan modellplanet och observationerna eftersom den avvikande riktningen inte är inkluderad i modellen. SPE är ett mått på hur observationerna varierar ortogonalt[24] mot modellplanet (Rosén, 1998:71) och kan ses som en residualvariabel som indikerar hur väl varje observation passar PCA-modellen. SPE är vidare summan av kvadraterna av varje rad i **E** (se ekvation 2) och beräknas enligt

[24] ortogonal = vinkelrät

$$SPE(k) = \mathbf{e}(k)\mathbf{e}^T(k) \tag{8}$$

där $\mathbf{e}(k)$ är den *k*:te raden i **E** (Wise et al., 2006:102). En enskild rad i **E**, $\mathbf{e}(k)$, representeras av SPE-bidragen från respektive variabel. Genom att studera SPE-bidragen för en specifik observation kan information om vilka variabler som bidrar mest till residualfelet erhållas (Wise et al., 2006:102). Ekvation 8 kan vid PCA även skrivas enligt ekvation 9 när $\mathbf{t}(k)$ byts ut mot $\mathbf{x}(k)\mathbf{P}$ enligt ekvation 7 (Rosén, 1998:70).

$$SPE(k) = \sum_{j=1}^{n}\left(x_j(k) - \mathbf{t}(k)\mathbf{p}_j^T\right)^2 = \sum_{j=1}^{n}\left(x_j(k) - \mathbf{x}(k)\mathbf{P}\mathbf{p}_j^T\right)^2 \tag{9}$$

För en PCA-modell kan, enligt Jackson (1991:36), konfidensgränsen för SPE räknas ut enligt

$$SPE_\alpha = \Theta_1\left[\frac{c_\alpha\sqrt{2\Theta_2 h_0^2}}{\Theta_1} + 1 + \frac{\Theta_2 h_0(h_0 - 1)}{\Theta_1^2}\right]^{\frac{1}{h_0}} \tag{10}$$

där

$$\Theta_i = \sum_{j=a+1}^{n}\lambda_j^i \qquad \text{för } i = 1, 2, 3 \tag{11}$$

samt

$$h_0 = 1 - \frac{2\Theta_1\Theta_3}{3\Theta_2^2} \tag{12}$$

c_α i ekvation 10 är $N_{1-\alpha}(0,1)$ där α anger signifikansnivån och *a* i ekvation 11 är antalet PC:er som inkluderats i modellen.

4.2.1.2 Hotellings T²

Ett annat mått på nya observationers passning till modellen är Hotellings[25] T² vilket är den normaliserade summan av scores (Rosén, 1998:71). T² är ett mått på avståndet mellan en observation och origo[26] för det koordinatsystem som PC:erna bygger upp. Därmed beskriver T² hur data varierar inom PCA-modellen (Ibid.). T² vid tidpunkt *k* beräknas enligt

$$T^2(k) = \mathbf{t}(k)\Lambda^{-1}\mathbf{t}^T(k) = \mathbf{x}(k)\mathbf{P}\Lambda^{-1}\mathbf{P}^T\mathbf{x}^T(k) \tag{13}$$

där $\mathbf{t}(k)$ anger scores vid tid *k* och Λ^{-1} är en diagonalmatris med egenvärden till de *a* PC:erna som inkluderas i modellen (Wise et al., 2006:103). Genom att studera vilken av score-vektorernas bidrag till T² som dominerar för en specifik observation är det möjligt att avgöra vilka variabler det är som framförallt avviker från vad som är modellerat.

[25] Harold Hotelling (1895-1973), amerikansk matematiker, statistiker och nationalekonom (Nationalencyklopedin, sökord ”Hotelling, Harold”, 2006-11-16)
[26] Snarare den punkt i vilken koordinataxlarna korsas, men eftersom **X** här antas vara medelcentrerad sammanfaller den punkten med origo.

Med andra ord kan avvikelser längs en specifik PC avslöja avvikande variabelvärden. Ekvation 14 visar hur PC:ernas bidrag till T^2 beräknas.

$$T^2(k) = \frac{t_1^2(k)}{\lambda_1} + \frac{t_2^2(k)}{\lambda_2} + \ldots + \frac{t_n^2(k)}{\lambda_n} \tag{14}$$

$t_i(k)$ är här score-värdet och λ_i är egenvärdet associerat med den *i*:te PC:en (Rosén, 1998:87). Den PC som bidrar mest till T^2 är den PC som förknippas med den största termen i ekvation 14. Det är nu möjligt att beräkna hur mycket respektive variabel bidrar till avvikelser längs de PC:er vars bidrag till T^2 är störst. Ekvation 15 förklarar hur mycket respektive variabel bidrar till avvikelserna längs en viss PC.

$$\mathbf{c}(k) = \mathbf{x}(k)\mathbf{P}_j \tag{15}$$

$\mathbf{c}(k)$ är en vektor med bidragen från variablerna vid tid *k* och $\mathbf{P}_j$ är en diagonalmatris av radvektorn $\mathbf{p}_j$ (Rosén, 1998:88). Konfidensgränsen för T^2 fås ur F-fördelningen enligt

$$T^2_{a,m,\alpha} = \frac{a(m-1)}{m-a} F_{k,m-a,\alpha} \tag{16}$$

där *m* är antalet observationer i modellen och *a* anger antalet PC:er (Wise et al., 2006:104).

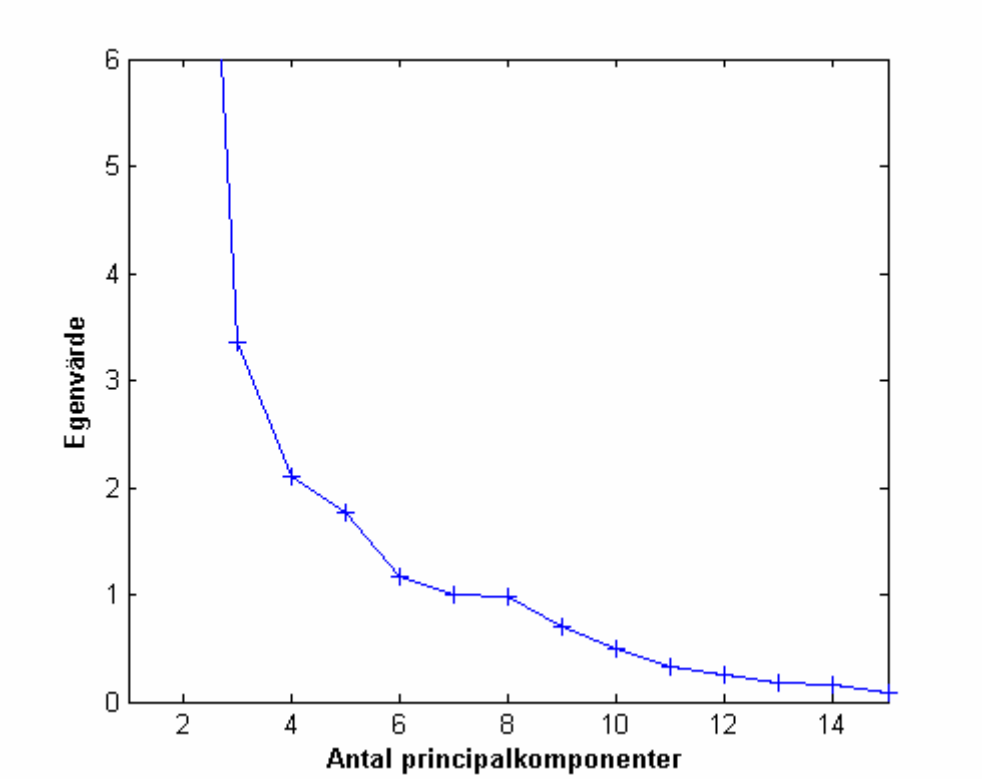
4.3 Antalet PC:er

Det främsta målet vid PCA är att ersätta de *n* elementen i **x** med ett mycket mindre antal, *a*, PC:er utan att försaka betydande information. Att använda *a* PC:er istället för *n* variabler reducerar kraftigt antalet dimensioner om $a << n$, men samtliga *n* variabler behövs i regel för att beräkna PC:erna eftersom varje PC kan vara en funktion av alla *n* variabler (Jolliffe, 1986:92). Ett övervägande som måste göras vid PCA är att avgöra hur många PC:er som PCA-modellen ska baseras på för att den mest betydande variationen i **X** ska bibehållas samtidigt som antalet dimensioner reduceras i så hög utsträckning som möjligt. Då PC:erna ordnas efter hur pass väl de beskriver variationen i **X** ersätts de *n* variablerna av de *a* första PC:erna. Det finns ett antal ad hoc-metoder som vanligtvis används som tumregler för bestämning av lämpligt värde på *a* (Jolliffe, 1986:93).

4.3.1 Scree-plot

Den metod som oftast[27] hänvisas till baseras på ett subjektivt val av operatören utifrån en så kallad scree-plot. Scree-plotten är helt enkelt ett diagram i vilket variansen som förklaras av varje PC, med andra ord egenvärdena till kovariansmatrisen av **X**, är representerad som en graf (The MathWorks, 2006b:8.10). Figur 4.1 visar ett exempel på en scree-plot.

[27] Se exempelvis Jolliffe (1986:96-97), Rosén (1998:79-80), Jackson (1991:45-46) samt Wise et al. (2006:115ff).



Figur 4.1 Scree-plot. Utifrån grafen kan antalet PC:er bestämmas. Grafen indikerar att 6 PC:er bör inkluderas i modellen.

Valet av a görs, enligt teorierna, med fördel utifrån scree-plotten. Enligt Jolliffe (1986:96) väljs a så att lutningen på linjen är stor till vänster om a och liten till höger om a. Anledningen till att det är ett lämpligt kriterium för valet av a är att det inte är eftersträvansvärt att inkludera ytterligare en PC om den inte bidrar märkbart till att förklara variationer i **X**. Ett problem som kan uppstå är om inget tydligt så kallat knä framkommer i scree-plotten. Det kan då vara användbart att konsultera ytterligare en metod för bestämning av värdet på a.

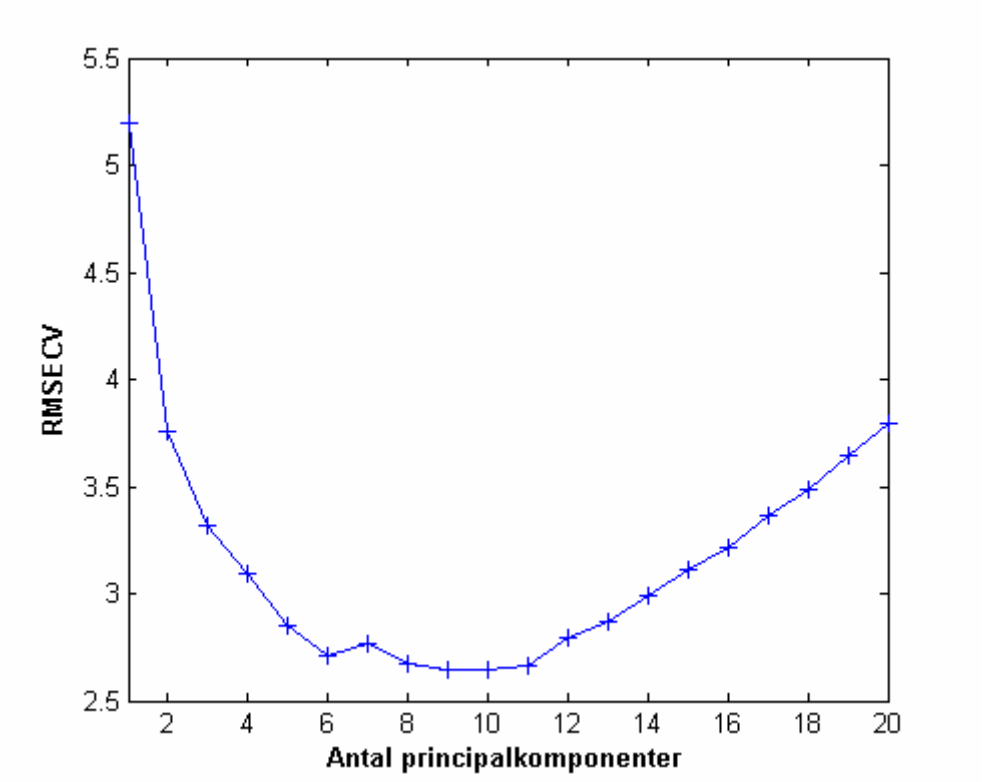
4.3.2 Korsvalidering

En annan vanligt förekommande metod är korsvalidering. Vid korsvalidering delas observationerna ur **X** upp i, säg g, grupper. En i taget utesluts sedan grupperna ur **X** varvid PCA genomförs på de övriga $g-1$ grupperna och samtliga n PC:er beräknas. Därefter bestäms SPE vid modellering med 1 till och med n PC:er för samtliga observationer i den uteslutna gruppen, g_i. Proceduren upprepas för samtliga g grupper varefter m SPE-värden för var och en av de n PCA-modellerna erhålls. SPE-värdena summeras därefter för respektive PCA-modell varefter SPE-summorna divideras med produkten av antalet variabler, n, och antalet observationer, m[28]. De erhållna värdena kallas för Predicted Residual Error Sum of Squares (PRESS-värden) (Jackson, 1992:353-354). Utifrån PRESS-värdena kan en graf över Root-Mean-Square Error of Cross-Validation (RMSECV) skapas. Förhållandet mellan PRESS- och RMSECV-värden ser ut enligt ekvation 17

[28] SPE-medelvärdet för respektive PCA-modell divideras därmed med antalet variabler, n.

$$RMSECV_k = \sqrt{\frac{PRESS_k}{n}} \tag{17}$$

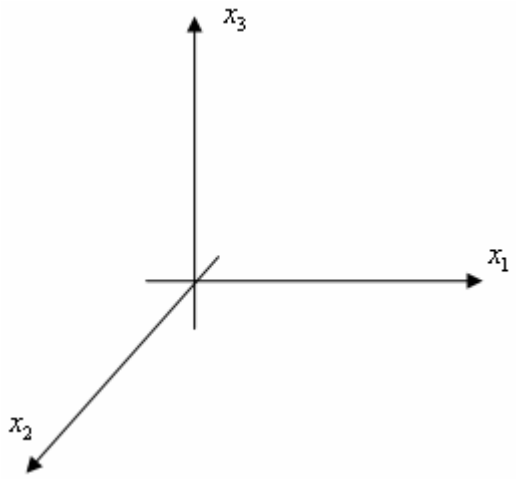
där k anger antalet PC:er inkluderade i modellen och n anger antalet observationer (Wise et al., 2006:156). I figur 4.2 visas ett exempel på en RMSECV-plot. RMSECV-värdena kan ses som ett mått på modellens lämplighet för observationer som inte använts vid modellkonstruktionen (Wise et al., 2006:156).



Figur 4.2 *RMSECV-plot. 6 PC:er vore enligt grafen ett bra val. Värdet på RMSECV är då så lågt som möjligt samtidigt som antalet PC:er minimeras.*

4.4 Geometrisk tolkning av PCA

Betänk en datamatris **X** med m observationer och n variabler. Varje variabel representerar en koordinataxel i variabelrymden. För klarhets skull kommer fallet med tre variabler, $n = 3$, att betraktas. Variablerna spänner då upp ett rum där variabelvärdena anger koordinater för observationerna. Figur 4.3 visar hur de tre variablerna utgör koordinatsystem i rummet.



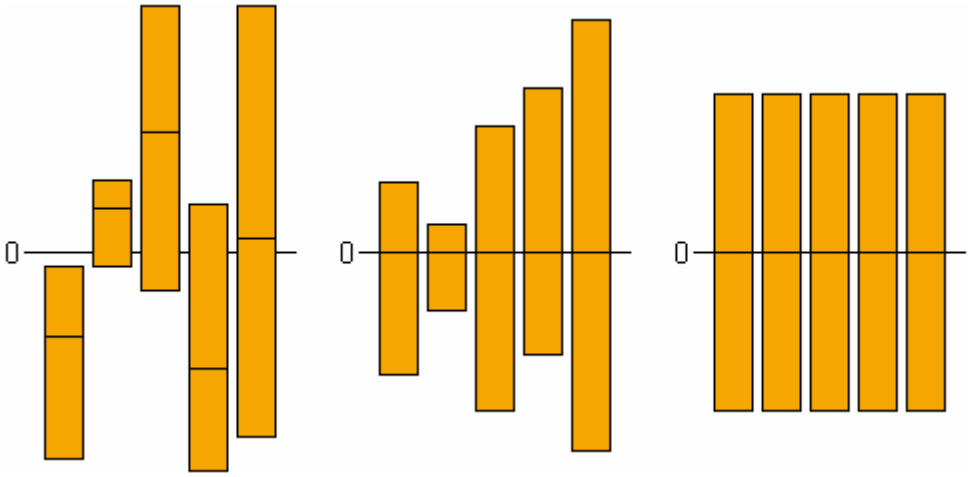
Figur 4.3 En n-dimensionell variabelrymd. För klarhets skull visas ett exempel med $n = 3$.

Raderna i datamatrisen, observationerna, representeras av en punktsvärm i variabelrymden. Innan observationerna projiceras på variabelrymden är det dock lämpligt att förbehandla data (Eriksson et al., 2006:40). Här kommer enhetsskalning och medelcentrering att beaktas. Se avsnitt 9.1 för en utförligare genomgång av dataförbehandling.

4.4.1 Enhetsskalning och medelcentrering av data

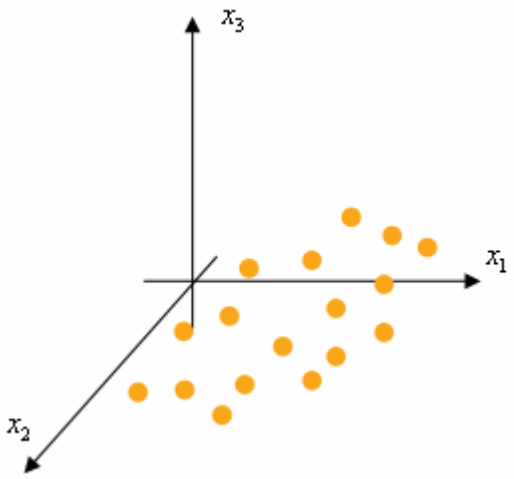
Skalning och medelcentrering av data är standard vid användande av PCA (Eriksson et al., 2006:40 samt Rosén, 1998:40). Inom examensarbetet enhetsskalades och medelcentrerades all data före samtliga analyser. Enhetsskalning används för att få en rättvis jämförelse mellan variabler med olika stor spridning. Eftersom syftet med PCA är att erhålla komponenter som beskriver så stor del av variationen i data som möjligt kommer variabler med stor varians annars att förklaras av PCA-modellen i större utsträckning än en variabel med liten varians. Enhetsskalning innebär att samtliga variabler, genom att dividera samtliga variabelvärden med standardavvikelsen för respektive variabel, skalas på så sätt att enhetsvarians erhålls.

Medelcentreringen utförs genom att subtrahera samtliga variabelvärden med respektive variabels medelvärde för att öka tydligheten hos modellen (Eriksson et al., 2006:44). Medelvärdet hos variablerna blir då lika med noll. Figur 4.4 visar en grafisk tolkning av medelcentrering och enhetsskalning.

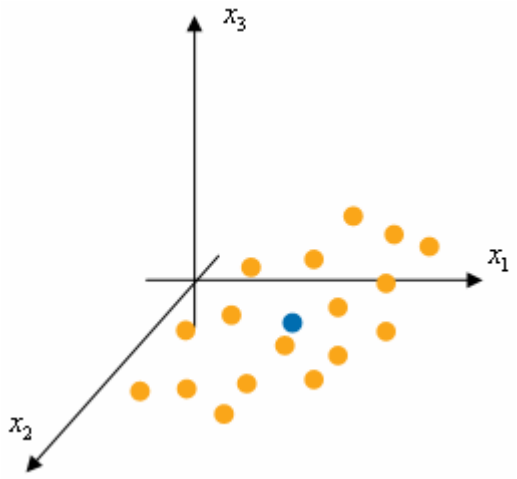


Figur 4.4 *Effekten av medelcentrering och enhetsskalning. Staplarna representerar variabler där medelvärdet visas som ett horisontellt streck. Medelcentrering gör att variablernas medelvärde blir lika med noll. Innan enhetsskalning har variablerna skilda varianser. Efter enhetsskalning är varianserna identiska.*

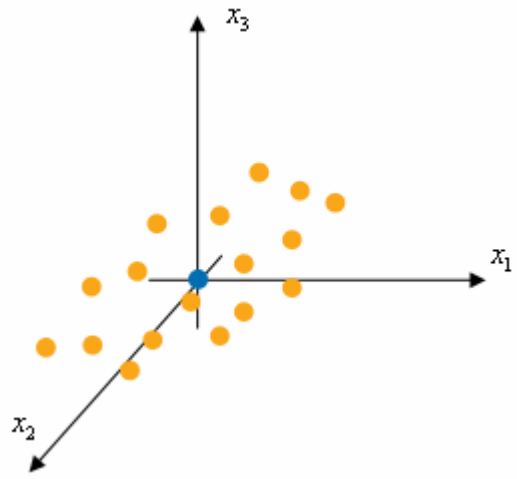
Figurerna 4.5-4.7 visar ett antal observationer som en punktsvärm i variabelrymden och hur data medelcentreras för att öka tydligheten hos modellen. Nästa steg av PCA:n är sedan att beräkna PC:erna.



Figur 4.5 *Raderna i datamatrisen, observationerna, representerar en punktsvärm i variabelrymden. För klarhets skull är respektive koordinataxel skalad till enhetsvarians.*



Figur 4.6 *Första steget i medelcentreringen är att beräkna variablernas medelvärden. Medelvärdesvektorn som bildas kan tolkas som en punkt i variabelrymden. Punkten hamnar i punktsvärmens mitt. I figuren representeras medelvärdesvektorn av den blå punkten.*

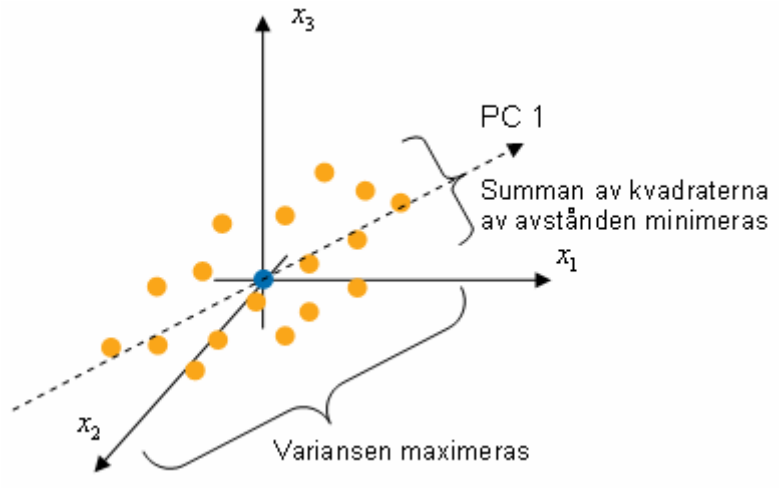


Figur 4.7 *Vid medelcentrering justeras variabelrymdens koordinatsystem så att punktsvärmens medelpunkt, här representerad av den blå punkten, sammanfaller med origo.*

4.4.2 Beräkning av principalkomponenterna

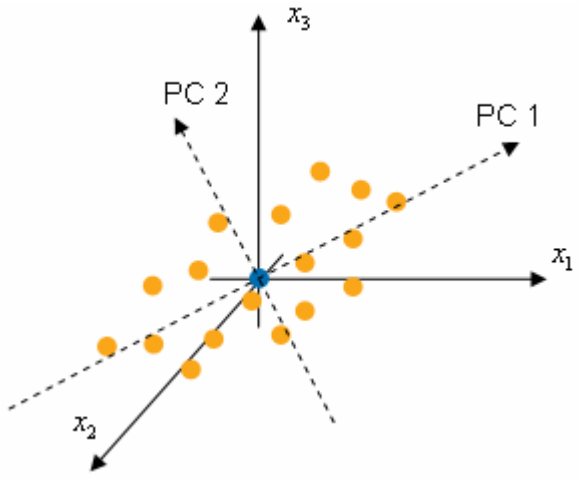
Den första PC:n, PC1, beräknas så att största möjliga variation i data förklaras. Figur 4.8 förklarar grafiskt hur PC1 bestäms utifrån befintlig data. Observationernas projektioner på PC:n bestämmer de så kallade score-värdena, vilket är observationernas koordinater längs PC:n. PC1 är linjen i variabelrymden som bäst approximerar data enligt minsta kvadrat-metoden[29] (MK-metoden) (Eriksson et al., 2006:46). Samtidigt, vilket går att utläsa ur figur 4.8, maximeras variansen hos score-värdena.

[29] Summan av kvadraterna av avstånden minimeras.



Figur 4.8 *PC1 beräknas så att formen av punktsvärmen på bästa sätt förklaras. Observationernas koordinater för PC:en fås genom att projicera observationerna ner på PC:en. En koordinat benämns score.*

En PC är i regel inte tillräckligt för att modellera den systematiska variationen i en samling multivariat data, då antalet variabler är stort. Ytterligare PC:er beräknas därför enligt samma princip som vid uträkningen av PC1. De andra PC:n, PC2, representeras, liksom PC1, av en linje i variabelrymden. PC2 dras ortogonalt mot PC1 och i den riktning som innebär att största möjliga variation hos data förklaras. I figur 4.9 visas hur två PC:er spänner upp ett plan. Resterande PC:er beräknas på samma sätt och dras ortogonalt mot tidigare dragna PC:er. För en datamängd **X** med *n* variabler kan *n* PC:er beräknas, men förhoppningen är att den mesta variationen i **X** kan förklaras av *a* PC:er, där $a << n$ (Jolliffe, 1986:2). När antalet PC:er sammanfaller med antalet variabler, $a = n$, förklaras naturligtvis 100 % av variationen i data.

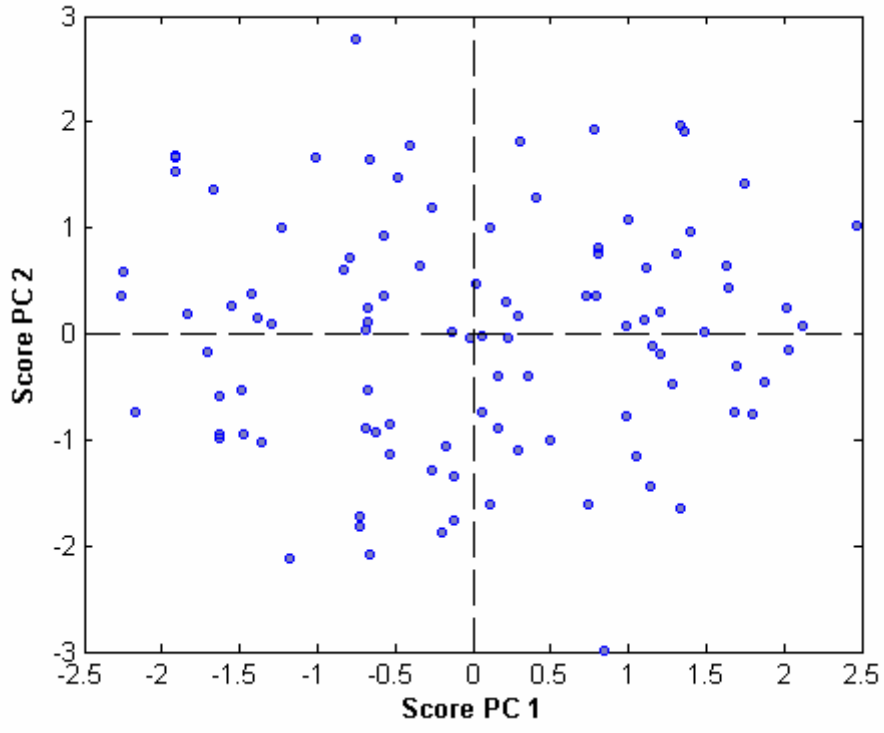


Figur 4.9 *Enligt samma metod som för PC1 dras resterande PC:er, ortogonalt mot tidigare dragna PC:er. PC2 förklarar därmed den näst största variationen i data och så vidare. Två PC:er spänner upp ett plan.*

4.4.3 Tolkning av score-plot och loading-plot

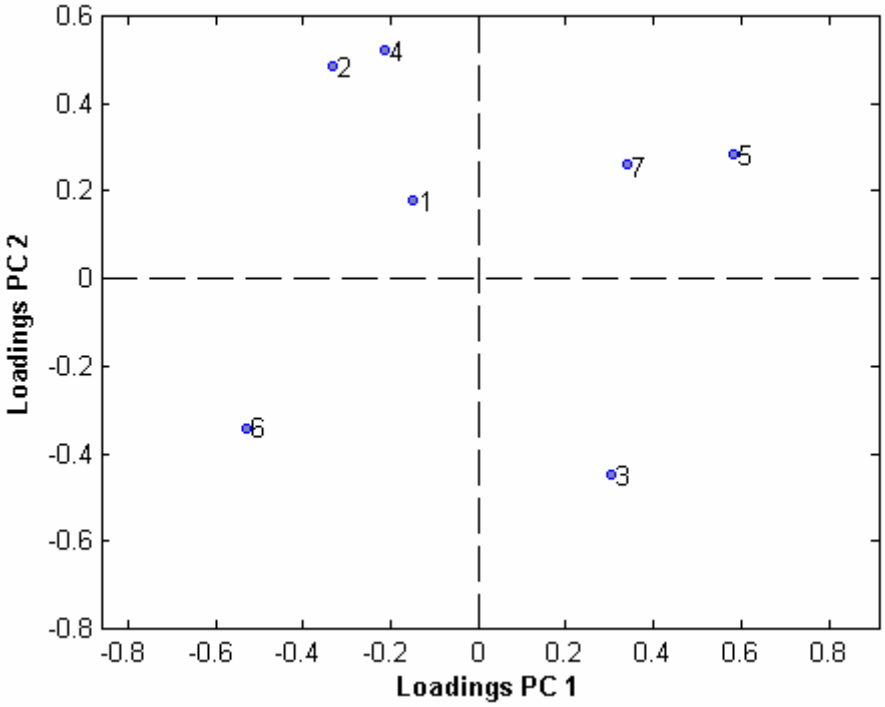
I enlighet med vad som nämnts tidigare kan varje punkt i variabelrymden sägas representera ett PT. Av klarhetsskäl, eftersom det är svårt att överblicka rymder i allmänhet och multidimensionella dito i synnerhet, presenterar samtliga figurer i examensarbetet en tvådimensionell bild av variabelrymden; nämligen planet som spänns upp av de två första PC:erna; PC1 och PC2. Det ska trots allt poängteras att samtliga beräkningar är gjorda i fler än de två dimensioner som presenteras i figurerna. Med andra ord är hänsyn tagen till fler än två PC:er och konstruerade modeller förklarar mer variation än vad de tvådimensionella figurerna ger sken av. Tolkningar av figurerna bör därför inte överskattas eftersom det endast är en förenklad bild av PT:en som presenteras, men figurerna kan ändå visa på intressanta skiftningar mellan olika PT, ty PC1 och PC2 är de PC:er som beskriver den största variationen i data. Vidare kommer avsnitt 5.1.4 att visa hur kompletterande figurer kan hjälpa till att öka analysmöjligheterna.

Genom att projicera samtliga observationer på det tvådimensionella PC-planet fås ett fönster mot den multivariata variabelrymden. Punkterna i planet kallas, som tidigare poängterats, för scores och figuren benämns därför score-plot (Rosén, 1998:84). Figur 4.10 visar en score-plot med data från det aktuella vattenkraftverket. Varje punkt representerar ett PT och närliggande punkter indikerar liknande PT. Enligt samma logik anger två kraftigt separerade punkter skilda PT. Processens normaltillstånd är kring origo (Ibid.).



Figur 4.10 Score-plot. Punkterna representerar PT. Närliggande punkter indikerar liknande PT och tydligt separerade punkter antyder skilda PT.

Genom att istället projicera observationerna på PC:ernas loading-vektorer ges en övergripande bild över relationerna mellan samtliga *n* variabler. Figur 4.11 visar ett exempel på en så kallad loading-plot och punkterna representerar här de olika variablerna. Närliggande punkter indikerar variabler som påverkar processen på liknande vis (Rosén, 1998:89). Genom att studera score- och loading-plottarna samtidigt kan kunskap om hur variabelvärdesförändringar driver processen åt olika håll i PC-rymden erhållas. I en process där flera av variablerna är kontrollerade, det vill säga styrs av en operatör, kan den här insikten i variablernas processpåverkan hjälpa operatören att styra tillbaka processen till den normala regionen kring origo om PT-förändringar skulle uppstå (Ibid.).



Figur 4.11 *Loading-plot. Punkterna representerar variabler. Närliggande punkter indikerar variabler som påverkar processen på liknande sätt. En ökning av exempelvis variabel 3 skulle få PT:en att driva snett nedåt höger i en score-plot.*

5 Applicering på problemet

Matlab är det program som använts för att utföra de, för PCA:erna, nödvändiga beräkningarna samt för att exekvera demonstrationsapplikationen[30]. Matlab är en förkortning av Matrix Laboratory (The Mathworks, 2006a:1.2) och är ett program utvecklat av The Mathworks som är särskilt utformat för matrisberäkningar. Till grundversionen av Matlab används även Statistics Toolbox, även den utvecklad av The Mathworks, för att förenkla genomförandet av beräkningarna bakom PCA:erna. Utöver egenhändigt komponerad kod används programmet PLS_Toolbox 4.0 från Eigenvector Research. PLS_Toolbox 4.0 är en sammansättning avancerade multivariata verktyg som används i kombination med Matlab (Wise et al., 2006:1). I enlighet med vad som nämnts tidigare har processdata hämtats från TK-databasen med hjälp av programmet ExportTKData. Data hämtades på matrisform och importerades enkelt till Matlab via det Windowsbaserade programmet Excel.

5.1 Arbetets upplägg

Med hjälp av PCA kommer en modell som underlättar driftövervakning av vattenkraftverk att utformas. Vid driftövervakning är det av intresse att detektera situationer då driften avviker från vad som anses normalt. Idag används larmgränser på respektive variabel för detektion av icke normal driftsituation. Larmgränserna är dock statiska på så vis att de endast ger ett mått på inom vilka gränser respektive variabelvärde tillåts variera. Det är emellertid möjligt att det, trots att samtliga variabelvärden hålls inom sina respektive gränser, råder ett drifttillstånd som inte kan anses vara normalt. Eftersom PCA tar hänsyn till samtliga variabler samtidigt skulle därför en väl kalibrerad PCA-modell kunna förbättra möjligheterna till en effektiv driftövervakning.

5.1.1 Normala vs. icke normala processtillstånd (PT)

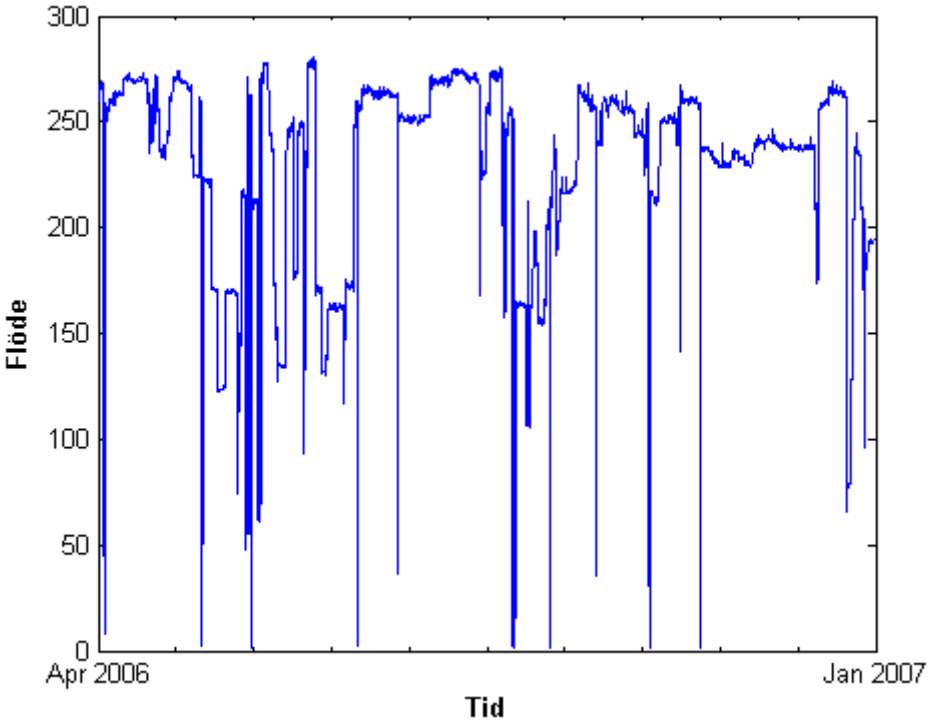
Eftersom PCA-modellen ska kunna detektera avvikande PT är det av yttersta vikt att modellen baseras på data från drift vid normala PT. Dock är det inte nödvändigtvis enkelt att definiera vad som kan anses vara normal drift. Inom ramen för examensarbetet har emellertid data från tidpunkter då inga larm getts betraktats som normala. I Conwide System III finns tillgång till driftjournaler där registrerade larm är listade. Data från tidpunkter då larm uppstått har undvikits, för att inte basera PCA-modellen på icke normala PT.

Likt tidigare påpekats kan dock PT vara icke normala trots att ingen larmgräns har passerats. Den distinktion mellan normala och icke normala PT som använts i examensarbetet är därför inte optimal eftersom PCA-modellen delvis kan vara baserad på PT som egentligen borde ha undvikits vid modellkonstruktionen. Emellertid existerar ingen fullständigt säker metod för att urskilja icke normala PT från normala dito. Den enda urskiljningsmetod som finns implementerad är just larmgränser för respektive variabelvärde. Larmgränserna får därför agera riktlinje vid kategoriseringen av normala och icke normala PT, trots ovan nämnda problematik. Tanken är att PCA-modelleringen ska underlätta urskiljningen av icke normala PT.

[30] Se kapitel 8.

5.1.2 Två fall – analys av driftsituationer vid olika flödesnivåer

Driften av det aktuella vattenkraftverket styrs av DC i Storuman. Med styrs menas att operatörer på DC kan reglera flödet av vatten in till de två aggregaten. Flödet är en av de variabler som registreras i TK-databasen[31] och det är således möjligt att se hur flödet har varierat över tid. Figur 5.1 visar flödet till en av de två generatorerna vid den aktuella kraftverksanläggning över en längre period.



Figur 5.1 Flödet över tid, från 2006-04-29 till 2007-01-29.

I figur 5.1 tydliggörs att flödet kan variera kraftigt från dag till dag. När flödet till generatorerna ökas eller minskas kommer även övriga variabler att påverkas. Det är därför lätt att inse att flödet i hög grad påverkar aktuellt PT. Lika lätt inses att PCA-modellen, vid modellering av alla typer av flödesnivåer, inte blir lika känslig för varierande variabelvärden. Alternativet till en PCA-modell som klarar av att ta hänsyn till flödesnivån är en PCA-modell som är tränad på data från en period med förhållandevis konstant flöde. PCA-modellen förväntas då bli känsligare för variabel-värdesavvikelser, men kommer å andra sidan inte att klara av olika typer av drift i form av olika flödesnivåer.

Därmed har två olika fall utkristalliserats. Det optimala vore om PCA-modellen klarar av att göra åtskillnad mellan normal och icke normal drift oavsett aktuellt flöde. Först kommer därför en PCA-modell baserad på träningsdata från perioder med skiftande flöde att beräknas. Därefter kommer en PCA-modell baserad på processdata från en period med mer stabil flödesnivå att beräknas. Jämförelser mellan de två tillvägagångs-

[31] Se bilaga 1 för en komplett förteckning över registrerade variabler.

sätten kommer då att möjliggöras varpå slutsatser om användbarheten hos PCA för vattenkraftanläggningar medges.

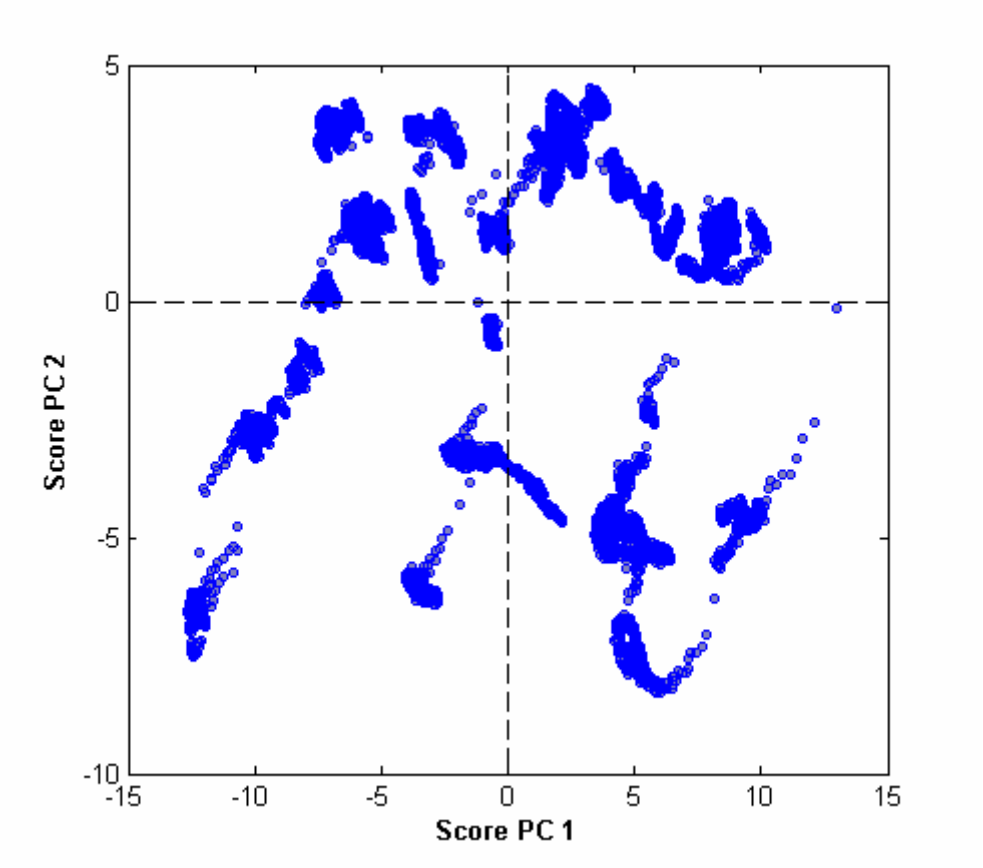
Efter att ha konsulterat driftjournalen i Conwide System III kunde lämpliga tidsperioder för uppbyggnad av PCA-modellerna väljas. Tidpunkter vid registrerade larm undveks samtidigt som flödeskurvan gav information om perioder med varierad respektive stabil drift. Fortsättningsvis kommer processdata, som använts till att konstruera en PCA-modell, att benämnas träningsdata. Data som sedan projiceras på den anpassade modellen kallas hädanefter för testdata. För en fullständig förteckning över tränings-respektive testdata hänvisas till bilaga 1 och bilaga 2.

5.1.3 Inkluderade variabler

När tidsperioder för träningsdata hade fastslagits kunde processdata hämtas från TK-databasen. En fråga som återstod var dock vilka variabler som skulle inkluderas i PCA-modellen. En av styrkorna hos PCA är den dimensionsreducerande effekten och det kan därför vara meningsfullt att, åtminstone initialt, inkludera så mycket data som möjligt i PCA-modellen (Carlsson, 2006-10-16). Endast de mest betydande variationerna kommer att modelleras när PC:erna är beräknade. Däremot kan det av andra skäl visa sig onödigt, eller rent av olyckligt, att inkludera somliga variabler i PCA-modellen. För förståelse kring hur det aktuella vattenkraftverket är konstruerat och hur olika variabel-värden därför påverkar varandra konsulterades underhållsingenjör Sören Ekström. I samråd med honom uteslöts några av variablerna från modelleringen. I bilaga 1 har uteslutna variabler markerats. Anledningen till uteslutningen kan exempelvis vara att givaren helt enkelt inte är inkopplad mot TK-databasen eller att givaren är säsongsbetingad och speciellt konstruerad för exempelvis negativa temperaturer som endast uppstår under årets kallaste månader.

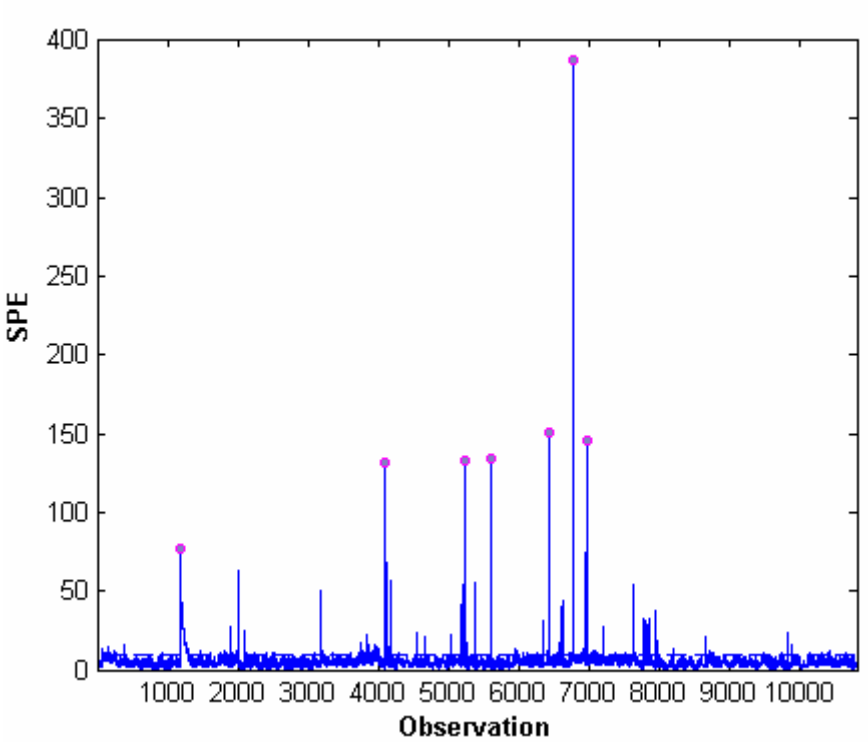
5.1.4 Outlier-detektion

När träningsdata är vald kan en första PCA utföras. Figur 5.2 visar en första score-plot vid modellering av träningsdata från fall 1 med varierande flödesnivå. I figuren går att se PT som ligger relativt långt från origo och övriga PT vilket innebär att de skiljer sig kraftigt från de andra observationerna.

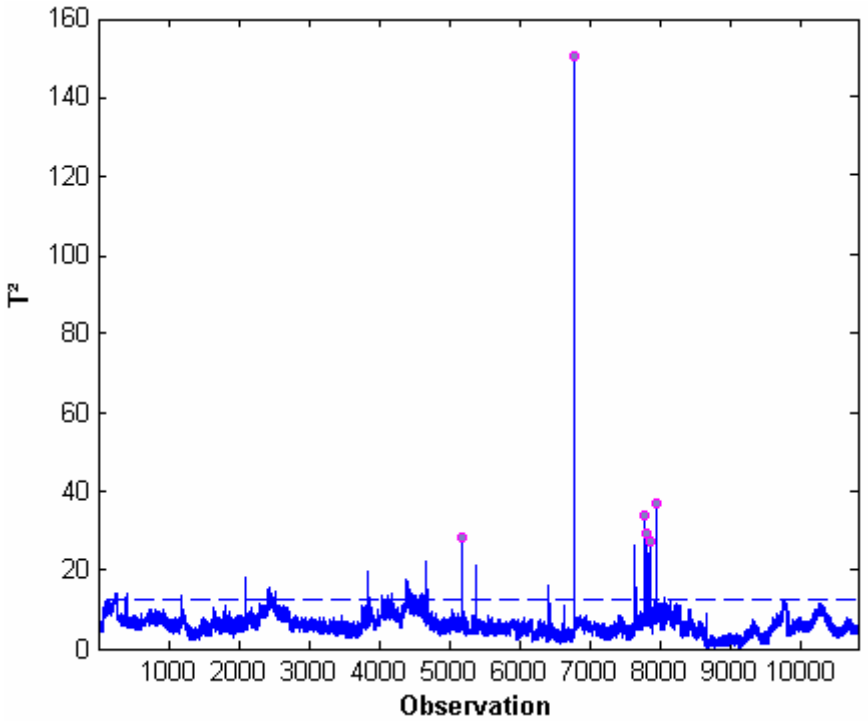


Figur 5.2 Score-plot för träningsdata från fall 1.

Genom att studera någon av de två tidigare nämnda residualplottarna tydliggörs vissa observationers avvikelse från den konstruerade modellen. Vid modellkonstruktionen är det viktigt att endast observationer som kan anses vara representativa behandlas (Rosén, 1998:81). Därför bör avvikande observationer studeras mer ingående. I figur 5.3 och figur 5.4 indikeras att några av observationerna avviker starkt från modellen.



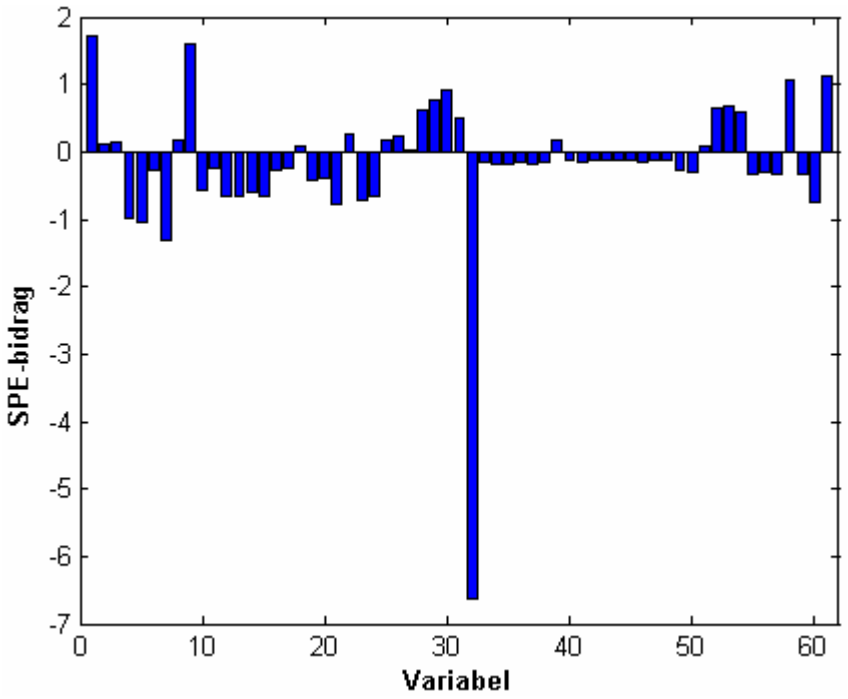
Figur 5.3 SPE-plot. Misstänkta outliers är markerade.



Figur 5.4 T²-plot. Misstänkta outliers är markerade.

Det är visserligen värt att notera att aktsamhet bör iakttas vid tolkning av figurerna ovan. Då antalet observationer är stort kommer somliga representativa observationer att avvika mer från modellen än vad gällande konfidensgräns anger (Rosén, 1998:80). Det bör därför påpekas att en till synes avvikande observation inte nödvändigtvis är en så kallad outlier. Varje misstänkt observation måste därför utredas individuellt.

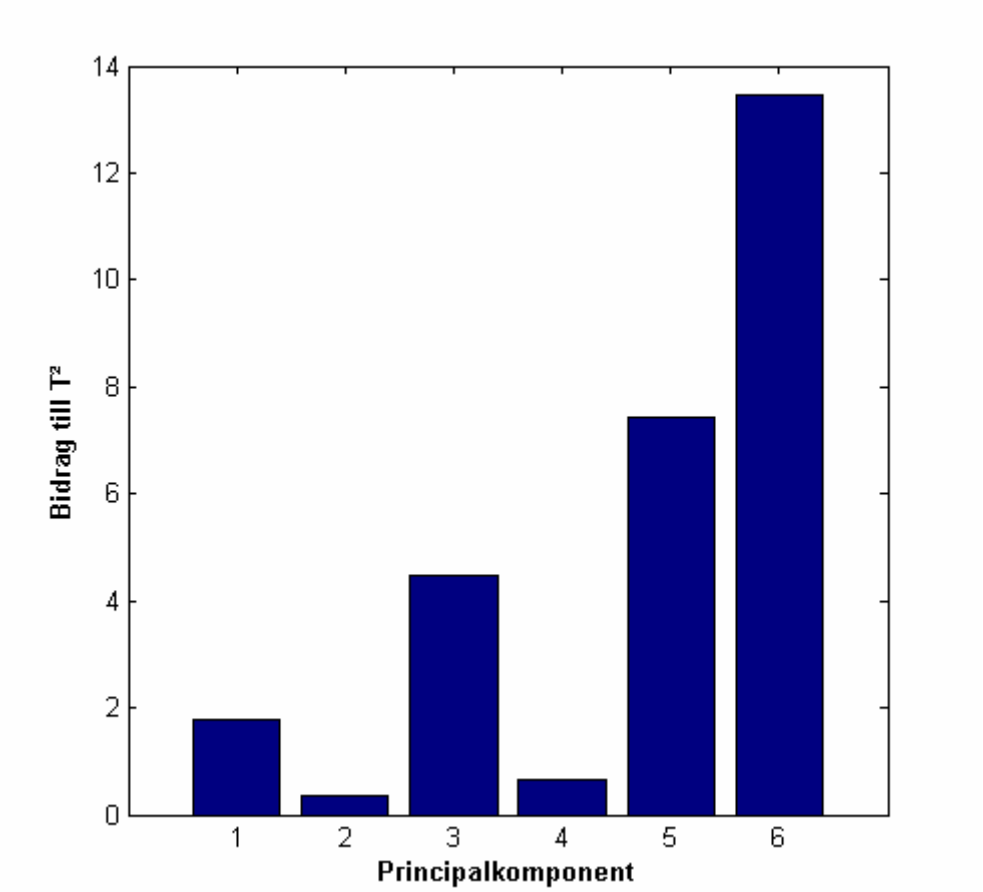
Det behöver inte nödvändigtvis vara enkelt att avgöra huruvida en observations avvikelse från modellen ska betraktas som naturlig eller ej. Somliga observationer avviker naturligtvis mer än andra men avvikelser kan bero antingen på naturliga eller på icke naturliga variationer i variabelvärden. Naturliga variationer ska definitivt inte uteslutas utan måste inkluderas i modelleringen om PCA-modellen ska bli representativ. Samtidigt måste, enligt samma princip, icke naturliga variationer uteslutas ur modelleringen. Icke naturliga variationer kan exempelvis bero på givarfel, där felaktiga variabelvärden registrerats i TK-databasen. En metod för att avgöra om starkt avvikande observationer ska inkluderas eller ej är, som tidigare diskuterats i kapitel 4, att studera respektive variabelvärdes bidrag till observationens avvikelse från PCA-modellen. En observations höga värde på SPE eller T^2 kan därmed avslöja den variabel eller de variabler som avviker vid en viss tidpunkt. För höga SPE-värden ges variablernas residualbidrag enkelt av en rad i **E** (se ekvation 8). I figur 5.5 visas variablernas residualbidrag för observation 1173 som ses markerad med en rosa ring till vänster i figur 5.3.



Figur 5.5 Respektive variabels bidrag till SPE vid observation 1173 från figur 5.3.

Figur 5.5 antyder att det framförallt är variabel 32 (Spärrvattenfilter Difftryck) som bidrar till det stora residualfelet med ett ovanligt lågt värde.

Undersökning av variablers bidrag till T^2 måste däremot ske i två steg. Eftersom värdet för T^2 anger en observations avstånd till origo inom det modellerade hyperplanet krävs först kunskap om vilken PC som bidrar mest till det höga T^2-värdet. Därefter är det möjligt att beräkna variablernas bidrag till avvikelsen längs den PC:en. Med hjälp av ekvation 14 kan PC:ernas bidrag till T^2 beräknas. Figur 5.6 visar PC:ernas bidrag till T^2 för observation 5185, som kan ses markerad i figur 5.4.



Figur 5.6 PC:ernas bidrag till T^2 vid observation 5185 från figur 5.4.

Nästa steg är att beräkna hur mycket respektive variabel bidrar till avvikelser längs de PC:er vars bidrag till T^2 är störst. Ekvation 15 avslöjar i vilken utsträckning respektive variabel bidrar till avvikelserna längs en viss PC. I en figur liknande figur 5.5 kan respektive variabels bidrag till avvikelser längs PC6 och PC5 därmed visas.

Genom att undersöka värden för de variabler som bidrar mest till residualfelet för en observation som avviker starkt från PCA-modellen är det möjligt att avgöra huruvida observationen bör betraktas som normal eller icke normal (Rosén, 1998:81). Det kan vara ett tidskrävande arbete eftersom det är ett övervägande som måste göras för varje enskild observation om den ska inkluderas i modellen eller ej. För exemplet angivet ovan måste exempelvis värdet på variabel 32 för observation 1173, där värdet på SPE vida översteg resterande värden, granskas mer noggrant för att det ska gå att avgöra huruvida observationen kan anses vara normal eller ej. Utifrån vilken typ av variabel det rör sig om kan naturligtvis värden variera naturligt på olika vis. Exempelvis bör inte en

temperatur variera alltför plötsligt för att sedan återgå till samma nivå igen sekunden senare. Variationen kan då inte sägas vara kännetecknande för normal driftsituation och skulle därför uteslutas ur PCA-modelleringen. För att få en rättvisande modell måste en liknande analys göras för samtliga observationer som avviker starkt från den initiala PCA-modellen.

5.1.4.1 Hantering av outliers

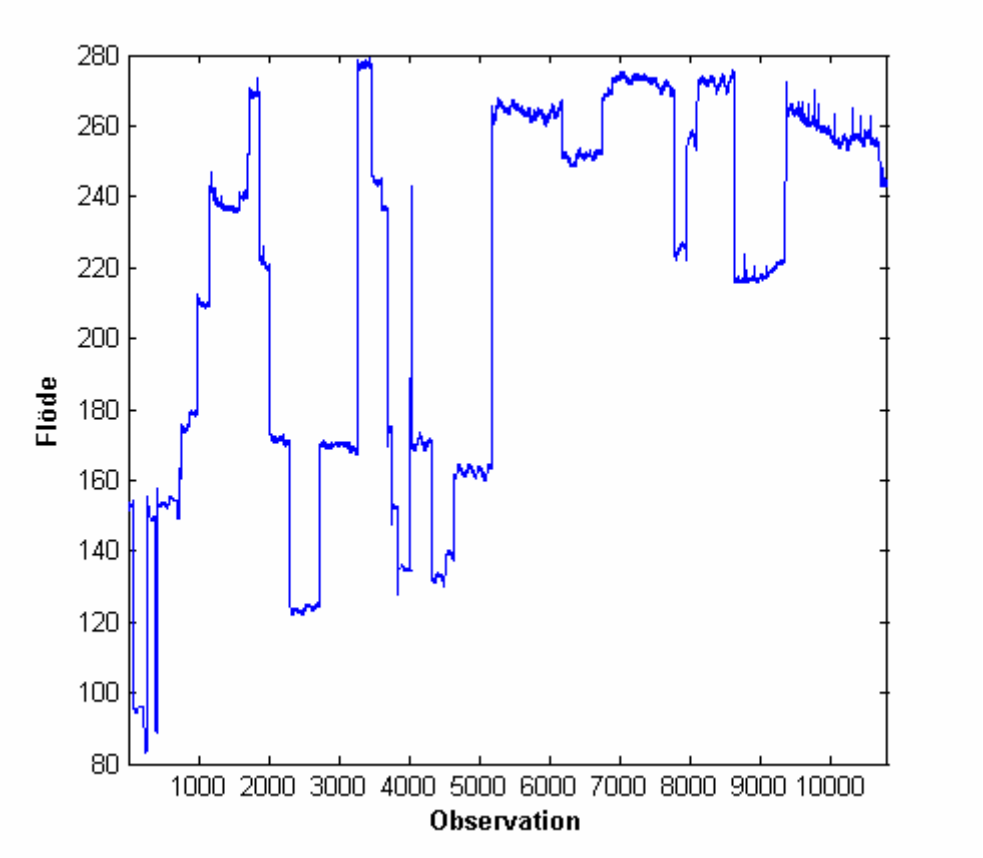
Hur outliers ska hanteras beror på typen av data som behandlas. Vid tidsserie-data kan observationer egentligen inte uteslutas utan vidare eftersom det då skulle uppstå en tids-lucka i serien. Inom examensarbetet utesluts dock outliers ur modelleringen. En förklaring till varför outliers kan uteslutas trots att det är tidsserie-data som ska modelleras är att processdata, inom examensarbetet, inte behandlas som tidsserie-data. Mer om tidsserie-data och hur den typen av data påverkar PCA:n behandlas i kapitel 9.

5.1.5 Slutgiltig PCA

Efter att data hade rensats från outliers, enligt tillvägagångssättet beskrivet i avsnitt 5.1.4, kunde den slutgiltiga PCA-modellen konstrueras. Genom att studera en scree-plot över variansen som förklaras av varje PC kunde lämpligt antal PC:er som ska inkluderas i modellen väljas. Ytterligare anvisning om hur många PC:er som bör inkluderas kan ges av en RMSECV-plot. När PCA-modellen väl är upprättad kan nya observationer projiceras på PC:erna varvid driftsituationen kan analyseras. Testdata bör bestå av observationer från en tidsperiod som är intressant ur övervakningssynpunkt. I realtidsfallet består testdata typiskt av den aktuella driftsituationen. Inom examensarbetet har dock ingen realtidsöverföring av data, från TK-databasen till någon PCA-implementation, konstruerats. Därför har testdata istället bestått av tidigare registrerade observationer från TK-databasen. Testdata måste bestå av lika många variabler som träningsdata. Dock ska ingen förbehandling, undantaget medelcentrering och enhetsskalning, av testdata utföras eftersom modellen är anpassad till normal driftsituation. De nya observationernas scores i den konstruerade PCA-modellen beräknas enligt tidigare angiven ekvation 7.

6 Fall 1 – driftsituation vid växlande flödesnivå

Först kommer det första av de två tidigare diskuterade fallen att betraktas. I det första fallet baserades PCA-modellen på träningsdata från perioder med växlande flöde. Tanken var att modellen på ett effektivt vis skulle klara av att göra en distinktion mellan normala PT och avvikande PT, oberoende av det aktuella flödet. För att välja lämpligt träningsdata studerades först driftjournalen i Conwide System III för att driftperioder vid larmregistreringar skulle kunna undvikas. Tillsammans med en graf över tidigare flöde kunde därmed träningsdata från perioder med skiftande flöde väljas. Figur 6.1 visar flödet över tiden för träningsdata och förklarar därmed skiftningarna i flödesnivån.

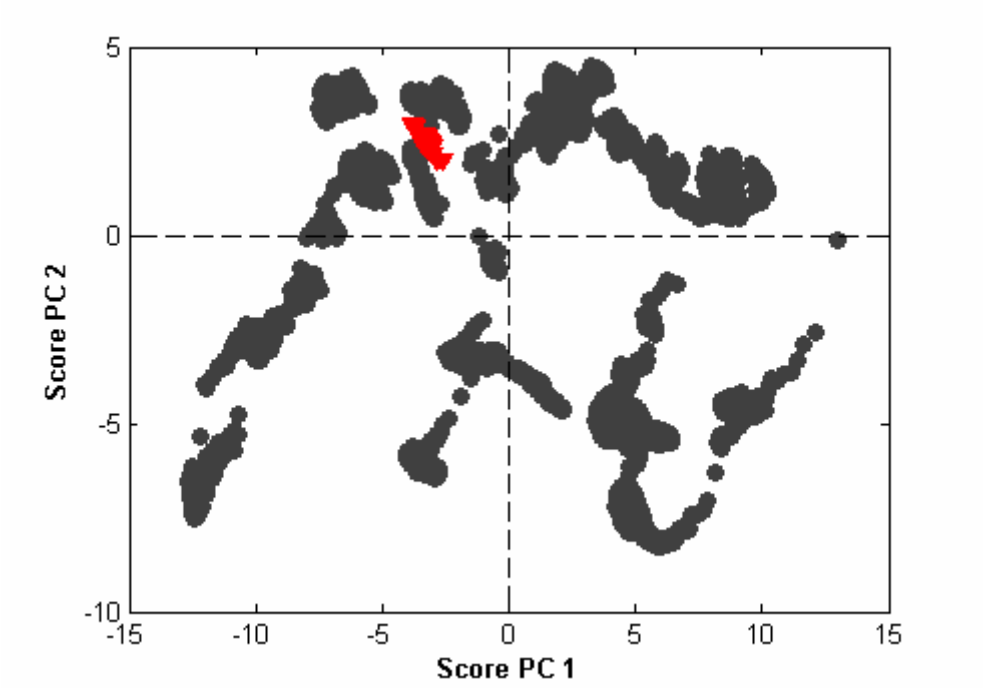


Figur 6.1 Flödesnivån över tiden för fall 1.

En sammanställning av vilka variabler som inkluderats i modellen ges i bilaga 1 och bilaga 2 utgör en förteckning över de perioder från vilka tränings- respektive testdata hämtats för de två fallen.

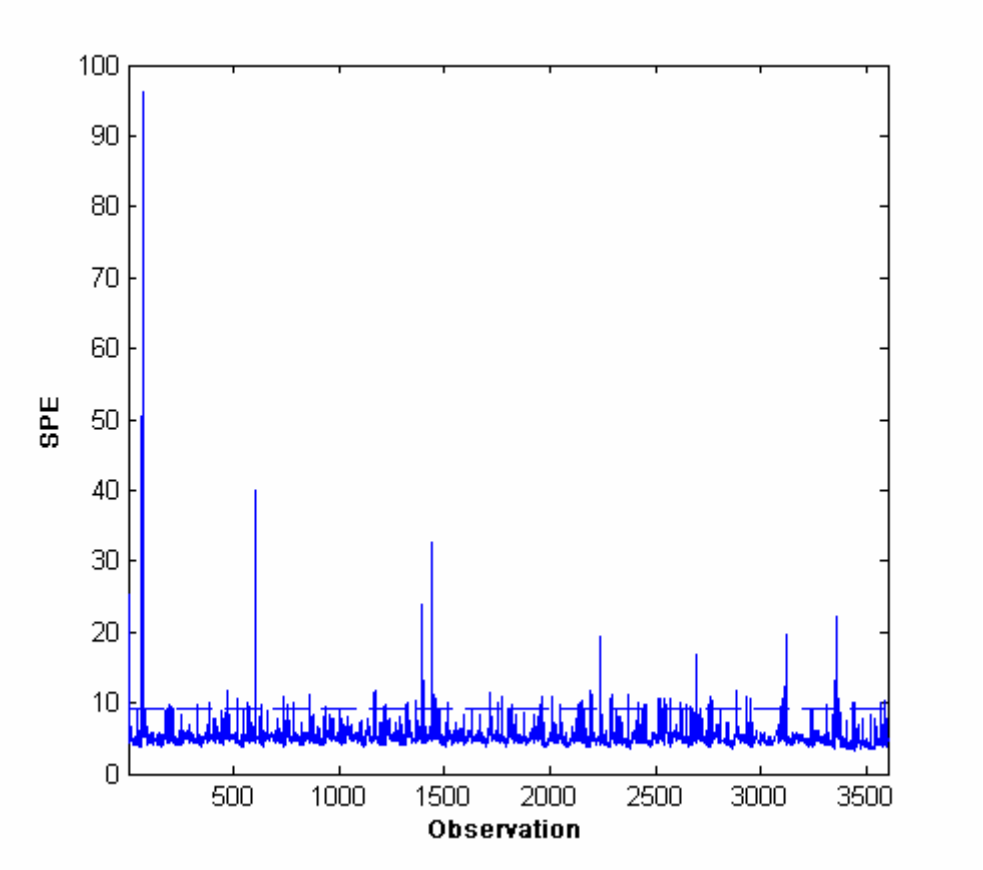
Träningsdata för fall 1 bestod initialt av 10816 observationer av 61 variabler. 12 outliers noterades och raderades. Bestämning av antalet PC:er att inkludera i modelleringen gjordes utifrån scree-plot samt korsvalidering vilka kan ses i figur 4.1 och i figur 4.2. PCA-modellen baserades på de 6 första PC:erna eftersom det enligt figurerna inte fanns mycket att vinna på att inkludera fler PC:er.

Figur 5.2 och figur 5.3 visar plottar över den, av träningsdata, konstruerade PCA-modellen respektive observationernas avstånd till modellen. Figur 6.2 visar testdata projicerad på PCA-modellens första två PC:er.

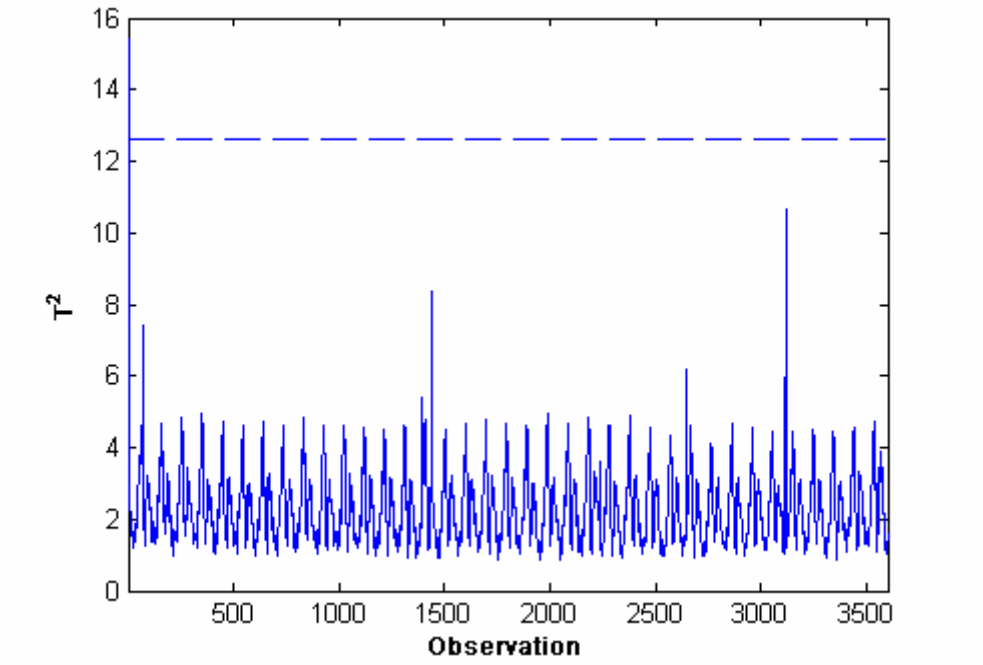


Figur 6.2 *Testdata från fall 1 projicerad på PCA-modellen. Observationer från testdata syns som röda trianglar i figuren. De mörka prickarna är observationer från träningsdata.*

Vid en jämförelse av figur 5.2 och figur 6.2 blir det tydligt att PCA-modellen verkar förklara testdata väl. Observationerna från testdata befinner sig nämligen förhållandevis nära origo. För ytterligare information om modellpassningen kan SPE- respektive T^2-plottar studeras. Figurerna 6.3 och 6.4 visar de två residualvektorerna för testdata grafiskt.



Figur 6.3 SPE-residualer för testdata från fall 1.



Figur 6.4 T^2-residualer för testdata från fall 1.

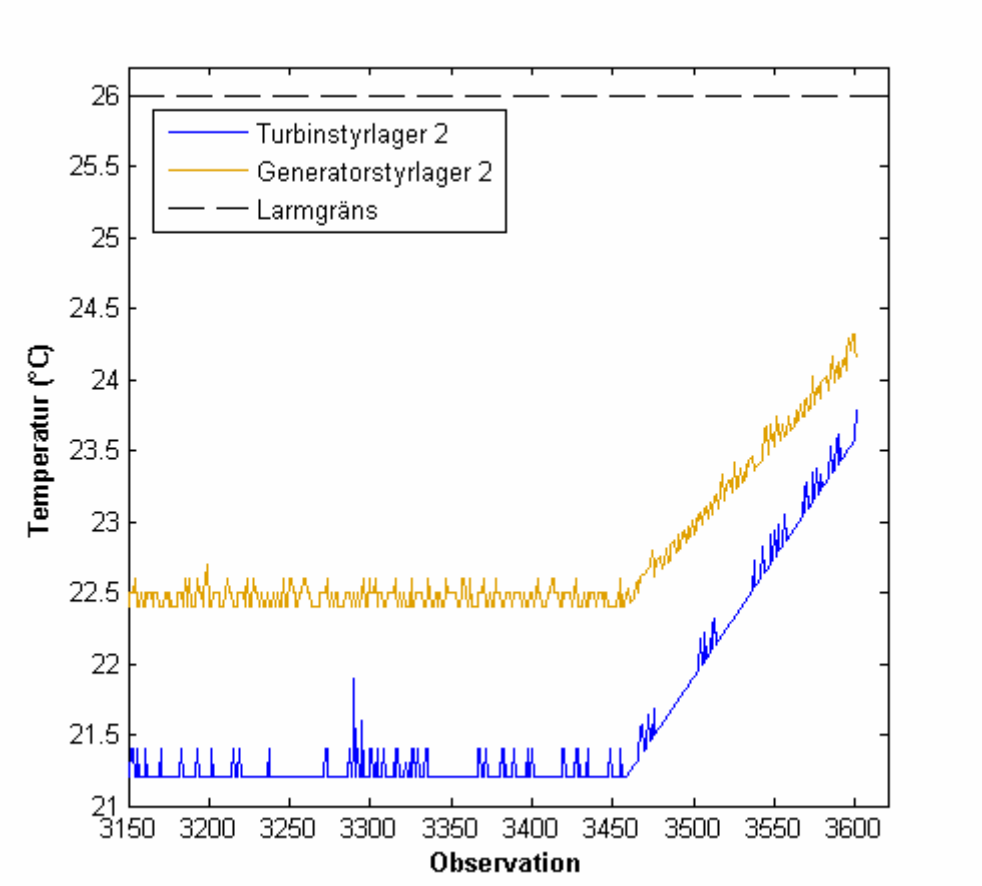
Ur figurerna 6.3 och 6.4 kan slutsatsen om att testdata passar PCA-modellen väl dras. Endast ett par enstaka spikar över konfidensgränsen för SPE kan urskiljas. Över tid

kommer residualvärden över konfidensgränserna att förekomma utan att det behöver innebära att PT:et nödvändigtvis är avvikande på ett alarmerande vis (Rosén, 1998:80). Eftersom spikarna direkt följs av värden klart under konfidensgränsen finns därför ingen egentlig anledning till att anta avvikande PT. Snarare handlar det då om naturliga avvikelser eller registreringsfel. Vad som däremot inte är önskvärt är om trender över ökande residualvärden kan skönjas. Sammanfattningsvis kan sägas att testdata sammanfaller väl med PCA-modellen vilket indikerar att driften under testperioden kan liknas vid driften under träningsperioden. Eftersom träningsdata baserades på perioder med normal driftsituation kan därför även driftsituationen under testperioden antas vara normal. Så långt tycks PCA-modellen fungera planenligt. Att driftsituationen under testperioden klassificerades som normal stämmer nämligen överens med vad som kunde väntas eftersom testdata valts från en period under vilken driftsituationen antagits vara normal.

För att modellen ska vara användbar i praktiken räcker det dock inte att normala PT klassificeras som normala. Modellen måste naturligtvis även kunna detektera avvikande PT. För att kontrollera modellens känslighet för avvikande PT har ett funktionsfel i anläggningen simulerats. Eftersom funktionsfel leder till avvikande driftsituation är det av vikt att tidigt detektera avvikelsen som felet medför. Det är inte alltid säkert att statiska larmgränser är tillräckligt för att upptäcka små avvikelser om de sker hos många variabler samtidigt. Tanken är att en PCA ska klara av att upptäcka även små förändringar som inte kan förklaras av den modell som baserats på processdata från så kallad normal drift.

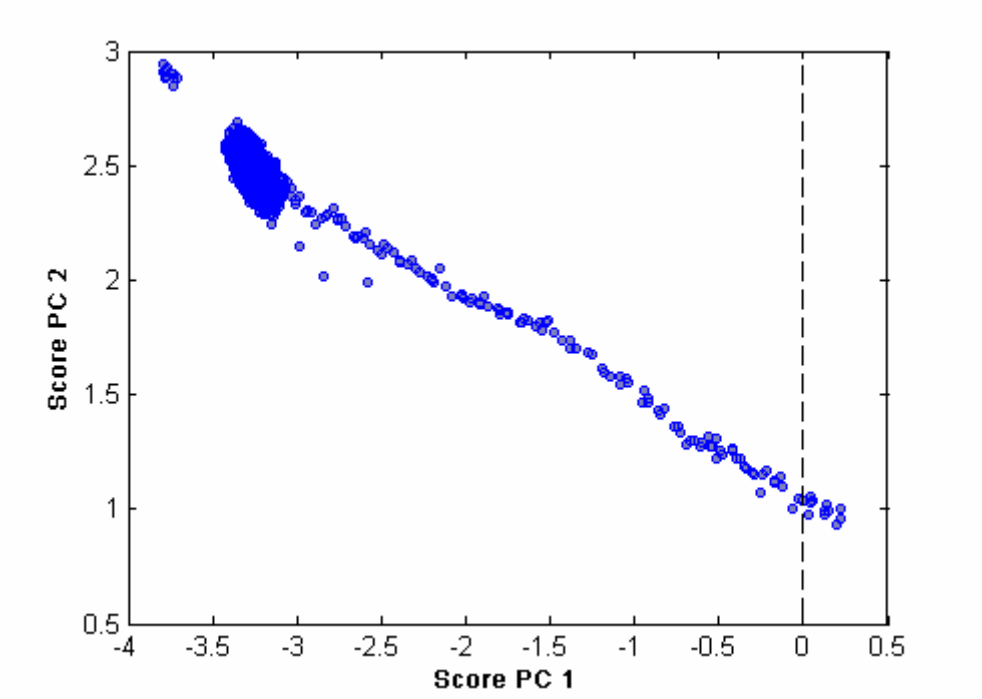
För att avgöra i vilken mån PCA-modellen klarade av att detektera avvikande beteende simulerades fel på kylvattenpumparna. Genom att förvanska testdata kunde en situation med kylvattenpumpfel efterliknas. Efter konsultation av Sören Ekström valdes variabler som påverkas av kylvattenpumpfel ut. För att bedöma hur väl PCA-modellen påvisar avvikelser i PT lades därefter trender in för de utvalda variablerna. Till en början var respektive avvikelse liten för de olika påverkade variablerna, men ökade sedan efter hand. På så vis var det möjligt att avgöra huruvida PCA-modellen kunde urskilja det avvikande PT:en och när de i så fall kunde skönjas. Noterbart är dock att ingen av variablernas larmgränser passerades, varken uppåt eller nedåt. Trenderna lades in ungefär två minuter[32] från slutet av testdata. I bilaga 1 ges en sammanställning av de variabler som Ekström anser påverkas av kylvattenpumpfel och som därmed förvanskats vid simuleringen. Utan undantag är det temperaturökningar till följd av sviktande kylning, i form av minskade värden hos variablerna Kylvattenledning flöde och Kylvattenledning tryck, som simulerats. Figur 6.5 visar ett exempel på inlagd trend för två av de utvalda variablerna, nämligen Generatorstyrlager 2 och Turbinstyrlager 2, samt de två variablernas gemensamma larmgräns.

[32] 120 observationer



Figur 6.5 *Manipulerad data. Grafen visar två av de variabler som manipulerats för att simulera kylvattenpumpfel. Trenderna inleds vid observation 3460 och fortlöper fram till den sista observationen. Noterbart är att ingen av de båda variablernas värde är i närheten av larmgränsen.*

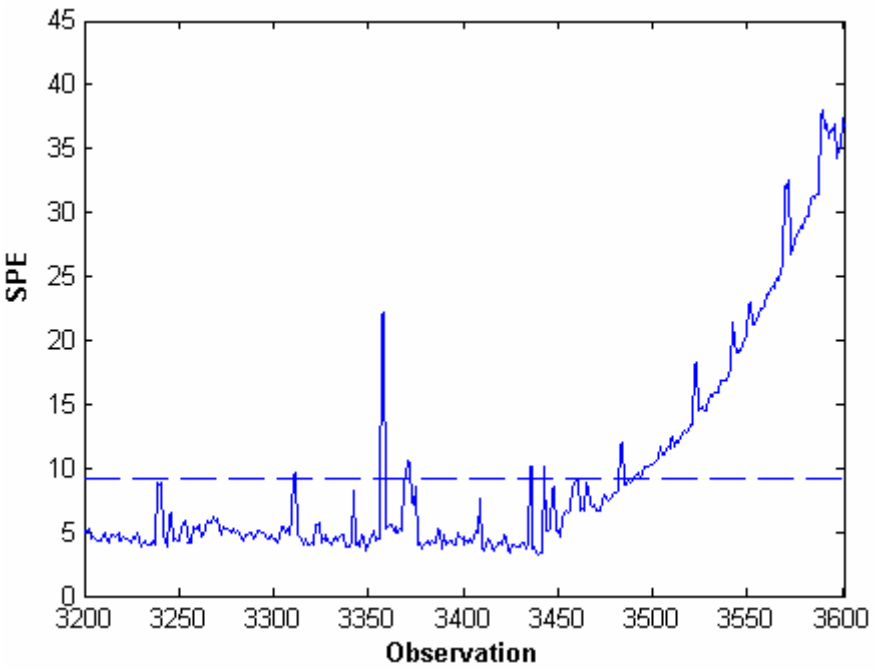
Vid projicering av förvanskad testdata för fall 1 ges score-plot enligt figur 6.6.



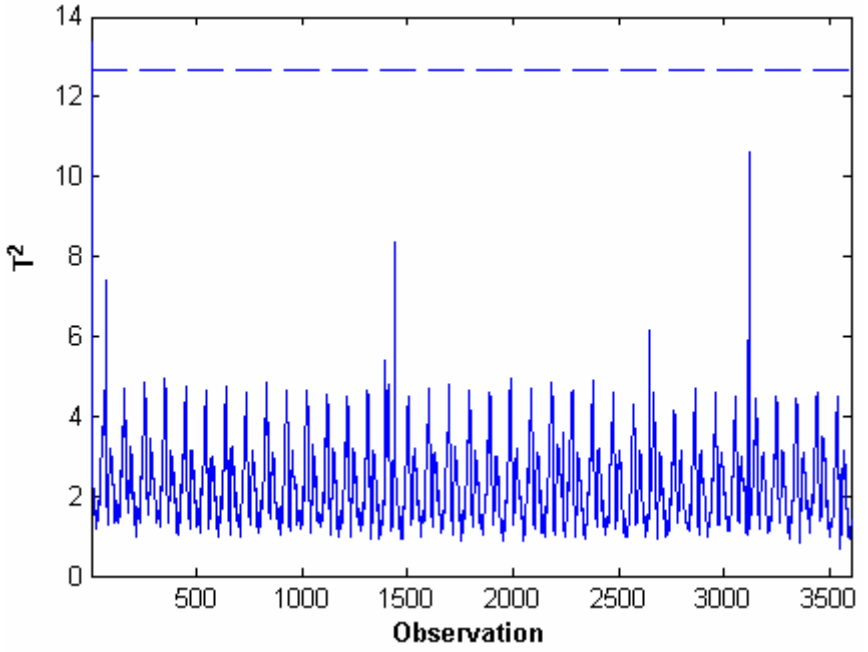
Figur 6.6 *Score-plot för testdata från fall 1 vid simulerat kylvattenpumpfel. Även här märks en tydlig trend. PT:en rör sig snett nedåt höger med tiden. Rörelsen indikerar att process-situationen håller på att ändras.*

Vid en jämförelse av figur 6.6 och figur 6.2 blir det tydligt att det simulerade felet har påverkat resultatet från PCA:n. En markant drift syns i score-plotten för testdata när felet simulerats. Att PT:en börjat driva indikerar att förändringar skett i processen. Visserligen bör inte några slutsatser dras från score-plotten allena, eftersom det naturligtvis skulle kunna röra sig om naturliga processförändringar. Information om passningen till PCA-modellen kan istället fås från figur 6.7 och figur 6.8 som visar residualvärdesvektorerna från PCA:n. Det simulerade felet inleds vid observation 3460 och blir snart därefter synligt i figur 6.7 på grund av ökande värden på SPE-residualerna. Efter en minut, strax efter observation 3500, är trenden påtaglig och misstankar om ändrad driftsituation skulle därmed kunna väckas på ett tidigt stadium. Återigen är det värt att påpeka att ingen larmgräns har passerats vid felsimuleringen, vilket innebär att processförändringen skulle undgå detektion om endast larmgränser användes för feldetektion.

Anledningen till att inget utslag noterats för T^2-residualerna är att avvikelserna skett utanför modellplanet (Rosén, 1998:119). Inom modellen kan uppenbarligen avvikelserna inte påvisas.



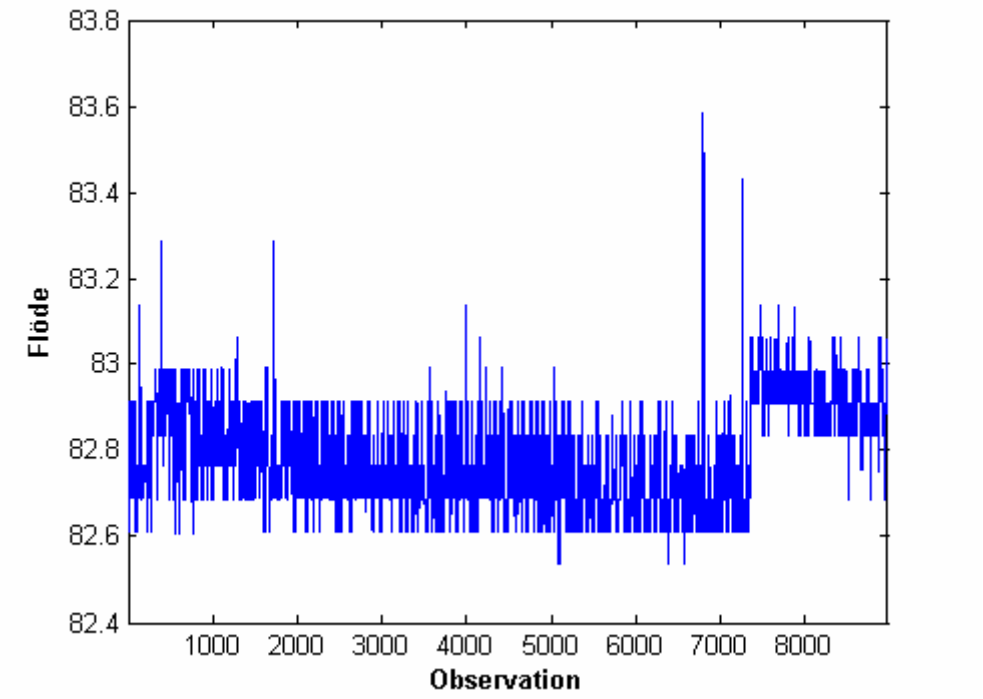
Figur 6.7 *SPE-residualer från kylvattenpumpfel-simuleringen för fall 1. 95 % konfidensgräns passeras strax före observation 3500 och trenden blir snart därefter tydlig. Simuleringen tog sin början vid observation 3460.*



Figur 6.8 *T²-residualer från kylvattenpumpfel-simuleringen för fall 1. I princip samtliga observationer befinner sig inom 95 % konfidensgräns.*

7 Fall 2 – driftsituation vid konstant flödesnivå

I det första fallet baserades PCA-modellen på träningsdata från perioder med växlande flöde för att modellen skulle bli så generell som möjligt. I ett försök att göra en mer specifik PCA-modell, som än bättre klarar av att modellera drift vid en särskild flödesnivå, utfördes ytterligare ett försök. I fall 2 kommer träningsdata respektive testdata istället att hämtas från en period där flödet hållits förhållandevis konstant. Även för fall 2 undveks driftperioder vid larmregistreringar genom studie av driftjournalen i Conwide System III. Samtidigt gav den tidigare studerade figur 5.1 information om perioder med relativt konstant flöde. Figur 7.1 visar flödet över den period som valts för fall 2.



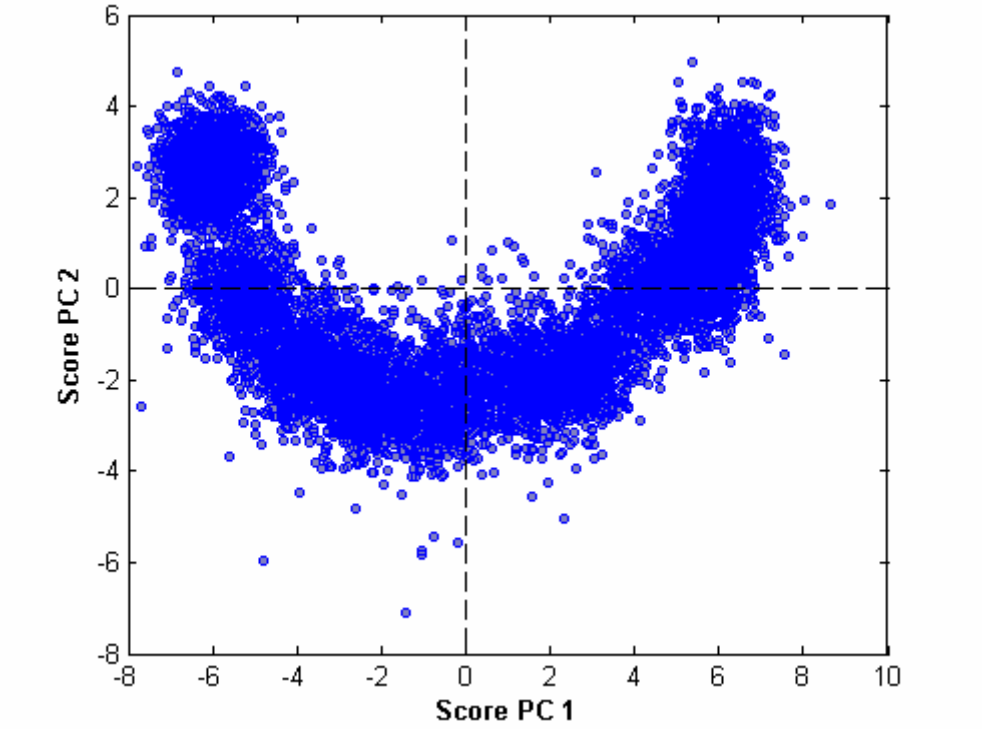
Figur 7.1 Flödet över perioden för fall 2. Flödet är synnerligen stabilt med endast små variationer.

Som tidigare nämnts ges en sammanställning av inkluderade variabler i bilaga 1 medan bilaga 2 utgör en förteckning över perioden från vilken tränings- respektive testdata hämtats.

För fall 2 bestod träningsdata inledningsvis av 9001 observationer av de 61 variabler som även inkluderats i fall 1. Då förekomsten av outliers undersöktes, och data granskades närmare, noterades att inga värden registrerats för variabel 50[33] under hela träningsperioden. Därför uteslöts variabel 50 ur PCA-modelleringen för fall 2. Efter en ny PCA, nu med 60 inkluderade variabler, noterades och raderades 26 outliers. Enligt samma princip som för fall 1 bestämdes sedan antalet inkluderade PC:er utifrån studier av scree- respektive RMSECV-plottarna. För fall 2 visade det sig tillräckligt att

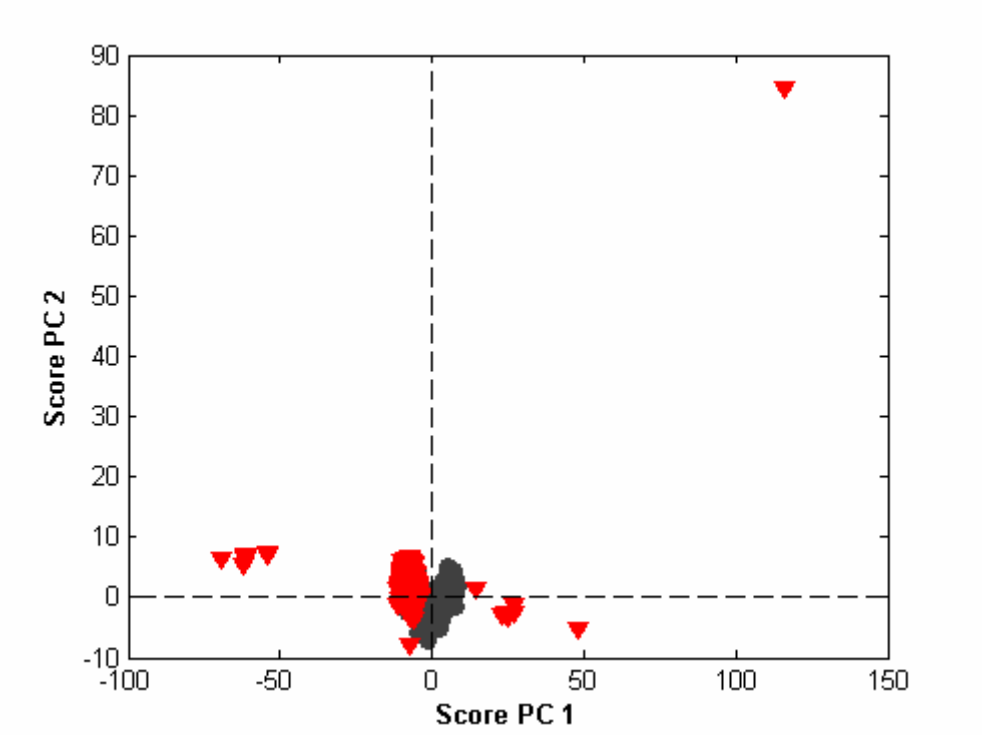
[33] Tätningsbox 2 temp

modellera de första 5 PC:erna för att få PCA-modellen att förklara den mesta variationen hos träningsdata. Figur 7.2 visar den PCA-modell som konstruerats utifrån träningsdata för fall 2 och figur 7.3 visar testdata projicerad på den av träningsdata konstruerade PCA-modellen (PC1 och PC2).

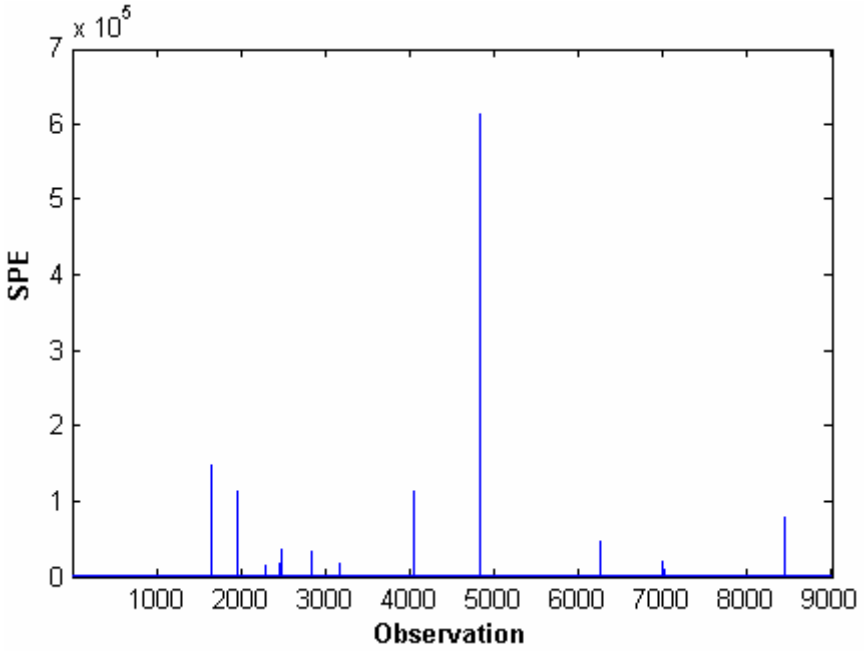


Figur 7.2 Score-plot över träningsdata projicerat på PCA-modellens första 2 PC:er vid fall 2.

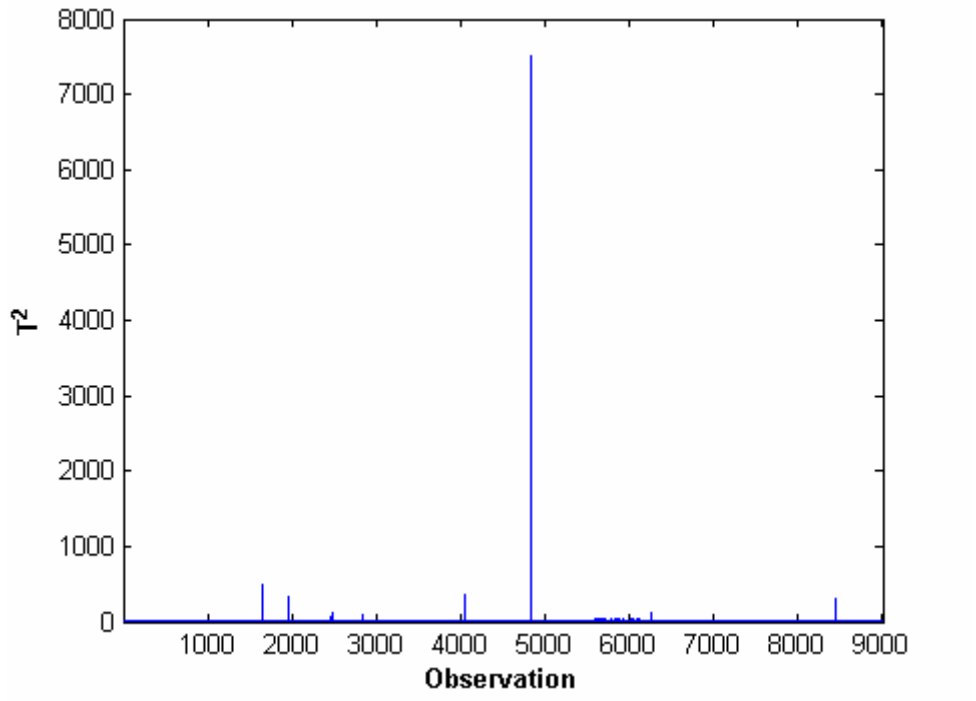
Skillnaderna mellan de två fallen blir tydlig vid en jämförelse av figur 7.3 och figur 6.2. Projiceringarna av testdata divergerar nämligen kraftigt för de två fallen. Observationerna från testdata för fall 2 projiceras långt från origo i förhållande till observationerna från träningsdata. De långa avstånden indikerar att observationerna från träningsdata och testdata skiljer sig från varandra. För att närmare studera skillnaderna granskas residualvärdesvektorerna i figurerna 7.4-7.7.



Figur 7.3 *Score-plot över testdata projicerat på PCA-modellens första 2 PC:er vid fall 2. Röda trianglar markerar observationer från testdata. Mörka prickar indikerar observationer från träningsdata.*

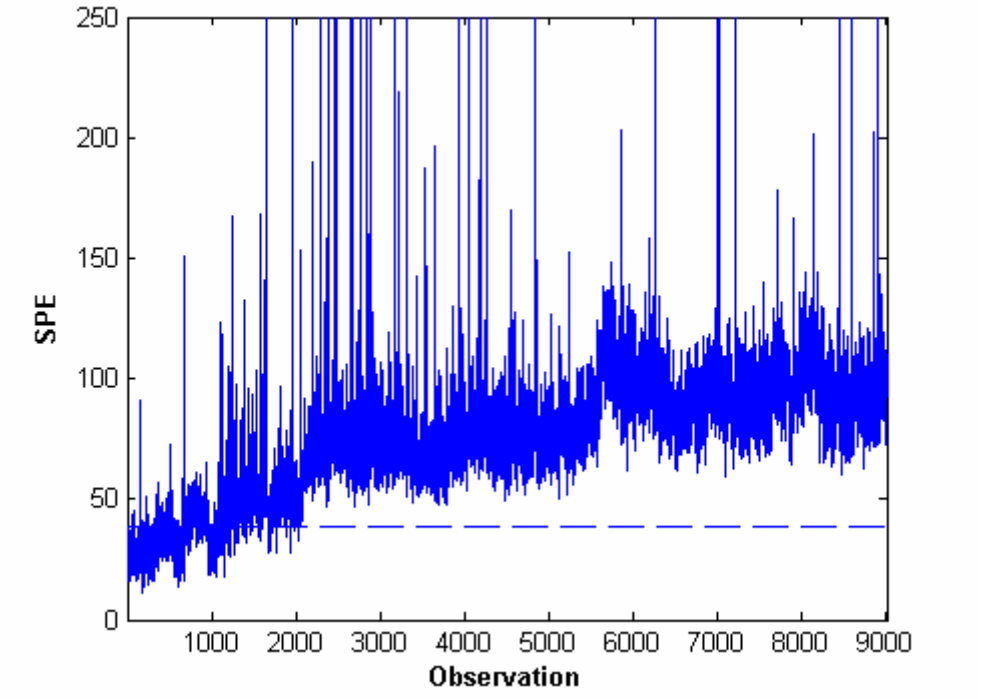


Figur 7.4 *SPE-residualer för testdata från fall 2. Extremt stora värden noteras för ett par observationer.*

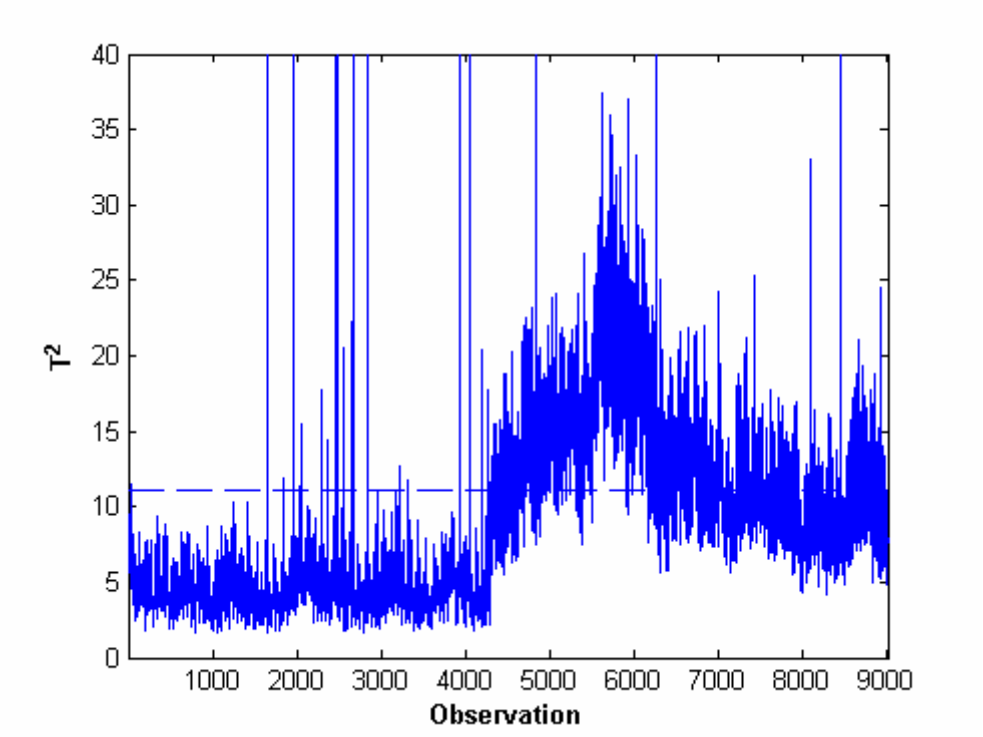


Figur 7.5 *T^2-residualer för testdata från fall 2. Även här noteras mycket stora värden för ett antal observationer.*

Redan vid en första anblick på residualplottarna i figur 7.4 och figur 7.5 noteras anmärkningsvärt höga värden på residualvärdena för ett par observationer. Särskilt SPE-värdena antar stundom kolossala värden. En närmare granskning av de båda residualvektorernas värden ges i figur 7.6 respektive figur 7.7.



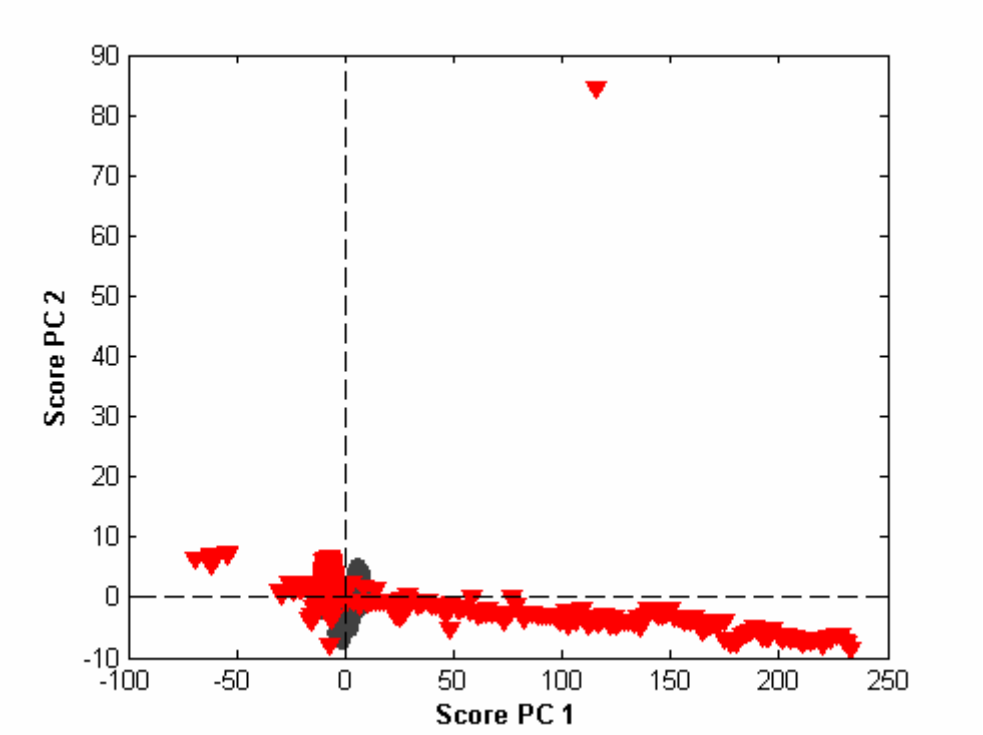
Figur 7.6 *Vid en närmare titt på SPE-residualerna inses att endast ett fåtal av observationerna befinner sig inom 95 % konfidensgräns.*



Figur 7.7 Även en stor del av T²-residualerna visar sig passera 95 % konfidensgräns.

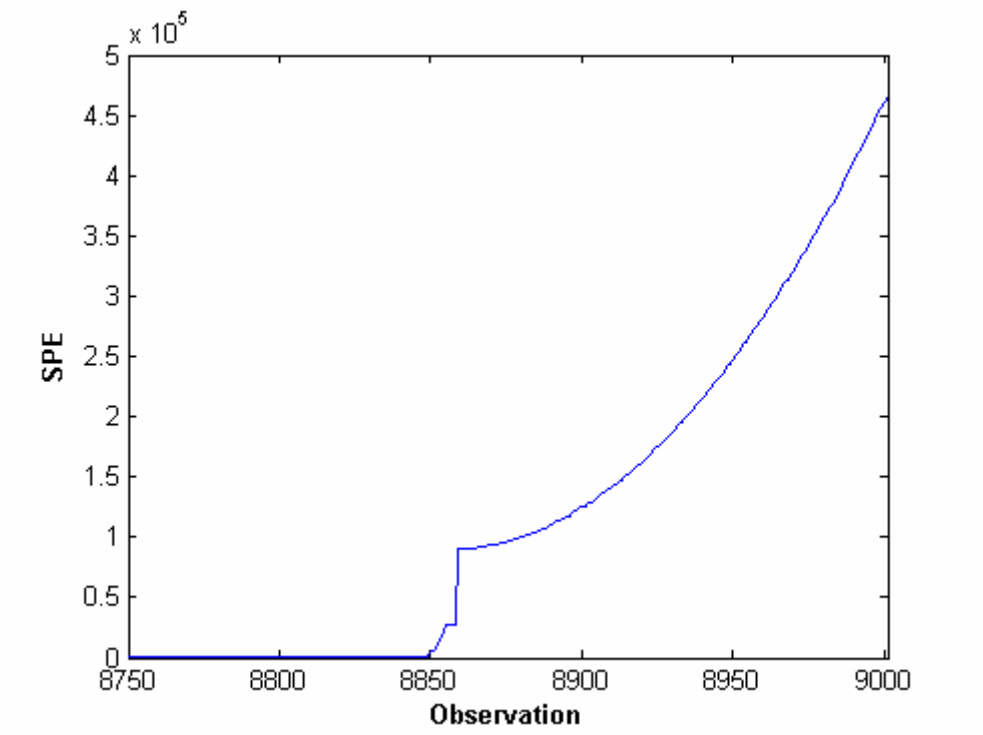
Granskningen tydliggör att testdata inte passar modellen särskilt väl. Som synes är det endast ett fåtal av observationernas residualvärden som understiger 95 % konfidensgräns. Med andra ord befinner sig merparten av observationerna från testdata för långt ifrån PCA-modellen för att de ska kunna klassificeras som normala. Trots att både träningsdata och testdata kommer från en period med synnerligen stabil flödesnivå skiljer sig alltså observationerna relativt kraftigt åt, enligt den konstruerade PCA-modellen. Modellen visade sig alltså vara i överkant känslig för variabelvärdes-avvikelser och frågan om den konstruerade PCA-modellen för fall 2 verkligen lämpar sig för driftövervakning är därmed berättigad.

Härnäst kommer emellertid samma simulering av kylvattenpumpfel som för fall 1 att studeras för fall 2. Eftersom PCA-modellen vid fall 2 baserats på data från drift med mer eller mindre konstant flödesnivå förväntas större avvikelser hos residualerna då förändringar uppstår i PT:et. Figur 7.8 anger det simulerade felets påverkan på PCA-resultatet vid fall 2.

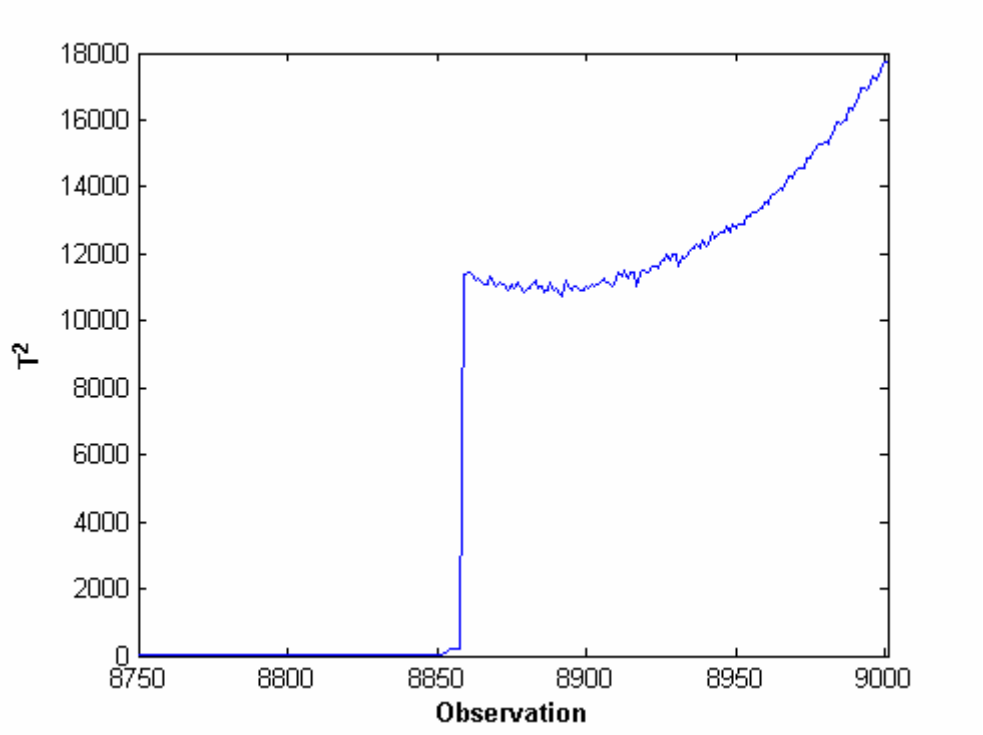


Figur 7.8 Score-plot över simulerat kylvattenpumpfel för fall 2. PT:en driver kraftigt åt höger med tiden.

Vid en jämförelse av figur 7.8 och figur 7.3 blir det tydligt att den simulerade feltrenden påverkat PT:en för testdata påtagligt. Så långt kan det anses att PCA-modellen fungerar planenligt, men det vore dock inte lätt, om ens möjligt, att göra distinktion mellan naturliga driftavvikelser och driftavvikelser till följd av någon form av fel eller onaturligt tillstånd hos processen. Avvikelserna från origo blir genast, även för mycket små variabelvärdesförändringar, stora och utrymme ges inte för några som helst variationer. SPE- och T^2-plottarna i figur 7.9 respektive figur 7.10, visar vidare att även residualvärdena ökar lavinartat för även små, till och med för naturliga, förändringar.



Figur 7.9 SPE-plot vid simulerat kylvattenpumpfel för fall 2.



Figur 7.10 T²-plot vid simulerat kylvattenpumpfel för fall 2.

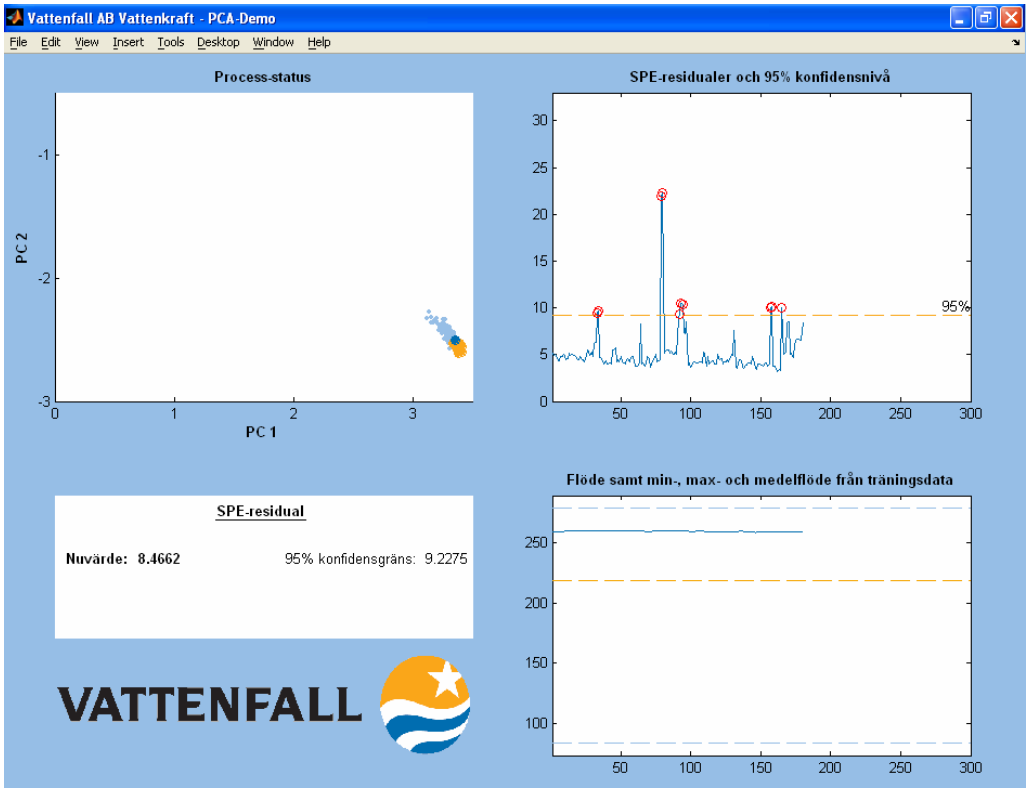
Vis från resultaten av PCA:n för fall 2 är det intressant att åter studera den inledande score-plotten för fall 2, vilken kan ses i figur 7.2. Kanske hade det varit möjligt att redan då förutspå att PCA-modellen för fall 2 skulle bli väl känslig för störningar. Score-plotten visar nämligen en tydlig rörelse hos PT:en, trots det faktum att processdata valts

från perioder med ytterst stabil flödesnivå. Att PT:en driver runt i modellplanet utan att flödesnivån ändrats mer än oansenligt tyder på att PCA:n i hög grad påverkas av variabelvärdesförändringar. Viss variation måste naturligtvis medges för att PCA-modellen ska fungera praktiskt.

8 Demonstrationsapplikation

För att åskådliggöra hur PCA skulle kunna användas för att övervaka det aktuella PT:et hos en vattenkraftsanläggning har en demonstrationsapplikation framtagits. Tanken med demonstrationsapplikationen var att, på ett mer lättbegripligt sätt, förklara hur en eventuell PCA-modellinstallation vid en av Vattenfall AB:s vattenkraftsanläggningar skulle kunna se ut och vilken information den i så fall skulle kunna bidra med. Som snart kommer poängteras, skulle mer gå att göra för att förbättra resultaten av PCA:erna och det ska därför understrykas att den framtagna applikationen endast är avsedd för demonstration, såsom namnet antyder. Det är emellertid så att demonstrations-applikationen tjänar till att indikera nyttan hos en väl utarbetad PCA.

Demonstrationsapplikationen avser att demonstrera ett möjligt övervakningsprogram som visar information om en vattenkraftsanläggnings aktuella PT och som indikerar eventuella avvikelser från det normala PT:et. Figur 8.1 är en bild över demonstrations-applikationen och visar den information som en operatör vid en DC eventuellt skulle kunna se om PCA utfördes kontinuerligt i realtid.



Figur 8.1 Demonstrationsapplikationen visar normal drift.

Demonstrationsapplikationen är i grunden en PCA-modell, baserad på samma träningsdata som användes vid fall 1, på vilken observationer från testdata projiceras med en sekunds intervall. Sekundfördröjningen innebär en slags realtidssimulering[34] av

[34] En riktig driftövervakningsapplikation skulle naturligtvis hämta processdata i realtid för att sedan presentera analysresultaten.

driftövervakning och applikationen blir därför inte enbart ett statiskt analysverktyg utan kan följa processen med tiden.

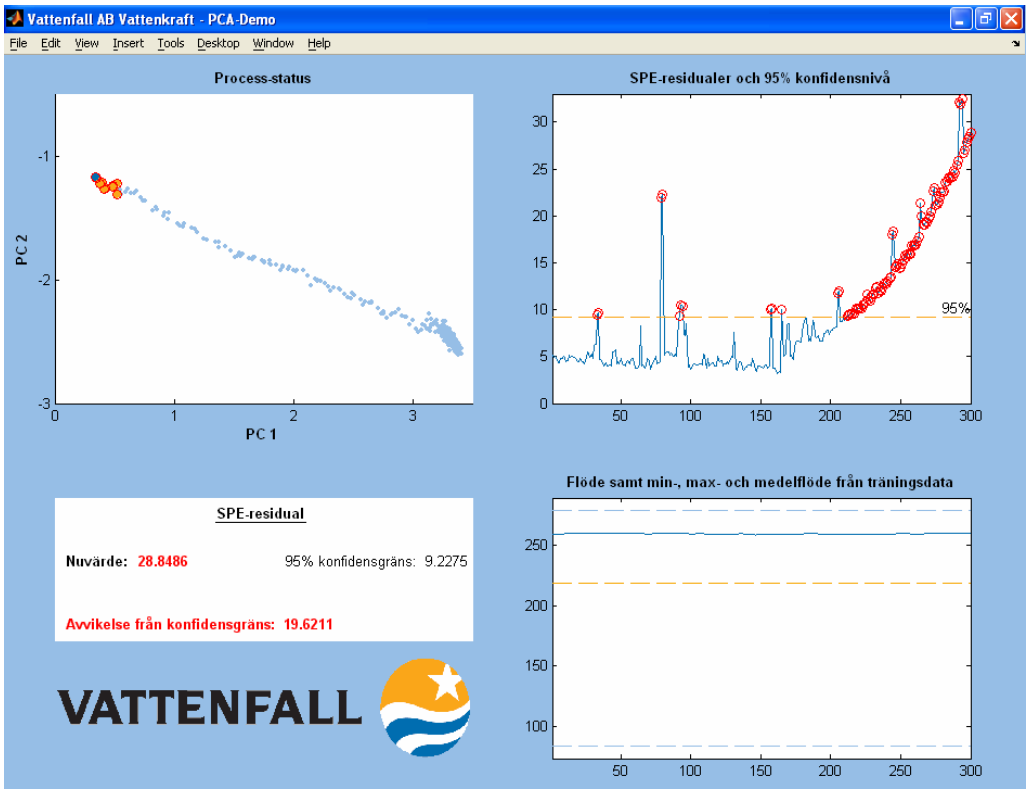
Vad demonstrationsapplikationen visar, i simulerad realtid, är information om det aktuella PT:et på flera olika vis. Genom att projicera de senaste tio observationerna i det plan[35] som spänns upp av de två första PC:erna, PC1och PC2, uppstår en så kallad mask vars rörelser visar hur PT:et ändras över tid. Utifrån maskens vandring i planet är det möjligt att dra slutsatser om förändringar i PT:et. Stabilt PT skulle resultera i att maskens rörelser minimeras. På samma vis innebär stora[36] rörelser hos masken att PT:et ändras avsevärt. Särskilt är vandringar i någon viss riktning intressanta eftersom de tyder på förändringar kopplade till någon form av trend. Genom att studera den loading-plot som förknippas med score-plotten kan information om respektive variabels påverkan på maskens vandring fås. Därmed skulle det vara möjligt att direkt se vilka variabler som påverkat PT:et att förändras i en viss riktning. Här kommer dock inte loading-plotten att beaktas.

Score-plotten visar på ett sätt observationernas avstånd till origo, vilken alltså kännetecknar processens normaltillstånd. Score-plotten är dock endast en två-dimensionell bild av modellen och döljer därför viss information. Korrekt information om observationernas avstånd till modellen ges istället av SPE- respektive T^2-plottarna som visar avstånden beräknade för samtliga PC:er inkluderade i modellen. Stadigt ökande residualvärden tyder på en trend av ökande avstånd mellan observationer och modell. Givet är också en så kallad informationsruta, vilken visar de båda residual-värdena, SPE och T^2, numeriskt. Observationer utanför angivna konfidensgränser markeras med röd färg i samtliga plottar samt i informationsrutan, där även avvikelsen från konfidensgränsen då visas.

Det aktuella PT:et, och därmed demonstrationsapplikationen, är, i enlighet med vad som tidigare konstateras, känslig för flödesnivåförändringar. Om flödesnivån ändras kommer hela processen påverkas och således kan inte längre ett stabilt PT förväntas. För att kunna få insikt i huruvida eventuella residualvärdesökningar beror på att en ny drift-situation råder eller på att andra, icke önskade, variationer uppstått presenteras även en flödes-plot. Flödes-plotten visar, utöver den aktuella flödesnivån, min-, max- och medelflödet över träningsperioden. Om den aktuella flödesnivån nära följer tränings-periodens medelflöde kan den konstruerade PCA-modellen antas utgöra en god mall för PT:ets rörelser och eventuella avvikelser från modellen förmodas därför komma av oönskade variationer hos driften. Vilka variationerna är kan sedan undersökas vidare så att eventuella åtgärder kan vidtas för att undvika allvarliga fel eller tillbud. Om den aktuella flödesnivån däremot börjar avvika från den flödesnivå på vilken PCA-modellen är baserad är förändringar hos PT:et att vänta. Då bör istället PCA-modellen förkastas till fördel för en ny modell.

[35] Projektionerna i planet utgör en score-plot.
[36] Med stora avses i det här sammanhanget stora i jämförelse med variationer hos projektionerna av träningsdata på PCA-modellen.

Figur 8.2 visar hur demonstrationsapplikationen först registrerat normal drift för att sedan, ungefär vid tidpunkt 200, indikera stadigt ökande residualer. De simulerade feltrenderna inleds vid tidpunkt 180. Notera att flödet hållits påfallande stabilt under hela testperioden, varför de ökade residualvärdena inte gärna kan bero på att driftsituationen har förändrats.



Figur 8.2 Demonstrationsapplikationen har noterat förändringar i PT.

De simulerade feltrenderna syns markant i figur 8.2. Samtidigt ska det noteras att inga av de modellerade variablernas larmgränser har överskridits. Med andra ord är de simulerade trenderna inget som skulle upptäckas av de nuvarande larmgränserna. Se åter figur 6.5 för en grafisk representation.

När ett misstänkt fel upptäckts i driftövervakningsapplikationen är tanken naturligtvis att operatören ska kunna ta reda på vilka variabler det är som bidrar till avvikelserna i plottarna. Genom att då studera variablernas inbördes bidrag till de ökande residualvärdena, SPE respektive T^2, skulle det därmed vara möjligt att urskilja variabler som bidrar mer än vanligt. Utifrån informationen om att det är vissa temperaturhöjningar som påverkat residualvärdenas ökningar kan misstankar om felande kylning väckas.

9 Förfining av PCA för vattenkraftsdata

Det finns ett flertal åtgärder med vilka, antingen enskilt eller i kombination med varandra, det vore möjligt att förbättra resultaten från PCA:n. Examensarbetet har visat hur PCA utförs samt förklarat hur analysen kan användas och bidra med information om det aktuella PT:et. Nästa steg i analysarbetet vore att individanpassa PCA:n till Vattenfall AB för applicering på processdata från vattenkraftsdrift. Nedan följer en sammanställning av åtgärder som, om de adresserades, skulle leda till att PCA:n på ett bättre sätt förklarar vattenkraftverksdriften.

9.1 Förbehandling av data

Förbehandling av data genom applicering av filter kan bidra till att förbättra resultaten av en PCA. GIGO innebär att indata noga måste kontrolleras för att utdata ska bli användbart (Aguilar et al., 1999:8). Ett vanligt problem vid mätning av variabelvärden är nämligen mätstörningar orsakade av bland annat elektromagnetiska störningar (Rosén, 1998:27). Rengöring och kalibrering av mätutrustningen är viktigt för att minimera uppkomsten av mätstörningar, men genom att filtrera data innan PCA:en kan mätstörningarnas påverkan på analysresultatet minskas ytterligare (Ibid.). Störningsreducering kan åstadkommas med hjälp av digitala filter och framför allt icke-linjära filter har visat sig vara effektiva. Rosén (1998:166) nämner särskilt medianfilter och Finite Response Impulse Median Hybrid-filter (FMH-filter) som han menar lämpar sig väl vid processdata-analyser, tack vare förmågan att bevara diskontinuiteter hos mätsignalerna. FMH-filtret beskrivs nedan av ekvation 18.

$$\hat{y}_{FMH,l}(k) = median\left(\frac{1}{l}\sum_{i=k-l}^{k-1} y(i),\ y(k),\ \frac{1}{l}\sum_{i=k+1}^{k+l} y(i)\right) \tag{18}$$

Med andra ord är filtrets utdata, $\hat{y}$, medianvärdet av tre uppskattningar av $y(k)$; medelvärdet av l tidigare värden, $y(k)$ själv och medelvärdet av l framtida värden. Filtret är därmed inte applicerbart i realtidssituationer. Längden på l anpassas till den givna situationen (Rosén, 1998:32).

9.2 Tidsserier

Vid PCA antas observationerna, $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$, vara oberoende (Jolliffe, 1986:205). Vid tidsserieanalys är emellertid observationerna inte helt oberoende eftersom en variabels nuvärde beror delvis på föregående värde. Beroendet mellan observationernas värden orsakas av deras relativa närhet i tiden. $\mathbf{x}_h$ och $\mathbf{x}_i$ är i regel starkt beroende om $|h-i|$ är litet, med minskande beroende allteftersom $|h-i|$ ökar (Ibid.). I examensarbetet har processdata, som består av ett flertal variablers tidsserier, därmed inte behandlats som tidsserier i sin rätta bemärkelse. Processdata har visserligen genomgående varit ordnad efter tid, vilket inneburit att eventuella trender varit möjliga att upptäcka vid applicering av testdata på konstruerad PCA-modell. Däremot har observationernas ordning inte spelat någon roll vid konstruktionen av PCA-modellen utifrån träningsdata. Med andra ord hade PCA-modellerna varit identiska även om observationerna i träningsdata kastats om. Anledningen till varför outliers helt enkelt kunde raderas vid konstruktion av PCA-

modellerna[37] var just att observationerna ansågs vara helt oberoende av varandra och att den inbördes ordningen därför inte spelade någon roll. Vid en tidsserieanalys, i ordets rätta bemärkelse, hade det däremot inte varit möjligt att radera outliers. Alternativet hade då istället varit att uppskatta värden för samtliga outliers (Rosén, 1998:36) för att inte förskjuta den inbördes ordningen mellan observationer i processdata.

Det finns ett par olika sätt genom vilka det är möjligt att ta hänsyn till beroendet mellan olika observationers värden. Exempelvis har tidigare konstaterats att PCA baseras på en kovariansmatris, men vid tidsserieanalys är det inte bara möjligt att beräkna kovarianser mellan variabler uppmätta vid samma tidpunkt, utan det är även möjligt att beräkna kovarianser mellan variabler vid olika tidpunkter[38] (Jolliffe, 1986:206). Att ta hänsyn till tidsseriedata vid PCA är, enligt Jolliffe (1986:209), emellertid en komplicerad uppgift som kräver sofistikerad matematik. Han menar vidare att praktisk erfarenhet kring teknikerna saknas varför det är svårt att utvärdera deras nytta. Genom att använda någon av de nedan nämnda metoderna skulle PCA:n hur som helst göras mer anpassad för applicering på processdata. Eftersom PCA främst är en deskriptiv metod påverkas resultatet inte i hög grad av ett visst beroende mellan observationernas värden (Jolliffe, 1986:205). Det är också precis därför som PCA i regel utförs utan hänsyn tagen till tidsberoendet (Carlsson, 2006-10-16). Trots allt vore det intressant att se huruvida PCA-resultaten skulle gå att förbättra om relevant tidsserieanalysmetod användes.

9.2.1 Frekvensdomän

Jolliffe (1986:207) menar att PCA med fördel utförs i frekvensdomänen vid tidsserieanalys och Brillinger (1975) ägnar ett helt kapitel[39] åt PCA i frekvensdomänen. Vidare ger Brillinger (1975:355) exempel på hur PCA i frekvensdomänen använts för att analysera temperaturmätningar från 14 olika väderstationer. Kopplingen mellan PC:er beräknade i frekvensdomänen och PC:er beräknade enligt den konventionella metoden inbegriper bland annat Hilbert-transformer och är, enligt Jolliffe (1986:208), inte helt lättbegriplig. Den intresserade hänvisas till Brillinger (1975).

9.2.2 Förbehandling av data vid tidsserieanalys

Vid hantering av tidsseriedata är det, som tidigare konstaterats, viktigt att vara medveten om att variablernas nuvärden beror av tidigare värden. Därför är det önskvärt att hänsyn tas till viss historik. Det kan göras med hjälp av så kallade Finite Impulse Response-filter[40] (FIR-filter) (Rosén, 1998:76). Ett FIR-filter är i grunden ett linjärt kausalt digitalt filter[41] enligt ekvation 19, där samtliga *a*-koefficienter är lika med noll (Rosén, 1998:29).

---

[37] Se avsnitt 5.1.4.1.

[38] I det icke tidsberoende fallet, där observationerna $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ är oberoende, är kovarianserna mellan element av $\mathbf{x}_i$ och $\mathbf{x}_j$ lika med noll då $i \neq j$ (Jolliffe, 1986:206).

[39] Kapitel 9.

[40] FIR-filter benämns ibland Moving Average-filter (MA-filter) (Rosén,1998:29).

[41] Ett digitalt filter som endast använder nuvärde och historiska värden för att beräkna utdata.

$$\hat{y}(k) = -a_1\hat{y}(k-1) - a_2\hat{y}(k-2) - \ldots - a_n\hat{y}(k-n) + b_0 y(k) + \ldots + b_m y(k-m) \qquad (19)$$

Ekvation 19 skrivs då enligt ekvation 20.

$$\hat{y}(k) = b_0 y(k) + \ldots + b_m y(k-m) \qquad (20)$$

En stor nackdel med FIR-filter är det ökade antalet variabler som då blir lika många som det ursprungliga antalet värden multiplicerat med längden på tidsfönstret, *m* (Rosén, 1998:79). Med begränsad datorkraft kan fördröjningen därför bli avsevärd (Rosén, 1998:29).

9.3 Val av data

Efter en jämförelse av resultaten från de två fallen diskuterade i kapitel 6 respektive kapitel 7 blir det tydligt att PCA-modellen är starkt beroende av valet av data. Att utföra fler tester, både på data från perioder med stabil flödesnivå och på data från perioder med skiftande flödesnivå, vore därför intressant och skulle kunna ge information om hur bred PCA-modellen bör göras för att så bra som möjligt detektera avvikande PT och samtidigt fungera i praktiken.

Även vilka variabler som bör inkluderas i, och vilka variabler som bör uteslutas ur, PCA-modellen vore möjligt att granska närmare. Inom examensarbetet har PCA-modellerna baserats på data från en hel vattenkraftsanläggning och det är tänkbart att resultaten från analyserna skulle bli både tydligare och lättare att tolka om modellerna istället baserades på data från en enskild del av samma anläggning. Driftsprocessen i anläggningen är högst dynamisk, med beroenden och tidsförskjutningar, och genom att isolera analysen till en mindre del skulle eventuellt mycket av komplexiteten undvikas.

9.4 Tänkbara övriga förbättringar

Vidare studier skulle även kunna göras på eventuell efterbehandling av resultaten från PCA:erna. Genom medelvärdesberäkning av exempelvis SPE- och T^2-värden över ett visst tidsintervall skulle till exempel trender i avvikande residualvärden kunna detekteras (Rosén, 1998:114). På så vis skulle ytterligare automation erhållas.

Närmare studier av loading-plotten skulle även kunna ge information om på vilket sätt respektive variabel påverkar driftsprocessen. Eventuella samband mellan variabler vore därmed möjliga att påvisa. Nästa steg vore därefter att undersöka på vilket sätt exempelvis regressionsmetoden PLS, beskriven i avsnitt 10.1, skulle kunna bidra med information om samband bland variablerna i vattenkraftverket.

10 Övriga tänkbara metoder

Examensarbetet har, i huvudsak, behandlat PCA. Vad metoden innebär och hur den går att applicera på processdata från ett vattenkraftverk är saker som diskuterats genom rapporten. I kapitel 3 nämndes emellertid att det inom litteraturen förekommer en uppsjö av metoder för att behandla stora datamängder. En stor andel av metoderna är utvecklade för ett visst specifikt ändamål och lämpar sig därför väl endast för tillämpningar inom ett ytterst snävt område. Likväl finns ett par, mer generella, metoder som frekvent refereras. Några av de mest brukade metoderna kommer i korthet att beskrivas nedan.

10.1 Partial Least Squares Projection to Latent Structures (PLS)

Regressionsmetoden Partial Least Squares Projection to Latent Structures (PLS) är en utvidgning av PCA (Eriksson et al., 2006:63). Även vid PLS fås scores och loadings, om än inte samma som vid PCA (Wise et al., 2006:149). Vid PLS är de istället ordnade så att de, givet indata, på bästa sätt förutspår utdata (Ibid.). Genom PLS är det möjligt att lära avancerade samband mellan olika variabler (Eriksson et al.,2006:63). Givet en viss processsituation går det därför, att med PLS, approximera ett specifikt variabelvärde. Således ges möjlighet att upptäcka fall då det observerade värdet i stor omfattning avviker från vad som är förväntat. På så sätt skulle potentiellt olämpliga situationer kunna undvikas.

PLS är en multivariat linjär regressionsmetod som, enligt exempelvis Rosén:168 och Jackson:263, med fördel kan användas tillsammans med PCA. PLS har tidigare tillämpats vid just processdata-analyser (Röttorp, 1999, Skagerberg & Sundin, 1993, Andersson et al., 2003) varför det vore tänkbart att studera hur PLS kan bidra vid analys av en vattenkraftanläggning.

10.2 Artificiella neurala nätverk (ANN)

Det har rått en viss hajp kring artificiella neurala nätverk (ANN) vilket gjort att missuppfattningar angående vad ett ANN kan åstadkomma har spridits (Anderson & McNeill, 1992:1). ANN är egentligen inget annat än en icke-linjär regressionsmetod som baseras på linjärkombinationer av indata (Hastie et al., 2001:347). ANN är samtidigt elektroniska modeller som söker efterlikna den neurala strukturen hos en människohjärna (Anderson & McNeill, 1992:2). Hjärnan lär av erfarenhet och detsamma gäller för ANN. Det bör noteras att träning av ANN kan vara kostsamt och att träningen i regel kräver stor datorkraft (Anderson & McNeill, 1992:26). Den mest använda ANN-strukturen är den så kallade feedforward back-propagation network (Anderson & McNeill, 1992:1) som ofta används exempelvis vid mönsterigenkänning (Anderson & McNeill, 1992:63). Även processövervakning är ett användningsområde inom vilket ANN visat sig användbara (Anderson & McNeill, 1992:64-65).

Enligt professor Bengt Carlsson[42] (2006-10-16) finns dock ingen egentlig anledning till att använda en icke-linjär regressionsmetod om inte linjära metoder först visat sig olämpliga. Eftersom ingen regression studerats inom examensarbetet vore det naturliga därför att först undersöka hur väl linjära regressionsmetoder lämpar sig för att analysera processdata. En tänkbar metod att börja med skulle kunna vara PLS.

10.3 Independent Component Analysis (ICA)

Independent Component Analysis (ICA) är en metod som på många sätt kan liknas vid PCA (Stone, 2004:9). ICA är emellertid en metod som används vid en viss typ av problemlösning, så kallad Blind Source Separation (BSS) (Stone, 2004:13). BSS innebär att väldigt lite angående ursprungssignalernas egenskaper är känt (Stone, 2004:6). Det främsta användningsområdet för ICA har hittills varit att urskilja olika ljudkällor ur en mixtur av ljud; det så kallade cocktailparty-problemet (Hyvärinen & Oja, 2000:1).

I huvudsak är den stora skillnaden mellan PCA och ICA[43] att PCA ger okorrelerade PC:er medan ICA ger oberoende IC:er (Stone, 2004:129). För att få ner antalet variabler användas ibland PCA som förbehandlingsmetod vid beräkning av ICA (Stone, 2004:179). ICA är en metod under utveckling och metodens begränsningar är inte till fullo utforskade (Stone, 2004:10) och det skulle, enligt professor Stefan Arnborg[44], eventuellt vara intressant att studera hur ICA kan bidra till fördjupad kunskap om driften av ett vattenkraftverk utifrån processdata. ICA behandlas dock inte vidare inom examensarbetet.

[42] Bengt Carlsson är professor i reglerteknik vid avdelningen för systemteknik, institutionen för informationsteknologi, Uppsala universitet.
[43] Se Stone (2004:129) för en mer ingående förklaring av skillnaderna mellan PCA och ICA.
[44] Stefan Arnborg är professor i datalogi vid institutionen numerisk analys och datalogi, KTH.

11 Slutsats

Det råder ingen tvekan om att PCA är ett hjälpmedel som kan vara användbart för att, ur processdata, utvinna tidigare okänd information angående industriell anläggningsdrift. Flera tidigare studier pekar på potentialen hos PCA (Andersson et al., 2003, Skagerberg & Sundin, 1993 samt Rosén, 1998) och examensarbetet har gett ett exempel på hur PCA skulle kunna användas praktiskt i en övervakningssituation. Användbarheten och förmågan att upptäcka avvikande PT ifrågasätts heller inte, men samtidigt har det visat sig att appliceringen av analysmetoden inte är helt problemfri.

Två olika PCA-modeller konstruerades inom examensarbetet. Vid det första fallet, fall 1, användes processdata från perioder med varierande flödesnivå för att ge en bred PCA-modell som klarar av att modellera olika typer av driftsituationer. Det visade sig att avvikelser från normalt PT syntes tydligt i resultaten från PCA:n. I alla fall kunde PCA-modellen tidigt detektera det simulerade kylvattenpumpfelet. Fler tester behöver naturligtvis göras innan praktisk implementation är möjlig, men det faktum att PCA-modellen ger utslag för avvikande driftsituation är i alla fall en indikation på att metoden skulle kunna fungera för att påvisa avvikande PT hos en vattenkraft-anläggning.

PCA-modellen som konstruerades för det andra fallet, fall 2, baserades istället på processdata från perioder med avsevärt mindre flödesnivåvariationer. Processdata valdes från perioder där flödesnivån varit så stabil som möjligt för att få en PCA-modell som på bästa sätt klarar av att modellera en viss typ av driftsituation med en viss specifik flödesnivå. Även vid fall 2 märktes det simulerade felet tydligt i form av avvikande PT, men modellen var samtidigt så känslig för störningar och avvikande variabelvärden att också observationer från drift vid normal driftsituation passade modellen dåligt.

Utifrån resultaten från de båda fallen inses lätt att PCA-modellen vid fall 1 är den modell, av de två, som skulle fungera bäst i en praktisk tillämpning. Däremot skulle en kombination av de båda fallen antagligen vara det mest lämpliga alternativet för vidare studier. Perioder av processdata bör väljas med största omsorg för att göra PCA-modellen så pass generell att den klarar av att modellera observationer från olika typer av driftsituationer, samtidigt som den är tillräckligt känslig för avvikande variabel-värden att den lyckas upptäcka när PT:et är på väg att ändras.

En vattenkraftsanläggning består av många olika delar som i olika hög grad påverkas av störningar och som samtidigt påverkar varandra. Driften varierar dessutom i hög utsträckning från dag till dag. På det hela taget är det en ytterst dynamisk process vilket innebär att den är komplicerad att övervaka. Det krävs därför vidare studier innan metoden kan användas praktiskt och bidra till övervakningsarbetet med sin fulla potential. Något som dessutom noggrant bör utredas är vad som egentligen eftersöks; det vill säga vilket syftet bakom processdataanalyserna är. Det råder en viss hajp kring begreppet Data Mining men ingen metod kommer automatiskt, utan hjälp från människans erfarenhet och processkännedom, att finna några sanningar. Genom en väl utarbetad plan där ett tydligt mål eftersträvas kan metoderna anpassas till

Vattenfall AB:s anläggningar och göras mer avgränsade för att ge upplysningar som kan bidra till en ökad processkunskap.

Referenslista

Tryckta referenser

Aguilar, Marc; Rautert, Tankred & Pater, Alexander J.D., 1999, *Business Process Simulation: A Fundamental Step Supporting Process Centered Management*, Andersen Consulting, Amsterdam

Anderson, Dave & McNeill, George, 1992, *Artificial Neural Networks Technology*, Data & Analysis Center for Software, New York

Andersson, Magnus; Olsson, Jenny; Röttorp, Jonas & Ek, Mats, 2003, *Tillämpning av multivariata metoder för övervakning av skogsindustriella biologiska reningsanläggningar*, IVL Svenska Miljöinstitutet AB, Stockholm

Berry, Michael J.A. & Linoff, Gordon, 1997, *Data Mining Techniques for Marketing, Sales and Customer Support*, Wiley Computer Publishing, Kanada

Berson, Alex; Smith, Stephen & Thearling, Kurt, 2000, *Building Data Mining Applications for CRM*, McGraw Hill

Bishop, Christopher M., 1995, *Neural Networks for Pattern Recognition*, Oxford University Press Inc., New York

Brillinger, David R., 1975, *Time Series – Data Analysis and Theory*, Holt, Rinehart and Winston, Inc., USA

Chen, An-Pin & Chen, Chia-Chen, 2006, *A new efficient approach for data clustering in electronic library using ant colony clustering algorithm*, Emerald Group Publishing Limited

Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Trygg, J.; Wikström, C. & Wold, S., 2006, *Multi- and Megavariate Data Analysis – Part I; Basic Principles and Applications*, Umetrics, Umeå

Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P. & Uthurusamy R., 1996, *Advances in knowledge discovery and data mining*, American Association for Artificial Intelligence, Menlo Park, Kalifornien

Giudici, Paulo, 2003, *Applied Data Mining – Statistical Methods for Business and Industry*, John Wiley & Sons Ltd, Chichester, England

Glemme, Minna, 2001, *Utveckling av larm för tekniska system i vattenkraftstationer*, Uppsala

Hastie, Trevor; Tibshirani, Robert & Friedman, Jerome, 2001, *The Elements of Statistical Learning – Data Mining, Inference and Prediction*, Springer Series in Statistics, Kanada

Hyvärinen, Aapo & Oja, Erkki, 2000, *Independent Component Analysis: Algorithms and Applications*, Neural Networks Research Centre, Helsingfors, Finland

Jackson, J. Edward, 1991, *A User's Guide to Principal Components*, John Wiley & Sons Inc., Kanada

Jolliffe, I.T., 1986, *Principal Component Analysis*, Springer-Verlag, New York

Masman, Fredrik & Blom, Daniel, 2000, *Conwide i Luleälven: En uppföljning av insamlad information och faktiska händelser*, Luleå

Rosén, Christian, 1998, *Monitoring Wastewater Treatment Systems*, Universitetstryckeriet, Lund

Råde, Lennart & Westergren, Bertil, 1998, *Mathematics Handbook for Science and Engineering*, Studentlitteratur, Lund

Röttorp, Jonas, 1999, *Multivariat modellering – ett kraftfullt verktyg för processoptimering*, IVL Svenska Miljöinstitutet AB, Stockholm

Skagerberg, Bert & Sundin, Lasse, 1993, *Övervakning och styrning av processförlopp i flera dimensioner – Integrerad processintelligens*, ur ABB Tidning 4/93

Stone, James V., 2004, *Independent Component Analysis*, The MIT press, Cambridge, Massachusetts

The MathWorks, 2006a, *MATLAB® - The Language of Technical Computing*, The MathWorks Inc., Natick

The MathWorks, 2006b, *Statistics Toolbox – For Use with MATLAB®*, The MathWorks Inc., Natick

Wise, Barry M.; Gallagher, Neil B.; Bro, Rasmus; Shaver, Jeremy M.; Windig, Willem & Koch, R. Scott, 2006, *PLS_Toolbox 4.0 – for use with Matlab™*, Eigenvector Research Incorporated, Wenatchee

Övriga referenser

Carlsson, Bengt, 2006, Intervju 2006-10-16

Conwide, http://www.conwide.se/

Nationalencyklopedin, http://www.ne.se/

Sagonas, Konstantinos, 2005, Föreläsningar i kursen Informationsutvinning, Uppsala universitet

Bilaga 1: Givare vid vattenkraftverket

Bilaga 1 har av sekretess-skäl uteslutits från den officiella versionen av rapporten.

Bilaga 2: Tränings- och testdata

Förteckning över tränings- och testdata för de två olika fallen. Se bilaga 1 för en förteckning över de variabler som inkluderats i modellerna.

Fall 1

Träningsdata	**Sampelintervall 10 minuter**
060412-060425	10.00-10.00
060603-060613	00.00-00.00
060621-060626	12.00-20.00
060708-060716	00.00-00.00
060818-060824	00.00-00.00
060903-060907	00.00-00.00
060915-060922	12.00-12.00
060926-061002	00.00-00.00
061021-061029	12.00-12.00
061107-061115	00.00-00.00

Testdata	**Sampelintervall 1 sekund**
060503	10.30-11.30

Fall 2

Träningsdata	**Sampelintervall 1 sekund**
060408	00.00-02.30

Testdata	**Sampelintervall 1 sekund**
060408	02.30-05.00

Bilaga 3: Matlab-kod

```
% pca_demo.m
% Applikation avsedd att demonstrera hur principalkomponentanalys
% (PCA) kan användas i en driftövervakningssituation vid ett av
% Vattenfall AB:s vattenkraftverk.
% Kräver utöver tränings- respektive testdata även Statistics Toolbox
% samt filen 'vattenfall.bmp'.
% Del av ett examensarbete (20p) utfört vid Uppsala universitet;
% civilingenjörsprogrammet System i Teknik och Samhälle (STS).

% Upphovsman: Björn Thorsélius


% Argument
% --------
% traindata: träningsdata på vilken PCA-modellen baseras.
% testdata:  testdata vilken projiceras på PCA-modellen.

function pca_demo(traindata,testdata)

% Färgkodning av plottar
bla=[0/255 109/255 176/255];
gul=[250/255 166/255 25/255];
ljusbla=[150/255 190/255 230/255];

mask=9;               % Masklängd-1 (huvudet kommer till)
signiv=0.05;          % Signifikansnivå för de statistiska testerna
tidstopp=0.1;         % Tidsfördröjning för realtids-simuleringen

medelv=mean(traindata);          % Medelvärden för träningsdata
stavv=std(traindata); % Standardavvikelse för träningsdata

[m n]=size(traindata);           % Storlek träningsdata
standard=zscore(traindata);      % Standardisering av träningsdata
[loads score varians T]=princomp(standard);% PCA-beräkningarna

egenv=sort(eig(cov(standard)),'descend');  % Egenvärden sorteras

m=length(testdata);               % Storlek av testdata
flowplot=mean(traindata(:,3));    % Medelflöde från träningsdata
flowplotMin=min(traindata(:,3)); % Minflöde från träningsdata
flowplotMax=max(traindata(:,3)); % Maxflöde från träningsdata

logo=imread('vattenfall.bmp','bmp');  % Inläsning av Vattenfalls
    logotyp

%% Fönster 1: Scree-plot för att bestämma #PC:er
figure(1)
set(gcf,'NumberTitle','off','Name','Vattenfall AB Vattenkraft - PCA-
    Demo','Color',[ljusbla],'Toolbar','none')
clf
subplot(12,12,[2:11 14:23 26:35 38:47 50:59 62:71 74:83 86:95 98:107])
plot(2:25,varians(2:25),'-o','Color',[gul])% Scree-plot
title('Scree-plot','Fontweight','bold')
xlabel('Antal principalkomponenter','Fontweight','bold')
ylabel('Egenvärde','Fontweight','bold')
axis([2 25 0 4])
set(gca,'xtick',[2:2:25],'xminortick','on','xgrid','on')
```

```
subplot(12,12,[121:126 133:138]) % Vattenfalls logotyp
image(logo)
set(gca,'XColor',[ljusbla],'YColor',[ljusbla],'xtick',[],'ytick',[])

%% Val av #PC:er
ruta1=uicontrol('Style','text','string','Ange det antal PCs som ska
    inkluderas i modellen:','Fontweight','bold','Position',[760 150 150
    30],'backgroundcolor',ljusbla);

slider1=uicontrol('Style','slider','value',2,'min',2,'max',n,'sliderst
    ep',[1/(n-2) 5/(n-2)],'Position',[735 120 250
    25],'callback',['uicontrol
    (''Style'',''text'',''string'',[int2str(get(gcbo,''value'')) ''
    PCs''],''Fontweight'',''bold'',''Position'',[880 80 40
    25],''backgroundcolor'',[150/255 190/255 230/255]);']);

ruta2=uicontrol('Style','text','string','2
    PCs','Fontweight','bold','Position',[880 80 40
    25],'backgroundcolor',ljusbla);

ruta3=uicontrol('Style','pushbutton','string','OK','Fontweight','bold'
    ,'Position',[925 85 60 30],'callback','uiresume(gcf);');

uiwait(gcf)

antalpc=get(slider1,'value');

% Konfidensgräns för SPE
theta1=sum(egenv(antalpc+1:n));
theta2=sum(egenv(antalpc+1:n).^2);
theta3=sum(egenv(antalpc+1:n).^3);
h0=1-(2*theta1*theta3)/(3*theta2^2);
ca=icdf('norm',1-signiv,0,1);
SPEkonf=theta1*((ca*sqrt(2*theta2*h0^2))/theta1+1+(theta2*h0*(h0-
    1))/theta1^2)^(1/h0);

% Projektionen av testdata på konstruerad PCA-modell
for i=1:m             %Standardiserar med värden från givna modellen
    standardtestdata(i,:)=(testdata(i,:)-medelv)./stavv;
    for j=1:n
         plottestdata(i,j)=standardtestdata(i,:)*loads(:,j);
    end
end

% SPE för observationer från testdata vid valt #PC:er
for i=1:m
    for j=1:n
 SPEantalpc(i,j+1)=standardtestdata(i,j)-
    standardtestdata(i,:)*loads(:,1:antalpc)*loads(j,1:antalpc)';
    end
    SPEantalpc(i,1)=sum(SPEantalpc(i,:).^2);
end

%% Fönster 2
for i=0:m-mask-1
    figure(1)
set(gcf,'NumberTitle','off','Name', 'Vattenfall AB Vattenkraft - PCA-
    Demo','Color',[ljusbla],'Toolbar','none')
```

```
    clf

    % Figur 1: Score-plot i form av mask
    subplot(12,2,[1 3 5 7 9 11])
    axis([0 3.5 -3 -0.5])
    set(gca,'xtick',[0:1:3.5],'ytick',[-3:1:-0.5])
    hold on
plot(plottestdata(1:i+1,1),plottestdata(1:i+1,2),'.','Color',[ljusbla]
    )
line(plottestdata(i+1:i+mask+1,1),plottestdata(i+1:i+mask+1,2),'Color'
    ,[gul])
    for j=i+1:i+mask+1
         hold on
         if j==i+mask+1
              if (SPEantalpc(j,1) > SPEkonf)
scatter(plottestdata(j,1),plottestdata(j,2),'MarkerEdgeColor','r','Mar
    kerFaceColor',[bla])
              else
scatter(plottestdata(j,1),plottestdata(j,2),'MarkerEdgeColor',[bla],'M
    arkerFaceColor',[bla])
              end
         else
              if (SPEantalpc(j,1) > SPEkonf)
scatter(plottestdata(j,1),plottestdata(j,2),'MarkerEdgeColor','r','Mar
    kerFaceColor',[gul])
              else
scatter(plottestdata(j,1),plottestdata(j,2),'MarkerEdgeColor',[gul],'M
    arkerFaceColor',[gul])
              end
         end
    end
    hold on
    title('Process-status','Fontweight','bold')
    xlabel('PC 1','Fontweight','bold')
    ylabel('PC 2','Fontweight','bold')

    %% Figur 3: SPE-residualer och 95% konfidensnivå
    subplot(12,2,[2 4 6 8 10 12])
    plot(1:i+mask+1,SPEantalpc(1:i+mask+1,1),'Color',[bla])
    hold on
    plot([1 m],[SPEkonf SPEkonf],'Color',[gul],'LineStyle','--')
text(m,SPEkonf,'95%','VerticalAlignment','Bottom','HorizontalAlignment
    ','Right')
    if max(SPEantalpc(:,1))>SPEkonf
         axis([1 m 0 ceil(max(SPEantalpc(:,1)))])
    else
         axis([1 m 0 ceil(SPEkonf)])
    end
    for j=1:i+mask+1
         if SPEantalpc(j,1)>SPEkonf          % Markera outliers
              hold on
              plot(j,SPEantalpc(j,1),'ro')
         end
    end
title(['SPE-residualer och ' num2str(100*(1-signiv)) '%
    konfidensnivå'],'Fontweight','bold')
    xlabel(' ')
    ylabel('')

    %% Figur 4: Flöde samt min-, max- och medelflöde från modell
    subplot(12,2,[16 18 20 22 24])
```

```
plot([1 m],[flowplot flowplot],'Color',[246/255 166/255
    0/255],'LineStyle','--')
    hold on
plot([1 m],[flowplotMin flowplotMin],'Color',[ljusbla],'LineStyle','--
    ')
plot([1 m],[flowplotMax flowplotMax],'Color',[ljusbla],'LineStyle','--
    ')
    plot(1:i+mask+1,testdata(1:i+mask+1,3),'Color',[bla])
    axis([1 m flowplotMin-10 flowplotMax+10])
title(['Flöde samt min-, max- och medelflöde från
    träningsdata'],'Fontweight','bold')
    xlabel(' ')
    ylabel('')

    %% Figur 5: Informationsruta
    subplot(12,2,[15 17 19])
    text(4.5,9,['SPE-residual'],'Fontweight','bold')
    line([4.475 6.425],[8.45 8.45],'Color','k')
    text(1.25,6,['Nuvärde:'],'Fontweight','bold')
text(6,6,[num2str(100*(1-signiv)) '% konfidensgräns:  '
    num2str(SPEkonf)])
    if SPEantalpc(i+mask+1,1)>SPEkonf
text(2.8,6,[num2str(SPEantalpc(i+mask+1,1))],'Color','r','Fontweight',
    'bold')
text(1.25,2,['Avvikelse från konfidensgräns:  '
    num2str(SPEantalpc(i+mask+1,1)-
    SPEkonf)],'Color','r','Fontweight','bold')
    else
text(2.8,6,[num2str(SPEantalpc(i+mask+1,1))],'Fontweight','bold')
    end
    title(' ')
    axis([1 10 1 10])
    set(gca,'XColor','w','YColor','w','xtick',[],'ytick',[])

    % Figur 6: Vattenfalls logotyp
    subplot(12,2,[21 23])
    image(logo)
    set(gca,'XColor',[ljusbla],'YColor',[ljusbla],'xtick',[],'ytick',[]
    )

% Vänta på musklick för start av mask samt innan start av
    trendsimulering
    if i==1 || ceil((m-mask+1)/2)
         waitforbuttonpress
    end
    pause(tidstopp)              % Kortare paus för
    realtidssimuleringen
end
```