UPTEC W 16023
Examensarbete 30 hp
Juni 2016

UPPSALA
UNIVERSITET

Kvalitetsgranskning av flödesdata från vattenledningsnätet

Victor Eliasson

REFERAT

Kvalitetsgranskning av flödesdata från vattenledningsnätet
Victor Eliasson

I ett försök att revidera de riktlinjer som ligger till grund för dagens dimensionering av vatten- och avloppsledningsnät, genomför Tyréns ett projekt för Svenskt Vatten Utveckling (SVU). Detta SVU-projekt syftar till att utifrån flödesdata från vattenledningsnätet kartlägga beteendemönster hos olika förbrukartypers vattenförbrukning, såsom storförbrukare och enskilda brukare, för att på så vis kunna erhålla bättre och mer aktuella underlag för dimensioneringen.

En förutsättning för SVU-studien är att det mätdata som ska analyseras håller god kvalitet. Därför kommer detta examensarbete att ingå som en understudie till SVU-projektet med syftet att finna en metod för att kvalitetsgranska flödesdata.

Detta examensarbete visade metoder för att statistiskt granska flödesdatat med avseende på outliers och anomalier. Bland annat presenterades en metod för att finna förändringar i mätseriernas statistiska fördelning. De förändringar som studerades var antingen förändring av medelvärde eller en förändring av variansen. Den senare visade sig kunna detektera sekvenser, framförallt under sommaren, där dygnsvariationen i vattenförbrukningen ofta var annorlunda mot resten av året. Vidare kunde outliers detekteras med hjälp av regressionsanalyser på mätserierna. Olika mätfel, exempelvis mätvärden som upprepas, kunde även identifieras med hjälp av skurlängdskodning.

Vid identifiering av anomalier eller outliers har vissa typer kunnat detekteras och förklarats som felaktiga värden eller icke representativa för mätserien för vidare analyser. Andra identifierade suspekta mätvärden har inte kunnat förklaras eller det har inte kunnat säkerställas att de är felaktiga. Dessa har markerats och bör studeras av en expert inom ämnesområdet.

Nyckelord: Kvalitetsgranskning, flödesdata, ledningsnät

Institutionen för informationsteknologi, Uppsala universitet
Box 337, SE-751 05 Uppsala
ISSN 1401-5765

ABSTRACT

Quality control of flow meter data from water supply networks
Victor Eliasson

In an attempt to revise the guidelines which are the basis of today's design of water and sewage systems, Tyréns carries out a project for Swedish Water & Wastewater Association (SWWA). This SWWA project aims to, based on flow meter data, map out the behavior pattern regarding water consumption for certain types of consumers, such as industrial or individual users. The overall goal is to get a better support for the dimensioning process.

A prerequisite for SWWA study is that the data to be analyzed are of good quality. Therefore this thesis will be included as a sub-study for the SWWA project with the goal to find and test methods for quality control of the flow meter data.

This thesis showed some methods to statistically validate the data. A method is for instance presented that is used to detect changes in the statistical distribution for the sets of flow meter data. The changes that were studied were either changes in mean or variance. The latter turned out to be able to detect sequences, mostly during the summer, when the daily variations in water demand is different than during the rest of the year. Furthermore, some outliers could be identified using regression analysis. Some measurement errors, such as repeated flow values, could also be detected using run length encoding.

In identifying anomalies and outliers, some types of those were possible to explain as false flow observation or non-representative for the data series as a whole for further analysis in the SWWA project. Other suspected observations were not possible to declare as true erroneous and have therefore been flagged for further investigation with support by an expert in the field.

Keywords: quality control, water mains, water supply network, flow meter data

Department of Information Technology, Uppsala University
Box 337, SE-751 05 Uppsala
ISSN 1401-5765

FÖRORD

Handledare: Hans Hammarlund, Tyréns AB, Stockholm

Ämnesgranskare: Bengt Carlsson, IT institutionen, Uppsala universitet

Examinator: Allan Rodhe, Institutionen för geovetenskaper, Luft-, vatten- och landskapslära, Uppsala universitet

Examensarbetet ingår som en understudie till Svenskt Vatten Utveckling projektet ”Studie om dimensioneringstal för vattenförbrukning” som genomförs av Tyréns.

Detta examensarbete avslutar en femårig utbildning som civilingenjör i miljö- och vattenteknik vid Uppsala universitet. Jag vill passa på att tacka Hans, Linnéa och Krister på Tyréns som jag har jobbat nära med under projektets gång. Jag vill även passa på att tacka övriga medarbetare på Tyréns för att de fått mig att känna mig välkommen vid kontoret. Vidare skulle jag vilja tacka min ämnesgranskare Bengt, för goda råd under projektets gång. Jag vill också tacka Mårten för tips under programmering samt för feedback på rapporten. Jag vill även tacka Jakob för utbyte av idéer.

Uppsala juni 2016

Victor Eliasson

Publicerad digitalt vid Institutionen för geovetenskaper, Geotryckeriet, Uppsala universitet, Uppsala, 2016.

POPULÄRVETENSKAPLIG SAMMANFATTNING

Kvalitetsgranskning av flödesdata från vattenledningsnätet

Victor Eliasson

Rent vatten är ett essentiellt livsmedel för människan. Detta ställer stora krav på både produktion och leverans av dricksvatten. Det är ett allmänt vedertaget faktum att hur människor konsumerar vatten skiljer sig både över dygnet och över året. Vid dimensionering av vattenledningsnät i Sverige idag tas hänsyn till dessa variationer i de så kallade max- och mintimfaktorerna samt max- och mindygnsfaktorerna. Dessa faktorer syftar att ta hänsyn till de tillfällen då konsumeringen av vatten är som störst. Studier tyder emellertid på att detta kan leda till överdimensionering av ledningsnätet. Studier visar även att obehövliga kapacitetsökningar av redan överdimensionerade ledningsnät görs, ofta till följd av att nya distributionsområden kopplas på befintliga ledningsnät. Ett annat problem enligt dagens standard är att det finns vissa osäkerheter kring vilka faktorer som ska nyttjas för olika typer av områden.

Med anledning av detta genomför konsultföretaget Tyréns ett Svenskt Vatten Utveckling-projekt kring den dimensionerade vattenförbrukningen, med syftet att kartlägga olika brukartypers vattenanvändning samt kontrollera dagens riktlinjer kring dimensionering av ledningsnätet.

För att eventuella mätfel inte ska påverka de analyser som görs i projektet ingår detta examensarbete som en understudie till Tyréns projekt med syftet att finna en metod för att kvalitetsgranska den flödesmätdata som finns tillgänglig samt undersöka vilka faktorer som påverkar vattenförbrukningen. Kvalitetsgranskad data kommer sedan användas för vidare studier i huvudprojektet.

Två olika typer av mätserier studerades. Dels analyserades mätserier tillhörande flödesmätare som har registrerat inflödet till ett distributionsområde, dels studerades mätserier från enskilda förbrukare. Den senare bestod i vissa fall av data från två eller flera parallellkopplade flödesmätare.

Analyserna som har gjorts i detta examensarbete visade att ett antal olika anomalier i mätserierna kunde identifieras med hjälp av statistiska analyser som upptäckte när den statistiska fördelningen i mätserien förändrades. Dessutom föreslogs metoder för att finna olika typer av mätfel, exempelvis värden som upprepas och mätningar som beror på att flera sammanhängande mätningar har summerats.

Resultaten visade att mätserierna kan kvalitetsgranskas med avseende på vissa typer av osäkerheter medan andra kräver mer platsspecifik information för att kunna säkerställas.

INNEHÅLLSFÖRTECKNING

1 INLEDNING

Vatten är en livsnödvändig resurs som behövs både för att vi ska överleva och för vi ska kunna bibehålla en god hygien. För att tillgodose Sveriges befolkning med denna resurs producerar de allmänna vattenverken årligen omkring 1 km^3 vatten, vilket motsvarar 300 liter per person och dygn. Detta ställer inte bara stora krav på vattenverken, det ställer även stora krav på distributionssystemet. Vattenförbrukningen varierar under året och är störst under sommaren, dessutom finns det en dygnsvariation med tydliga förbrukningstoppar under förmiddagen och eftermiddagen (Lidström, 2013).

Då drickvattensystem i Sverige ska dimensioneras används max- och mindygnsfaktorerna samt max- och mintimfaktorerna enligt Svenskt Vattens riktlinjer i publikationen P83 (VAV P83, 2001). Här tar timfaktorerna hänsyn till de variationer som sker över dygnet, medan dygnsfaktorerna tar hänsyn till största och minsta vattenförbrukning över året. Faktorerna läses av i ett diagram som schablonvärden och är baserade på antalet invånare som är anslutna på ledningsnätet. Faktorerna är dock endast giltiga för distributionsområden med minst 500 brukare. Vid mindre områden får istället momentanförbrukningen beräknas för dimensioneringen. Detta görs genom att beräkna hur stort normflöde samtliga tappställen i fastigheterna ger upphov till, där normflödet bestäms av Boverket och beskriver hur stort flöde en specifik typ av tappställe har (Lidström, 2013). Alla summerade normflöden omvandlas sedan till det sannolika flödet efter vilket dimensionering sker med hjälp av en tabell (VAV P83, 2001).

Tidigare studier visar att då tim- och dygnsfaktorerna väljs onödigt stora kan det leda till stora kostnader då nya områden ansluts och ledningsnätet måste dimensioneras om för att klara den ökade belastningen (Näsman Melander, 2012). En sådan åtgärd kanske inte hade behövt göras eftersom det finns indikationer på att faktorerna är för höga och behöver därför ses över (Abdu & Ullén, 2014).

Baserat på resultaten från ovan nämnda studier har Tyréns påbörjat ett Svenskt Vatten Utveckling (SVU) projekt som syftar till att undersöka hur det dimensionerade vattenflödets riktlinjer kan komma att utvecklas för framtiden (SVU projektförslag, 2015). Det här examensarbetet kommer att ingå som en understudie till SVU-projektet.

1.1 SYFTE

Examensarbetets syfte var att utveckla en metod för att kvalitetsgranska flödesdata samt testa metodiken på ett antal mätserier med ursprung från vattenledningsnätet. Ambitionen var även hitta och studera faktorer som styrde vattenförbrukningen och hur dessa faktorer påverkade mätserien.

1.2 AVGRÄNSNINGAR

Examensarbetet har skrivits parallellt med ett annat examensarbete på Tyréns med samma problembeskrivning (Ekwall, 2016). Detta innebar en uppdelning av tillgängliga mätserier. I detta projekt har mätserier från Borås och Karlstad studerats.

2 BAKGRUND

Sedan 1983 klassas dricksvatten som ett livsmedel i Sverige. Därför är det Livsmedelsverket som bestämmer vilken kvalité som ska gälla för det vatten som distribueras till brukare. Det vatten som används till dricksvatten kan komma ifrån antingen ytvatten eller grundvatten. Ofta klassas inte råvattnet som tjänligt att dricka enligt Livsmedelsverket och måste därför först renas i ett vattenverk (Lidström, 2013).

Vattentjänstlagen säkerställer att distributionen av rent vatten ordnas inom Sveriges kommuner. För att det dricksvatten som produceras i vattenverken ska nå ut till brukarna behövs ett distributionssystem som klarar av att tillgodose vattenförsörjningen i samhället. Distributionssystemet består dels av ett ledningsnät, inom vilket vattnet transporteras, dels av vattenreservoarer, tryckstegringsstationer och andra anordningar såsom avstängningsventiler, anordningar för tömning av vatten eller avluftningsanordningar. Vattenreservoarernas funktion är bland annat är att reglera för de timmar på dygnet då den största respektive minsta förbrukningen sker, så att vattenverken kan producera en konstant mängd vatten utan att ta hänsyn till förbrukningens dygnsvariationer. Tryckstegringsstationer består av pumpar för att få vattnet att gå runt i systemet där självfall på vattenflödet inte är möjligt. Ledningsnätet utformas i praktiken idag enligt två olika metoder. Den ena kallas förgreningsnät, vilken innebär att vattnet fördelas i nät så att vattenflödet endast har en riktning och ledningarna avslutas i ytterkanterna av nätet. Den andra metoden kallas cirkulationsnät, inom vilken alla ledningssträckor är sammankopplade och där vattenflödet även kan byta riktning. Rent generellt består ett vattenledningsnät av olika typer av ledningar. De största ledningarna kallas huvudledningar, vilka fördelar vattnet från vattenverket till olika områden i samhället. Mer finmaskig fördelning sker med hjälp av distributionsledningarna, vilka oftast följer gator och vägar. Från distributionsledningarna utgår så kallade servisledningar, vilka förser enskilda brukare med vatten. För att säkerställa att distributionsnätet klarar av att tillgodose samhället med tillräckligt mycket vatten krävs det att ledningsnätet inte är underdimensionerat. Eftersom det är omöjligt att vid projektering av nya distributionsområden veta exakt hur mycket vatten som kommer att förbrukas, används idag generella faktorer angående hur mycket vatten som statistiskt sett används för en typ av fastighet. Förutom att reglera för dygnsvariationer och säsongsberoende variationer med tim- och dygnsfaktorerna (se sektion 1) brukar uttrycket specifik förbrukning användas. Specifik förbrukning är ett approximativt mått på den förbrukning en viss typ av förbrukartyp antas stå för. Det är viktigt att dimensioneringen är flexibel och tillräckligt robust för att klara av olika samhällsutvecklingar. Trender tyder på att vattenförbrukningen för hushåll minskar i Sverige. En aktuell frågeställning är därför hur länge denna minskning kommer att ske (Lidström, 2013).

Flödesmätningar är ett nyttigt verktyg för att kontrollera den mängd vatten som förbrukas inom ett distributionsområde. Bland annat är det ett rättvist verktyg för vattenabonnenterna som möjliggör att de endast betalar för det vatten de har förbrukat. Det finns även miljömässiga fördelar med flödesmätningar, eftersom vatten är en naturtillgång och alltför stora uttag av vatten kan rubba den ekologiska balansen. Genom att kontrollera vattenförbrukningen kan således insatser göras för att förhindra att ekosystemet påverkas alltför negativt. Dessutom kan flödesmätningar användas som ett verktyg för att upptäcka storleken på läckage i ledningsnätet. Sammanfattningsvis kan det sägas att flödesmätningar i ledningsnätet oftast görs av ekonomiska

skäl, men genom att ha god kontroll på förbrukningen erhålls också en bättre kontroll av den miljöpåverkan ett vattenuttag innebär (VAV P100, 2009).

Tyréns har startat ett Svenskt Vatten Utveckling (SVU) projekt med syftet att förnya de riktlinjer angående dimensioneringen av vattenledningssystem som finns angiva i publikationen Svenskt vatten P83 (VAV P83, 2001). I SVU-projektets projektplan finns ett antal moment definierade. Dessa är:

- Utredning av riktlinjer av dimensionerande faktorer.
- Utreda den specifika vattenförbrukningen, det vill säga den förbrukning som olika förbrukartyper står för.
- Utreda påverkan av trädgårdsbevattning och om det är möjligt att producera statistik angående hur stor påverkan detta har. I dagens läge finns trädgårdsbevattnings effekt inbakad i maxdygnsfaktorn (VAV P83, 2001).
- Undersöka huruvida den idag använda metoden med momentanförbrukning som används för områden med mindre än 500 brukare kan standardiseras till en liknande modell som den för större områden med tim- och dygnsfaktorer.
- Klargöra nyttan med regelbundna flödesmätningar med syftet att få fler kommuner att utnyttja detta.
- Kvalitetsgranskning av mätdata som ska användas vid analyser.

Tidigare examensarbeten, ej kopplade till detta SVU-projekt, som har gjorts på Tyréns kring ämnet vattenförbrukning visar att det finns ett behov av att tydliggöra hur tim- och dygnsfaktorerna ska väljas för olika typer av områden (Näsman Melander, 2012). Då det visade sig att detta kan vara aktuellt att studera mer i detalj gjordes ytterligare ett examensarbete. Där gjordes ett försök att från mätområden hos Norrvatten specificera hur olika förbrukartyper påverkade den specifika förbrukningen och hur då tim- och dygnsfaktorerna kunde behöva förändras för att tillfredsställa det behov som de ämnade lösa (SVU projektförslag, 2015). Det visade sig emellertid att det mätdata som användes innehöll flera extremvärden som inte kunde förklaras. Då den studien inte ämnade studera enskilda observationers riktighet togs dessa således bort från mätserierna då de starkt påverkade resultatet. Extremvärden är dock av stort intresse vid analyser av den maximala förbrukningen, ifall de är korrekta och inte orsakade av mätfel. För att komma till bukt med detta problem identifierades därför behovet av att först kvalitetsgranska det mätdata som ska användas till analyser kring vattenförbrukningen i SVU-projektet. Syftet med denna rapport är därför att föreslå en metod kring kvalitetsgranskningen av mätdatat samt att testa den på det data som tillgängliggjorts. Eftersom flera av de analyser som görs i kvalitetsgranskningen även kan vara intressanta för andra moment i SVU-projektet har även detta tagits i beaktande. Exempel på detta är hur den specifika förbrukningen skiljer sig för olika förbrukartyper samt hur sommarmånaderna påverkar det totala flödet inom aktuellt mätområde.

2.1 KVALITETSGRANSKNING

I SVU-projektets syfte ingår att kartlägga hur de dimensionerade faktorerna kan förändras. Detta görs genom att analysera ett antal mätserier kopplade till olika distributionsområden. Särskilt intressant för studien är lokala och globala extremvärden i mätserierna. Dessutom finns ambitionen att identifiera olika gruppers vattenanvändning. Därför är det av största vikt att det går att lita på alla observationers tillförlitlighet, då icke-identifierade felaktiga mätningar kan påverka slutresultatet av olika analyser avsevärt. Det här examensarbetet syftar till detta genom att identifiera förmodade felaktigheter och antingen ta bort dessa från mätserien, om det finns starka bevis för att mätvärdet är oanvändbart, eller markera mätvärdet som felaktigt. Informationen om de observationer som förmodas vara felaktiga finns då sparade och en expert kan därefter granska dessa data och ta ett slutgiltigt beslut om dess riktighet (Tiberg m.fl., 2014).

En kvalitetsgranskning definieras som en aktivitet som syftar till att försäkra om att data håller tillräckligt hög kvalitet. Det mer processinriktade arbetet, kvalitetssäkring, syftar till att säkerställa att det data som kontinuerligt produceras håller hög kvalité. Då detta projekt endast behandlar redan producerad data används begreppet kvalitetsgranskning (IPCC, 2000).

3 METOD

Metodiken för hur hela kvalitetsgranskningen i projektet genomfördes beskrivs med hjälp av ett flödesschema (Figur 1). Indatat rensades i ett första steg från summerade mätvärden (se sektion 3.2.2) i Excel. Därefter importerades datasetet till programpaketet R (sektion 0) där identifiering av upprepade värden gjordes med hjälp av skurlängdskodning (sektion 3.3) och med hjälp av statistiska analyser detekterades sekvenser där medelvärdet förändrades (sektion 3.5.4). För att kunna identifiera outliers (sektion 3.5.1) samt initiera metoder som detekterade förändringar i variansen (sektion 3.5.4) behövde tidserierna rensas från periodicitet och trend (sektion 3.4.2). Detta steg har i denna rapport benämnts som *säsongsrensning*. För att kunna ta bort periodiciteten från mätserierna gjordes först en spektralanalys med syftet att studera vilka frekvenser som dominerade (sektion 3.4.1). För vissa mätserier bestående av flera parallellkopplade flödesmätare identifierades outliers med hjälp av regressionsanalyser (Sektion 3.5.2). Alla resultat jämfördes med vilken säsong observationerna tillhörde samt data beträffande rörbrott. Slutligen gjordes en kvalitetsklassning vilket innebar att alla observationer fick en klassning som beskrev hur pålitlig varje observationen var. Kvalitetsgranskad data med felaktiga mätvärden borttagna exporterades som utdata tillsammans med tillhörande kvalitetsklassning.

Figur 1. Metodiken för hur kvalitetsgranskningen har genomförts och i vilka steg anomalier har identifierats

3.1 DATORPROGRAM

I detta projekt användes två olika datorprogram. För den första bearbetningen och sorteringen av rådatat användes Microsoft Excel. Samtliga statistiska analyser och tester gjordes i statistikprogrammet R.

3.1.1 Microsoft Excel

Den första sorteringen och bearbetningen av all rådata har gjorts i Microsoft Excel (utgåva 2013), då det erbjuder en överskådlig layout till tidsserierna som använts i projektet. För att på ett enkelt sätt kunna lagra information angående vilken typ av säsong en observation tillhörde skapades dessutom med hjälp av Excels inbyggda formelverktyg ett antal dummyvektorer, som innehåller binär information för varje observation. Bearbetad rådata och konstruerade vektorer har därefter exporterats till R för vidare analyser.

3.1.2 Programpaketet R

Alla statistiska tester och analyser av tidsserierna har gjorts i R (R Core Team, 2015), vilket är ett gratis open source-program med programmeringsspråket R som grundar sig i programmeringsspråket S (Morandat, 2012). R är vida använt för olika statistiska analyser, varför det också valdes i detta projekt som det primära verktyget (Muenchen, 2015)

I R används förutom ett antal inbyggda funktioner även så kallade paket. Dessa paket har ofta mer specifika användningsområden och laddas enkelt ner till programmet (R Core Team, 2015).

3.2 DATA

3.2.1 Datainsamling

De mätdata som har använts för kvalitetsgranskningen i detta projekt kommer från Borås kommun och Karlstad kommun. De flesta mätserier har givits som univariata tidsserier, det vill säga det finns endast en variabel i tiden, och den sträcker sig för de flesta mätserier mellan början av 2013 till hösten 2015. Vattenflödet har i regel givits per timme men på ett fåtal av tidsserierna finns även data per minut och/eller per sex minuter (Tabell 1). I vissa fall saknades information om antalet brukare tillhörande ett mätområde.

Mätdata från flerbostadshus i Borås härstammar i många fall från två mätare. I dessa fall var mätarna parallellkopplade i vattenmätarbygeln. Det innebär att summan av mätarnas värde vid varje tidpunkt utgjorde den totala vattenförbrukningen.

Tabell 1. Jämförelse över de olika mätserier som har varit tillgängliga under projektet. Mätserierna för Henstad-Hultsberg samt Alster härstammar från Karlstad. Övriga mätserier kommer från Borås.

Mätområde	Brukare	Mätare	Tidsupplösning	Inflöde till zon	Enskild brukare
Sjöbo	6490	1	1 min, 6 min	X	
Tårpilsgatan 7	14	1	1 h		X
Vejlegatan 6	52	2	1 h		X
Moldegatan2	74	2	1 h		X
Moldegatan 6	64	2	1 h		X
Vejlegatan 11	40	1	1 h		X
Vejlegatan 5	54	1	1 h		X
Vejlegatan 7	46	1	1 h		X
Vejlegatan 9	41	1	1 h		X
Islandsgatan	-	1	1 h		X
Vitingsgatan 1-13	549	2	1 h		X
Billdalsgatan 2-8	143	2	1 h		X
Tosseryd	189	1	6 min	X	
Tyllgatan	1652	1	6 min	X	
Mårdgatan	475	1	6 min	X	
Klintessväng 10	12	2	1 h		X
Hällegatan 22	-	2	1 h		X
Hällegatan 24	-	2	1 h		X
Hällegatan 26	-	3	1 h		X
Henstad-Hultberg	2036	1	1 min, 6 min, 1 h	X	
Alster	498	1	6 min, 1 h	X	

3.2.2 Databearbetning

Det första steget i projektet var att bearbeta rådata. I vissa fall saknades timvärden vid några observationer, så tidsvektorerna kompletterades med de tidpunkter som saknades. Vidare fanns på några av de större distributionszonerna data angående rörbrott. Tillfället för dessa rörbrott jämfördes med berörda observationer i mätserien.

På vissa av tidsserierna identifierades perioder där vattenflödet registrerats som noll i ett antal observationer för att sedan stiga till en, relativt serien som helhet, kraftig topp. Törneke[1] menade att detta kunde förklaras med en egenskap hos flödesmätaren att då en registrering av flödet för en eller flera observationer saknades så sparades istället värdet av dessa i minnet. När registrering åter var möjlig summerades alla missade observationers värden samt den nya i en enstaka observation. Detta gav tidsserien utseendet av att ha flera lokala extremvärden, vilket egentligen inte var fallet. Eftersom dessa summerade värden endast var av intresse då totalförbrukningen studerades och samtidigt gav missvisande egenskaper för tidsserien, har dessa värden flaggats som felaktiga och raderats som ett steg i valideringsprocessen. Identifieringen av dessa har skett automatiskt i Excel genom att sätta villkoren att om en eller flera observationer saknas och den första observationen efter saknade värden är större än ett ansatt tröskelvärde, så flaggades observationen.

3.3 SKURLÄNGDSKODNING

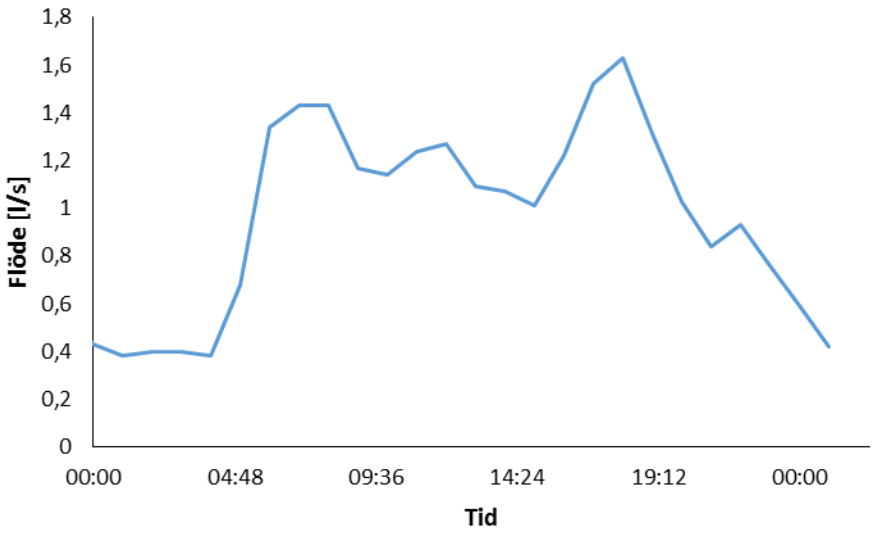
I projektet noterades att det fanns sekvenser i tidsserierna i vilka värden upprepades. Orsaken till att varför värdena upprepades var okänd, men värdena ansågs inte vara representativa för mätserien och antogs bero på mätfel. Därför identifierades behovet att märka ut de sekvenser som har detta fenomen och detta gjordes med en skurlängdskodning. Skurlängdskodning är en komprimeringsmetod med målet att finna sekvenser där värden upprepas. Kompressionen går till enligt följande. Antag en dataföljd enligt 11;20;13;11;5;5;5;62;7;7;8. Skurlängdskodningen i R komprimerar då denna sekvens till 1;1;1;1;3;1;2;1. Siffran i den komprimerade sekvensen anger hur många gånger värdet återkommer i följd. Här beskriver den komprimerade sekvensen att först kommer fyra stycken observationer som ej upprepas följt av 3 stycken upprepade observationer och så vidare (Smith, 2001).

I R gjordes skurlängdskodningen med hjälp av paketet *data.table* (Dowle m.fl., 2015).

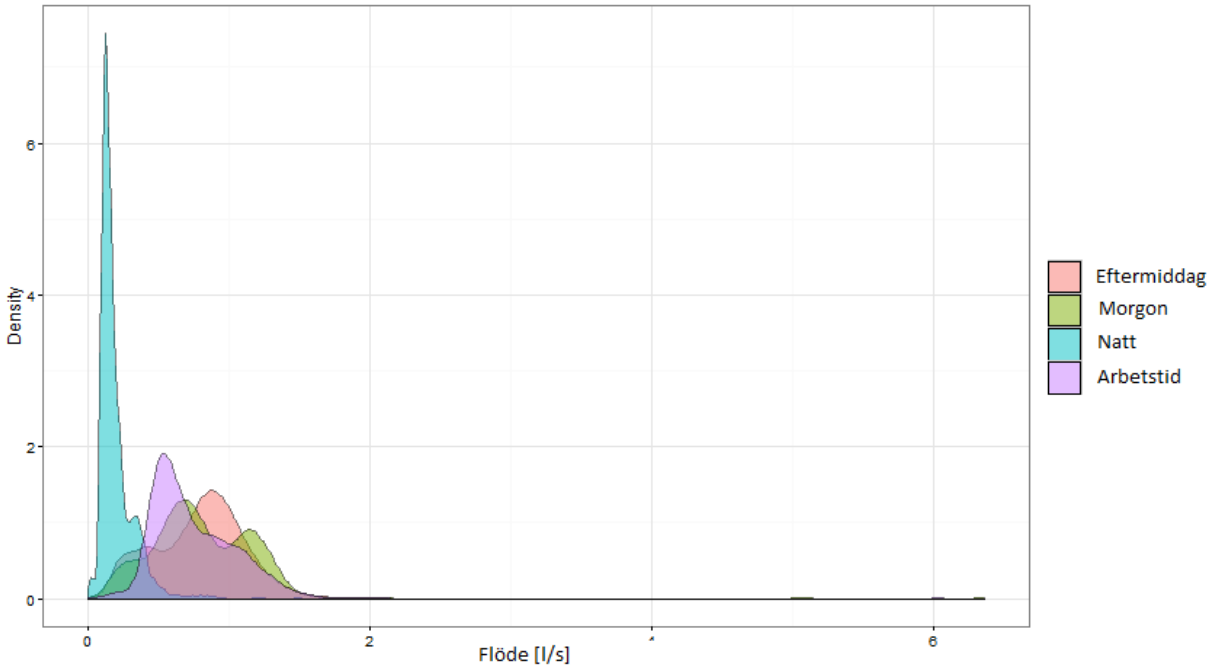
3.4 PERIODICITET

Vattenförbrukningen inom ett distributionsområde har en tydlig dygnsvariation. Från given mätdata gick det att se en tydlig topp i förbrukningen på förmiddagar och eftermiddagar (Figur 2). Detta sammanhänger med de tider då folk går till jobbet respektive när de kommer hem. På nätterna återfinns dygnets minimivärden (Lidström, 2013). De nattliga flödena har dessutom en betydligt mindre variation än de under övriga dygnet (Figur 3).

[1] Krister Törneke, VA-utredare, Tyréns, möte den 17 december 2015

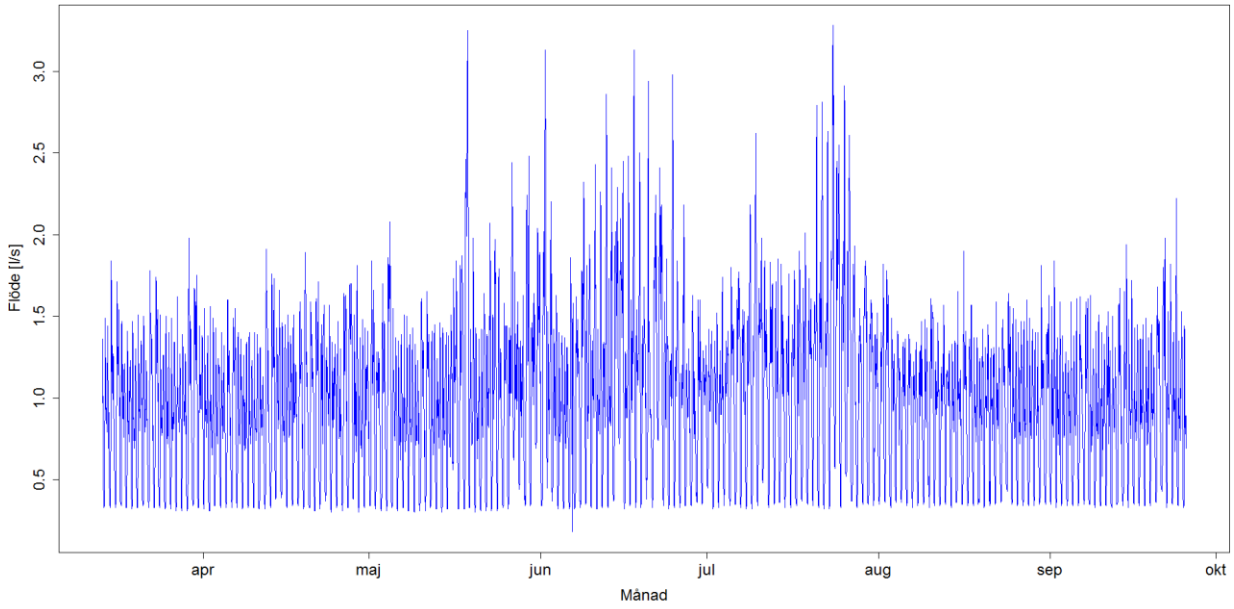


Figur 2. Vattenförbrukningen för Alster 2015-03-15. Exempel på hur flödet varierar över dygnet.



Figur 3 – Variationen hos vattenförbrukningen för Mårdgatan för olika delar av dygnet.

I vattenförbrukningen fanns det dessutom en variation under året (Figur 4). Detta kan vara i form av ökad förbrukning på sommaren eller vid vissa storhelger (Lidström, 2013).



Figur 4. Flöde för Alster april-oktober 2014. Dygnsvariationerna ökade under sommarmånaderna.

3.4.1 Spektralanalys

I mätserierna fanns en periodicitet både i form av dygnsvariationer och variationer över tid, såsom en veckovis och en kvartalsvis period. Med syftet att undersöka vilken periodicitet dataseten hade gjordes en spektralanalys. Spektralanalysen beräknades med den diskreta fouriertransformen på data och med en Daniell utjämnare, vilken är en typ av flytande medelvärde som förenklar utläsningen av dominerande frekvenstoppar. Detta ger ett utjämnat spektrum för data (Metcalfe och Cowpertwait, 2009).

3.4.2 Säsongsrensning

För att bättre kunna identifiera outliers som eventuellt låg dolda i seriens säsonger samt anomalier kopplade till förändringar variansen hos tidsserierna delades mätserien upp i komponenterna trend, säsong samt ”kvarvarande brus” (ekvation 1) (Brockwell och Davis, 1991).

$$X_t = m_t + s_t + Y_t \qquad (1)$$

där X_t betecknar originalserien, m_t är trendkomponenten, s_t är säsongskomponenten och Y_t är den kvarvarande delen när trend och säsong har subtraherats från mätserien.

För att genomföra säsongsrensning krävdes att en modell av tillgänglig data först skapades, vilket gjordes med tre olika metoder. Dessa var säsongsrensning med hjälp av ett flytande medelvärde, säsongsrensning med hjälp av lokal regression, samt en avancerad metod för säsongsrensning, med syfte att rensa mätserien med avseende på flera olika perioder, som baserades på exponentiellt utjämnade tillståndsmodeller.

Med kommandet *STL* i R (R Core Team, 2015) delades mätserien in i komponenterna trend, säsong samt kvarvarande brus, då trend och säsong subtraherats från datasetet. Detta gjordes med hjälp av utslätning med lokal regression, LOESS. Här passades ett polynom av låg ordning på varje observation i en delmängd av dataserien. Passningen skedde genom att använda viktade minstakvadratmetoden vilket innebär att punkter som låg nära den observation som skulle estimeras gavs en större vikt medan punkter längre ifrån fick en lägre vikt. Detta gjorde att modellen kunde göras robust mot eventuella outliers. Funktionen STL består av två loopar, en inre och yttre. I den inre loopen uppdateras säsong- och trendkomponenten med hjälp av LOESS-utjämning. I den yttre loopen beräknas viktningen, för att minska abnormt beteende hos de olika komponenterna och därmed ge ett mer robust resultat. Efter att den yttre loopen beräknats görs slutligen ytterligare en iteration på den inre loopen, varvid en skattad trend och säsong erhålls (Cleveland m.fl., 1990).

Säsongsrensning med flytande medelvärde är den klassiska metoden för säsongsrensning. Den sker genom att ett flytande medelvärde ansätts för att beräkna trendkomponenten i dataserien. Därefter beräknas säsongen genom att ta medelvärdet för varje period i serien. Säsongskomponenten byggs sedan upp genom att sammansätta alla perioder i en vektor. När den estimerade säsongs- och trendkomponenten sedan subtraheras från mätserien fås den kvarvarande komponenten. Denna typ av säsongsrensning antar att periodiciteten inte förändras över tiden. Om detta inte är fallet kan dessa förändringar i periodicitet över tiden inte fångas. En annan svaghet är att metoden inte är särskilt robust mot värden som inte följer normalfallet inom mätserien. Dock ansågs det i det här projektet att säsongsvariationerna inte förändrades särskilt mycket med tiden samt att det kunde vara intressant att jämföra den här metoden med andra tillgängliga metoder, framförallt då beräkningstiden är betydligt snabbare än för andra mer avancerade algoritmer (Kendall m.fl., 1983).

Eftersom tidserierna verkade ha en komplex periodicitet med flera olika säsonger, gjordes även en säsongsrensning med hjälp av funktionen TBATS i R som finns tillgänglig i paketet *forecast* (Hyndman, 2015). Algoritmen tillåter användaren att ange flera olika säsonger för dataserien. TBATS är en förenklad variant av funktionen BATS som använder sig av Holt-Winters modeller för exponentiell utjämning. Detta görs genom att ansätta repeterade lågpassfilter som filtrerar bort högfrekvent brus från datasetet, viktningen för estimeringen av säsongerna ökar därigenom med en exponentiell faktor. Närliggande punkter får en låg viktning som sedan ökar med avståndet. En låg viktning innebär att det estimerade värdet för modellen är snarlikt originalvärdet. Om det finns en trend i data används dubbelexponentiell utjämning och om det dessutom finns förändringar i säsongen används trippelexponentiell utjämning (Hyndman, 2013). I TBATS byts säsongskomponenten som används i BATS mot en trigonometrisk formulering av säsongen som baseras på en fouriertransformering. En fördel med denna metod gentemot BATS är att estimeringsprocessen förenklas (De Livera, 2011).

En enkel typ av säsongsrensning gjordes även. Här differentieras dataserien med avseende på föregående dygns värde (ekvation 2). Fördelen med denna metod är att ingen hänsyn behöver tas till någon modellanpassning (Killick och Eckley, 2014)

$$Z_t = \Delta^{24} y_t \qquad (2)$$

Eftersom differentieringen beror på det 24:e föregående värdet innebär det att det nya datasetet kommer vara 24 observationer kortare än originalserien, vilket måste tas i beaktande.

3.5 STATISTISKA ANALYSER

Ett antal statistiska analyser gjordes även med syfte att finna outliers samt för att finna sekvenser i mätserierna där den statistiska fördelningen förändras. Ett första steg för dessa analyser var att undersöka om tillgänglig data följde någon fördelning. Om så var fallet hade särskilda villkor för denna fördelning kunnat utnyttjas. Då det kunde ifrågasättas om behandlade tidsserier i detta projekt följde normalfördelningen söktes istället metoder för kvalitetsgranskningen som inte tog hänsyn till statistisk fördelning.

3.5.1 Outliers

Om en observation påträffas som ligger långt ifrån andra observationer eller det förväntade värdet kallas denna observation för en outlier. Om en sådan observation påträffas måste den analyseras för att avgöra om det verkligen är ett inkorrekt värde. Om detta kan bekräftas statistiskt eller fysikaliskt kan denna observation korrigeras då den påverkar hela seriens egenskaper. Om det inte kan bekräftas att det är en inkorrekt observation finns möjligheten att den faktiskt tillhör en riktig observation trots oregelbundna egenskaper. Det kan då vara riskabelt att rensa den då det bidrar till ett partiskt resultat (Grubbs, 1969).

Olika typer av anomalier har studerats. I detta projekt benämns dessa som outliers och övriga anomalier. Med outliers menas additiva outliers, vilka kan förklaras som en puls i mätserien med ett plötsligt onormalt stort (eller litet) värde för att i efterföljande observation åter vara normalt igen (Zainol m.fl., 2010). De övriga anomalier som studerades var abrupta förändringar i mätseriens medelvärde, varians (se sektion 3.5.4) eller observationer där mätvärden upprepas (se sektion 3.3). Dessutom studerades anomalier härrörande mätfel eller flödesmätarnas egenskaper (se sektion 3.2.2)

Identifieringen av additiva outliers för de mätserierna med endast en variabel i tiden gick ut på att göra en residualanalys. Då residualerna från den trend- och säsongsrensade serien blev för stora så betraktades motsvarande observationer som en outlier (Acevedo, 2012). Det tröskelvärde för de residualanalyser som användes i detta projekt var tre standardavvikelser från medelvärdet på säsongsrensad data (NIST/SEMATECH, 2013). Markerade outliers jämfördes sedan med tillgänglig information från den berörda tidsserien. Denna information var i form av observationens säsongstillhörighet samt data över då rörbrott inträffat. Hypotesen var att extrempunkter skulle uppstå under dessa tidpunkter.

För mätserier härrörande från enskilda brukare med parallellkopplade mätare kunde additiva outliers identifieras med hjälp av regressionsanalys.

3.5.2 Regressionsanalys

För mätserier bestående av data från parallellkopplade mätare gjordes en linjär regressionsanalys. Hypotesen var att mätserierna skulle visa samma värde eftersom flödesmätarna mätte i samma mätpunkt. Analysen gjordes i R med genom att jämföra den ena mätaren med den andra och anpassa en linjär trendmodell till dataserierna. Från denna linjära modell går det sedan att göra en enkel regressionsanalys (Acevedo, 2012).

I regressionsanalysen identifierades observationer som har höga residualer eller högt inflytande på skattningen av den linjära modellen. Då dessa blir för stora kan de betraktas som outliers. Detta gjordes med hjälp av Cook's avstånd (Acevedo, 2012). När sambandet mellan högt inflytande och höga residualer blev för stort betraktades den observationen som en outlier. Enligt Acevedo (2012) kan en observation betraktas som en outlier när Cook's avstånd är större än 1.

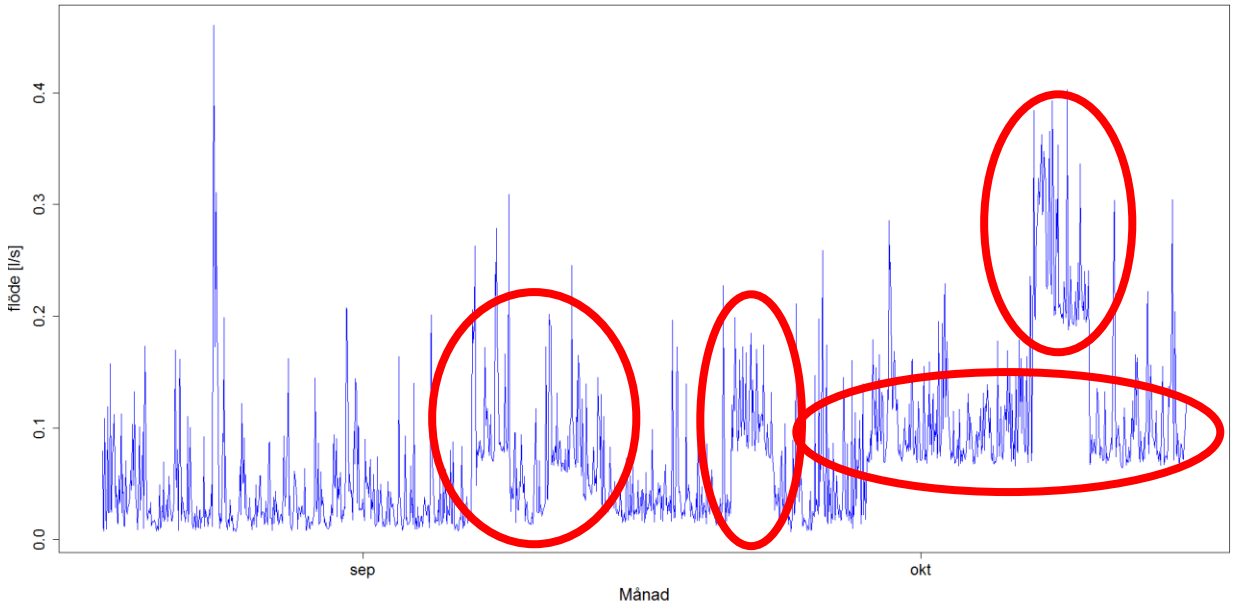
3.5.3 ARIMA

ARIMA är ett statistiskt verktyg, vida använt för sin förmåga att prognostisera tidsserier. Det kan också användas som ett verktyg för att finna anomalier och outliers i tidsserierna (De Nadai och van Someren, 2015). I detta projekt var prognostisering inte av intresse men metoder för att finna anomalier med hjälp av modeller gjorda med ARIMA testades. ARIMA är en vidareutveckling av klassiska ARMA-modeller. En ARMA-modell består av en autoregressiv process (AR), som beskriver ett linjärt samband mellan en observation och föregående observation, samt en process av flytande medelvärde (MA). ARMA modelleringen kräver att tidsserien är stationär, vilket innebär att periodicitet och trend saknas. För icke-stationär mätdata kan istället ARIMA användas, detta görs genom att till ARMA processen dessutom tillföra en integrerade del som anger antalet gånger en differentiering behöver göras för att erhålla stationäritet (Gustavsson, 2009).

I paketet *tsoutliers* i R finns en funktion som automatiskt identifierar olika typer av outliers baserat på en ARIMA-modell som väljs automatiskt i programmet (López-de-Lacalle, 2015). För det mätdata som användes i detta projekt visade det sig emellertid att denna metod inte fungerade då beräkningstiderna var extrema. Huruvida de långa beräkningstiderna berodde på att mätserierna var väldigt långa eller att ARIMA-modellerna var felaktigt formulerade i programmet undersöktes inte vidare eftersom parametervalet till formuleringen av ARIMA-modellerna visade sig vara svåra att genomföra. Dessa parameterval krävde viss erfarenhet av att tolka data samt antagande av särskilda villkor. Sammanfattningsvis kan sägas att då trovärdigheten för de ARIMA-modeller som lyckades skapas i detta projekt inte ansågs kunna motiveras i kombination med de långa beräkningstiderna, valdes andra verktyg till detektionen av outliers och anomalier.

3.5.4 Detektion av förändringar

I projektet identifierades behovet av att hitta sekvenser i dataseten där variansen eller medelvärdet plötsligt förändrades. I Figur 5 ses ett antal exempel på sekvenser från slutet av september där förändringar i medelvärdet sker. Detektion av abrupta förändringar (eng. Change detection) är en metod för statistiska analyser med syftet att finna observationer i en tidsserie där seriens statistiska egenskaper förändras. En sådan förändringspunkt (eng. change point) som delar upp en serie observationer i två segment, ett segment innan förändringen sker och en efter. Inom varje segment har distributionen liknande egenskaper. Målet med detektion av abrupta förändringar är att hitta den tidpunkt då dataseriens egenskaper plötsligt förändras (Ross, 2015).



Figur 5. Flödesdata från Tårpilsgatan 2015. Röda markeringar illustrerar exempel där en förändring i medelvärdet sker.

Nollhypotesen för detektion av abrupta förändringar är att det inte finns någon förändringspunkt i dataserien med den alternativa hypotesen att det finns en förändringspunkt. En enligt litteraturstudien vanlig metod att detektera abrupta förändringar ges av:

$$\sum_{i=1}^{m+1}\left[C\left(y_{\tau_{i-1}+1):\tau_i}\right)\right] + \beta f(m) \qquad (3)$$

Här ges datat av $y_{1:n}=(y_1,\ldots,y_n)$ och antalet förändringspunkter är *m*. Dessa återfinns i observationerna $\tau_{1:m} = (\tau_1,\ldots\ \tau_m)$. Av detta följer att antalet segment datalängden delas in i är m+1, det vill säga att varje förändringspunkt delar datat i två delar. $C(y_{\tau(i-1)+1):\tau i})$ är en kostnadsfunktion för den sekvens i mätserien som sträcker sig mellan förändringspunkterna τ_{i-1} och τ_i. Kostnadsfunktionen ska minimeras och för den observation i sekvensen där uttrycket blir som störst finns indikationer på att det sker en förändring. βf(m) är en straffunktion med tröskelvärdet β. Straffunktionen beror av vilken typ av statistisk metod som används och denna har syftet att skydda resultat från att överdriva små förändringar i datat så att endast förändringspunkter större än ett ansatt tröskelvärde detekteras (Killick m.fl 2012).

De straffsatser som användes i detta projekt var den linjära straffsatsen Akaikes informations kriterium, AIC, som beskrivs av $\beta = 2p$, där p är antalet parametrar som tillkommer vid förekomsten av en förändringspunkt. Genom att minimera AIC fås de förändringspunkter som var mest sannolika enligt detta kritrium. Därtill testades algoritmen CROPS där istället straffvärden väljs inom ett intervall sådant att $\beta \in [\beta_{min}, \beta_{max}]$ (Haynes m fl, 2014)

Två olika förändringar av tidseriernas statistiska egenskaper har studerats. Dessa är förändringar i medelvärdet samt förändringar i variansen. För analysering av förändringar i varians för periodiska tidsserier behövdes som förberedande databearbetning även en säsongsrensning göras, eftersom metoden kräver att ingen periodicitet förekommer i datasetet (Killick och Eckley, 2014).

I detta projekt användes och testades algoritmer från två olika R paket för att undersöka och hitta abrupta förändringar. I paketet *changepoint* (Killick m fl, 2015) testades binär segmentering, segment neighborhood samt PELT (Killick m fl, 2012).

Binär segmentering är en approximativ metod som söker igenom datasetet efter en enstaka förändringspunkter. Hittas en sådan punkt delas datasetet in i två nya segment, en före och en efter förändringen. Därefter görs en ny sökning efter en abrupt förändring i de två nya segmenten och så vidare till dess att ingen ny förändring hittas (Scott och Knott, 1974).

Segment neighborhood är en exakt metod som delar upp serien i ett bestämt antal segment. Först beräknas alla möjliga segment med dynamisk programmering, vilket innebär att beräkningen av de optimala segmenten beräknas genom att ett antal delfunktioner definieras. Lösningen till varje delfunktion fås genom att utnyttja lösningarna av föregående delfunktioner (Auger m.fl., 1989).

PELT är en beräkningsmetod som även den ger en exakt lösning. Förutom att PELT liksom segment neighborhood utnyttjar dynamisk programmering använder den sig av beskärning. Detta innebär att det träd av delfunktioner som används i den dynamiska programmeringen beskärs så att delfunktioner som inte innehåller tillräckligt relevant information tas bort från beräkningarna. Detta innebär att beräkningstiden kan förkortas avsevärt. Detta gäller emellertid endast då antalet förändringspunkter ökar linjärt med datasetets längd. I annat fall kommer inte optimal beskärning kunna göras (Killick m.fl., 2012).

Med paketet *cpm* (Ross, 2015) detekteras abrupta förändringar med hjälp av olika statistiska metoder. De olika statistiska metoder som går att välja bedömer vilken typ av förändring som sker i en observation samt för vilken typ av fördelning den är anpassad för. De statistiska metoder som användes i det här projektet för algoritmen i *cpm* var Mann-Whitney som letar efter förändringar i medelvärde samt Mood som letar efter förändringar i variansen. Dessa är icke-parametriska test vilket innebär att den statistiska fördelningen inte behöver vara känd. Eftersom det gick att ifrågasätta normalfördelning på de dataserier som fanns tillgängliga i det här projektet undveks statistiska metoder baserade på normalfördelning (Ross, 2015).

Nollhypotesen för Mann-Whitney och Mood är att varje punkt kommer från samma statistiska fördelning. Testerna tilldelar varje punkt i mätserien ett statistikt värde, sedan mäts till vilken grad varje punkt avviker från det förväntade värdet. En observation markeras som förändringspunkt i algoritmen om den observationen får det största värdet i det statistiska testet samt om detta statistiska värde är större än tröskelvärdet. Detta tröskelvärde kan väljas att beräknas automatiskt i algoritmen. Detta görs genom att ansätta α så att false positive probability (FPP), $P(N_{max}>h_t) = \alpha$, gäller för alla t. Där h_t ger tröskelvärdet. (Ross och Adams, 2012)

3.5.5 Rekursiva filter

Som ett verktyg för att finna anomalier utnyttjades metoder för att filtrera mätserierna. Genom att lägga på ett linjärt rekursivt filter på mätseriernas rådata erhölls en vektor med filtrerad utdata. Rekursiva filter är vanligt förekommande inom signalbehandling och det filtret gör är att den återanvänder föregående utdata som indata. Filtret kan beskrivas med ekvation 4, där f betecknar filtret. Den nya observationen x_i beror av den tidigare filtrerade utdatat y_{i-1} (Smith, 2001).

$$x_i = \sum_{m=1}^{n1}(g_m x_{i-m}) + \sum_{m=1}^{n2}(f_m y_{i-m}) \quad (4)$$

Då responsen av det linjära filtret studerades med hjälp av detektion av förändringspunkter kunde anomalier hittas.

4 RESULTAT

I detta avsnitt presenteras resultaten från de analyser som gjorts i projektet. Kontrollen har gått till enligt följande, ifall ett misstänkt värde har påträffats har det flaggats, det vill säga observationen har markerats med avseende på vilket typ av fel som har identifierats. Alla genomarbetade observationer har tilldelats en av tre klasser. Dessa är:

0. Inga anmärkningar har hittats för observationen.
1. Observationen bör användas med försiktighet då den markerats som suspekt.
2. Observationen har markerats som felaktig och bör inte användas i vidare analyser.

4.1 MÄTFEL

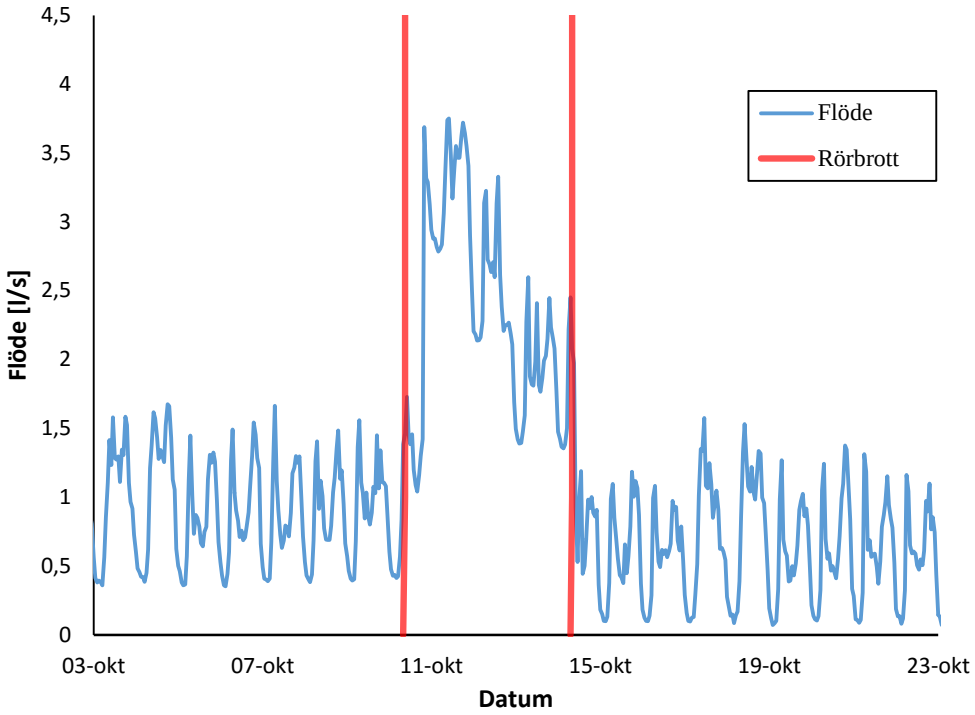
Det första steget i kvalitetsgranskningen var att identifiera uppenbara felaktigheter i mätserierna. Detta handlade främst om mätvärden som saknades helt för vissa tidpunkter (sektion 4.1.1), flera mätvärden som summerats i en och samma observation (sektion 4.1.3) samt jämförelse av rådata med data på tidpunkter då rörbrott inträffat (sektion 4.1.2).

4.1.1 Saknade mätvärden

I valideringsanalysen identifierades och flaggades observationer inom tidsserien som saknades eller var strikt noll värda. Dessa härstammar från tillfällen då flödesmätaren av olika anledningar misslyckats med att registrera värdet i en observation. Då en observation är 0 kan det innebära att mätningen understigit flödesmätarens inställning för *cut off* flöde, det vill säga att flödet är negativt eller så lågt att mätaren inte registrerar det. Saknade värden markerades som felaktiga av den enkla anledningen att det inte skulle skapas luckor i tidsvektorn, medan 0-värden markerades som suspekta då det inte ansågs rimligt att det faktiska flödet var exakt noll. Värt att notera är att för de mindre områdena med endast ett fåtal brukare är det ingen omöjlighet att det inte sker någon förbrukning överhuvudtaget under vissa tider på dygnet. Då det fanns en osäkerhet i om det faktiska flödet var noll eller om det är mätvärden som understigit *cut off* valdes dessa ändock att markeras som suspekta.

4.1.2 Rörbrott

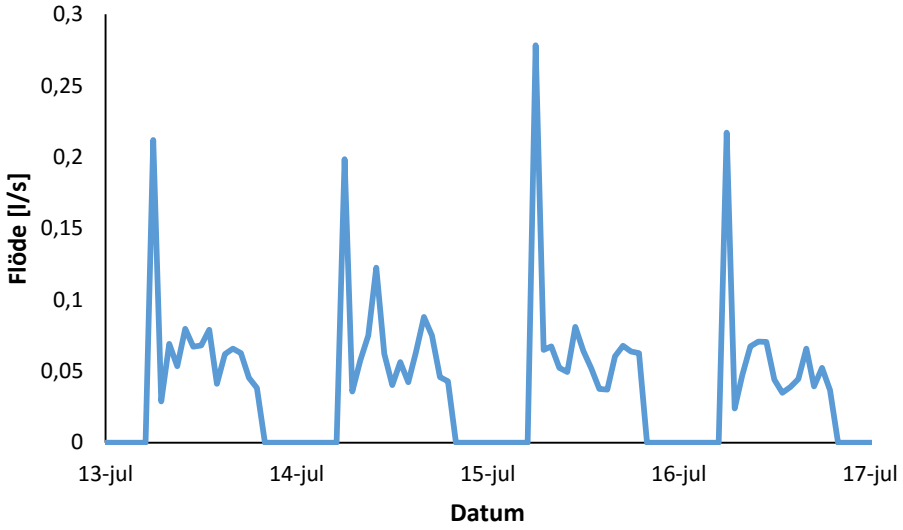
För några av de studerade dataseten har även information tillhandahållits angående datum när rörbrott har upptäckts. Vid en okulär kontroll över serierna vid dessa datum (Figur 6) kan det i vissa fall tydligt ses ett oregelbundet beteende. Det har därför med avseende på syftet att kvalitetsgranska data antagits att de observationer som härrör till datumen för rörbrott anses som opålitliga och de har därmed flaggats som varningar. Noterbart är att det inte finns någon information tillgänglig huruvida en inrapportering av rörbrott avser ett upptäckt rörbrott eller ett lagat rörbrott. Från figuren kunde det antas att den första inrapporteringen avser uppkomsten av rörbrottet och den nästkommande inrapporteringen avser åtgärdandet av detta. Men till detta kunde ingen hänsyn tas till följd av ovan nämnd orsak. Vidare har det inte bifogats någon ytterligare information, angående exakta tidpunkten när rörbrotten upptäckts eller när dessa är åtgärdade. I och med detta har alla observationer under aktuella dygn, enligt tillgänglig data på rörbrott, flaggats som suspekta observationer. Ingen ytterligare hänsyn har tagits till om själva rörbrottet eller effekten av denna varat längre en det dygn som flaggats.



Figur 6. Rapporterade rörbrott för Mårdgatan 2015-10-10 respektive 2015-10-14

4.1.3 Summerade mätvärden

Det visade sig att det var möjligt att identifiera sekvenser där saknade värden summerats i en senare observation (Figur 7). Genom att rensa data från dessa värden förändrades tidsserien utseende drastiskt och många lokala extrempunkter kunde flaggas. Mätvärden som härstammar från denna typ av kategori innehöll fortfarande information som kunde utnyttjas med avseende på det totala flödet för en period. Men eftersom dessa observationer gav ett missvisande utseende hos mätseriens, har de markerats som felaktiga.

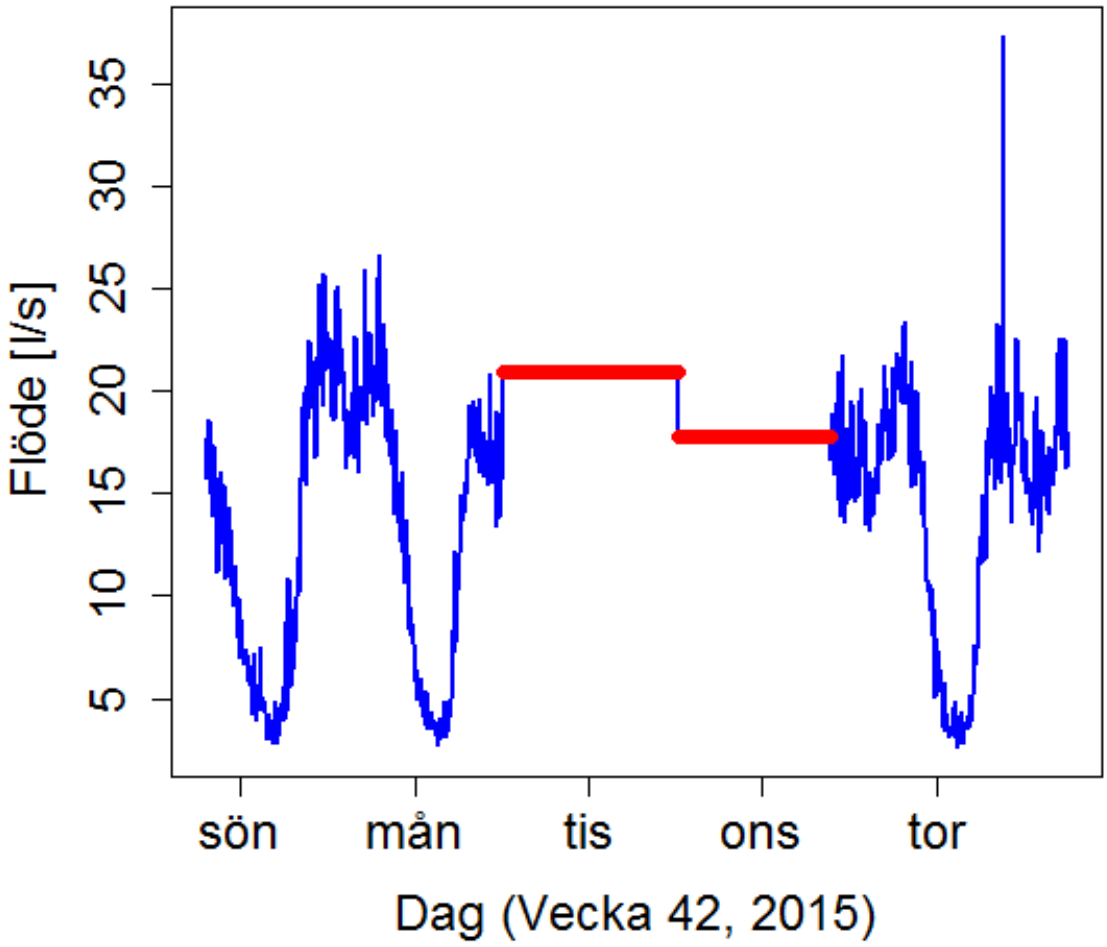


Figur 7. Flöde för Moldegatan 2 under 2013.

4.2 SKURLÄNGDSKODNING

Eftersom det finns en möjlighet att två efterföljande värden faktiskt är snarlika sattes ett kriterium att om tre eller flera efterföljande värden ger en skurlängdssekvens flaggas de observationerna. Särskild hänsyn tog dessutom till antalet angivna värdesiffror i respektive tidsserie. I en tidsserie med få värdesiffror är möjligheten stor att få flera skurlängdssekvenser med längden 2, särskilt under natten då förändringen i varians når dygnsminimum. Ett exempel för resultatet av skurlängdskodningen presenteras i Figur 8 för mätområdet Sjöbo. Röda punkter i figuren motsvarar detekterade upprepade värden.

I kvalitetsgranskningen har upprepade värden markerats som felaktiga då det inte ansågs troligt att flödet ska vara detsamma över en sammanhängande sekvens.

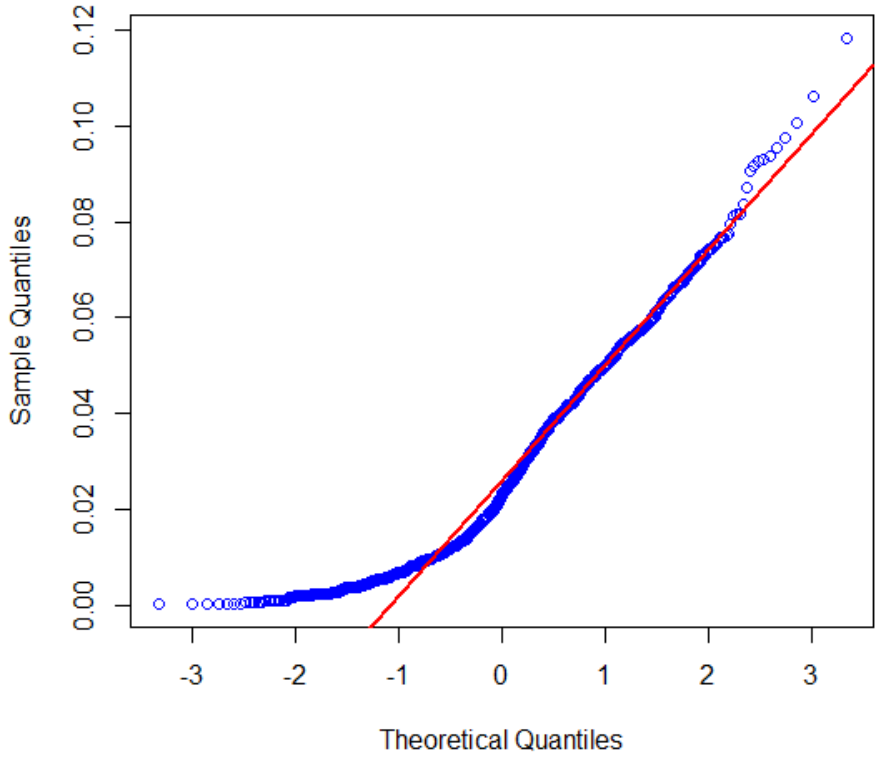


Figur 8 Resultat av skurlängdskodning för Sjöbo. Figuren visar två perioder i oktober 2015 med konstanta, varandra efterföljande värden (markerat i rött).

4.3 PERIODICITET

4.3.1 Fördelning

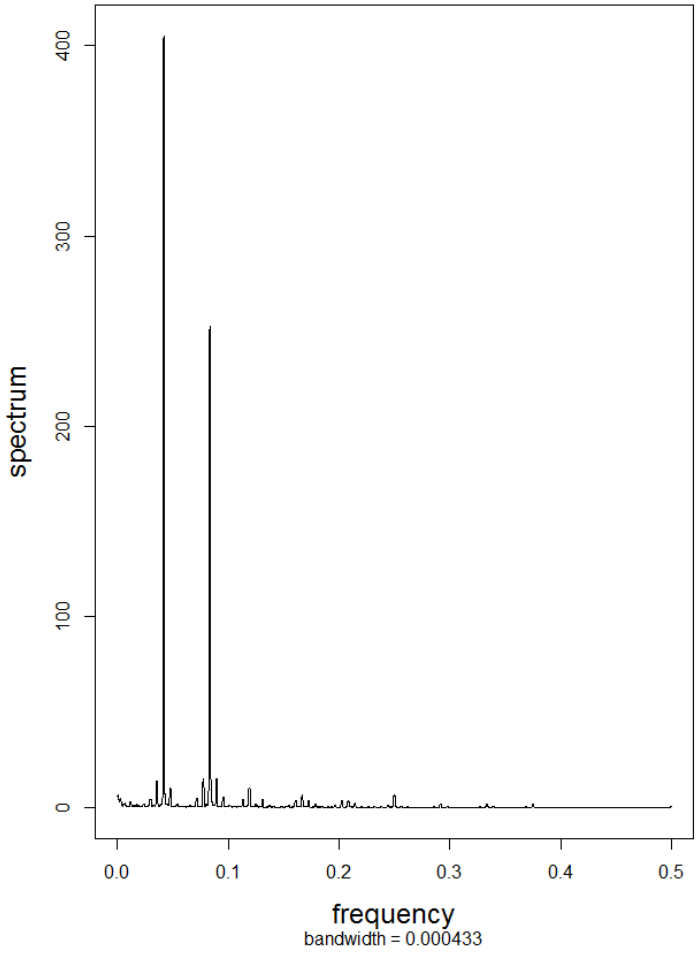
Gemensamt för rådatat för alla mätserier var att det inte verkade följa normalfördelningen. Ett Shapiro Wilk-test, som testar för normalfördelning, på mätserien för Klintesväng gav exempelvis ett lågt p-värde, detta gav anledning att anta nollhypotesen att rådatat inte var normalfördelad. Detta kunde även illustreras med en normalitetsritning, som visar kvantilerna för mätserien plottade mot de teoretiska kvantilerna för normalfördelning (Figur 9). Normalitetsritningen tyder på att det inte finns ett tillräckligt bra samband mellan mätseriens värden och de teoretiska värdena, framförallt vid låga mätvärden, för att normalfördelning ska kunna antas. Vi kan därför misstänka rådatat inte är normalfördelad.



Figur 9. Normalitetsritning för Klintesväng. Den raka linjen visar det förväntade sambandet mellan teoretiska och samplade kvantiler. Punkterna visar de samband som gäller för mätserien.

4.3.2 Spektralanalys

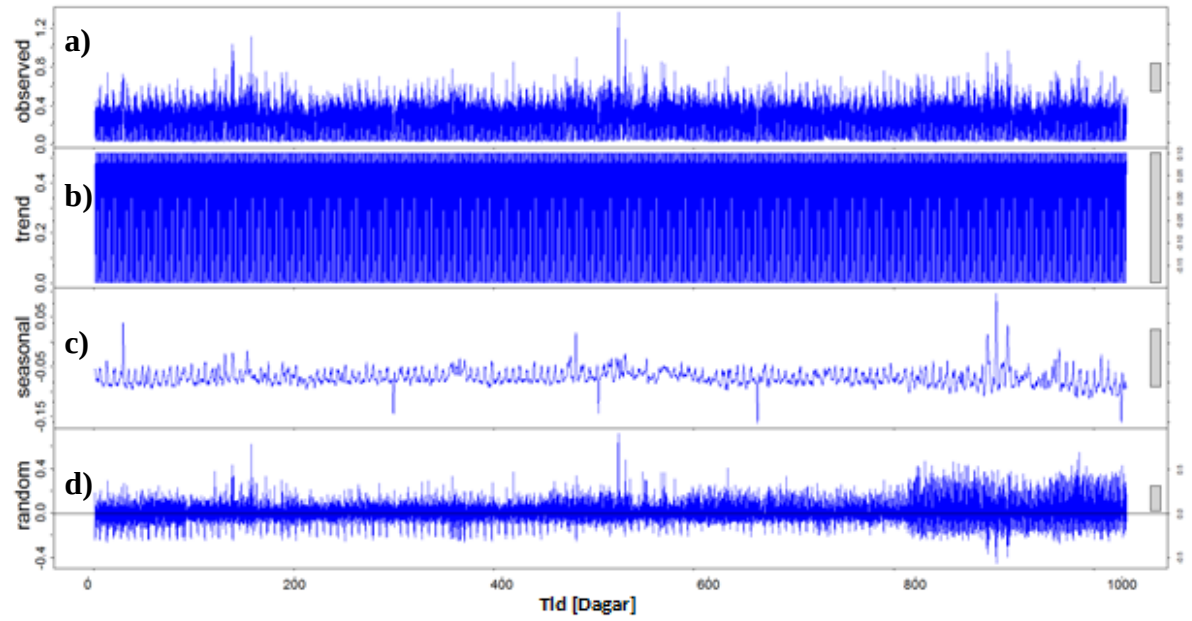
Resultatet från spektralanalysen visar frekvenser som motsvarar 24 timmar dominerar, alltså dagliga säsongen. Dessutom går det utläsa en periodicitet på 12 timmar, vilket troligen motsvarar de olika toppar som sker över dygnet (Figur 10).



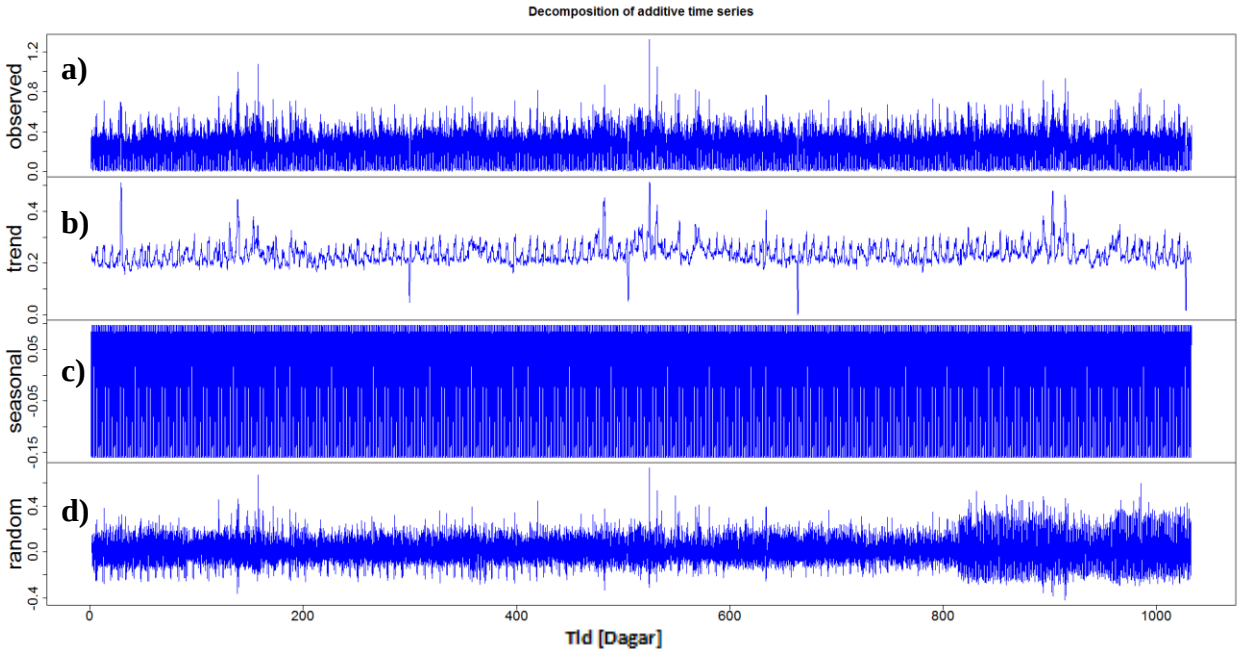
Figur 10. Peridogram över mätserien för Henstad-Hultsberg. Periodogrammet visar tydliga frekvenstoppar vid ungefär 0,4Hz och 0,8 Hz, vilket motsvarar 1/f ~ 24 respektive 1/f ~ 12 timmar

4.3.3 Säsongsrensning

Resultatet från säsongsrensningen med *STL* och flytande medelvärde visar att det inbakat i trendkomponenten verkar finnas en fortsatt periodicitet. Detta beror antagligen på att mätserierna har en komplex periodicitet med flera säsonger. Dessutom skedde en felskattning av trenden då en anomalie påträffades, vilket kunde ses som pulser i trendkomponenten (Figur 11 och Figur 12).

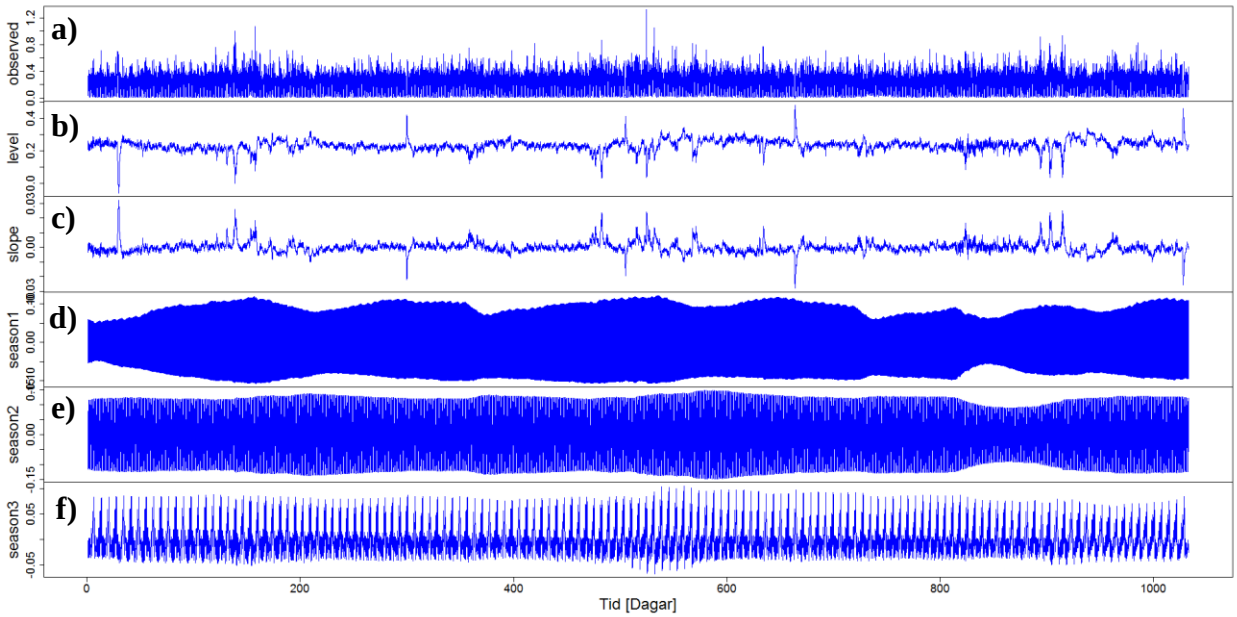


Figur 11. Säsongsrensad data för Tosseryd med hjälp av STL. Figurerna visar rådatat (a), följt av säsongskomponent (b) och trendkomponent (c). Nedersta grafen visar det kvarvarande bruset (d).



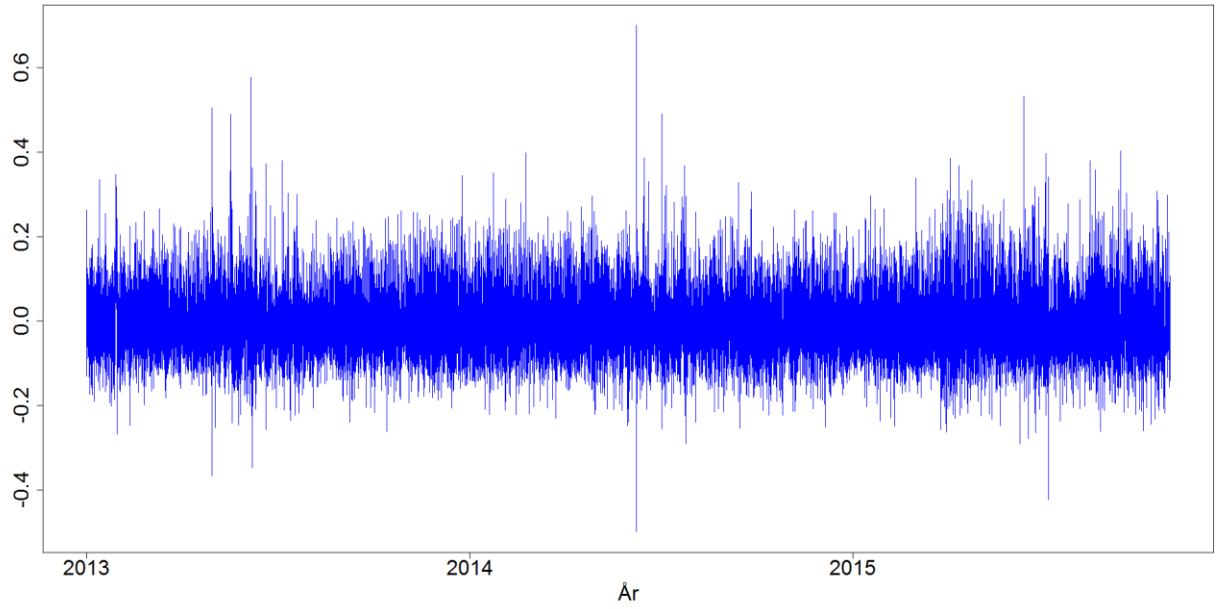
Figur 12. Säsongsrensad data för Tosseryd med hjälp av flytande medelvärde. Figurerna visar rådatat (a), följt av trendkomponenten (b) och säsongskomponenten med en periodicitet på 24 timmar (c). Den nedersta grafen visar det ”kvarvarande bruset” (d).

Med TBATS fanns möjligheten att rensa mätserierna från olika periodiska variationer. I Figur 13 har mätserien rensats från periodiciteten 12 timmar, 24 timmar samt veckovis periodicitet. Denna typ av modellering hade väldigt lång beräkningstid.

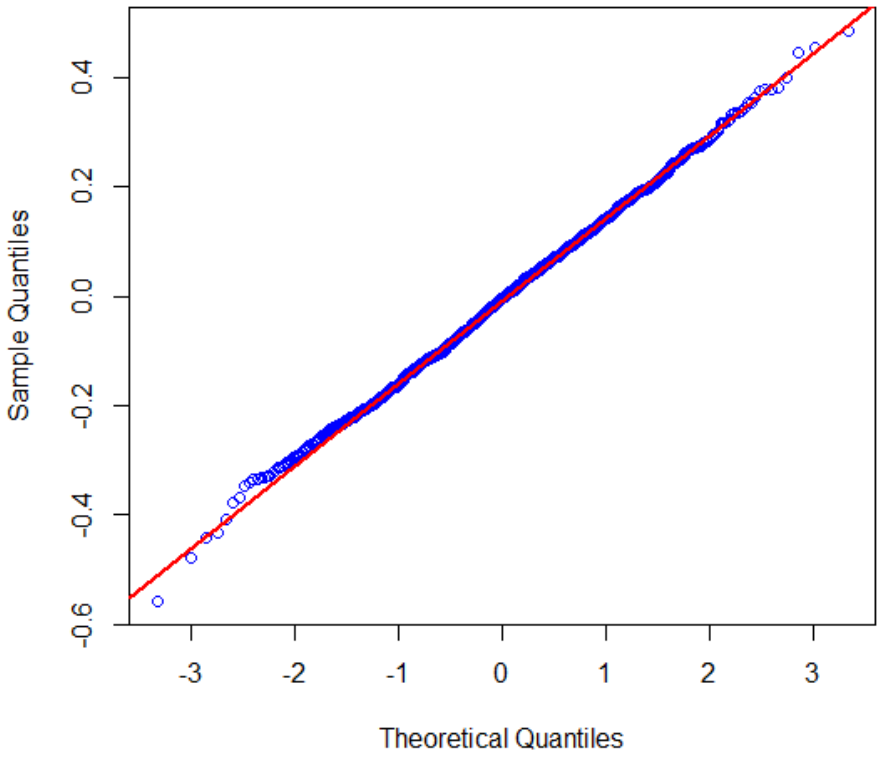


Figur 13. Säsongsrensning med TBATS-modellen. Figurerna visar rådatat (a) följt av level (b) och slope (c) som beskriver trenden över serien. De fyra understa graferna visar säsongskomponenterna för 12 timmar (d), 24 timmar (e) samt 168 timmar (f).

Figur 14 visar det "kvarvarande bruset" efter säsongsrensning med TBATS. Det gjordes även ett test huruvida resultatet från TBATS modellen följde normalfördelning. Med ett högt p-värde kunde vi inte förkasta normalfördelningen enligt Shapiro-Wilk testet. Normalitetsritningen visade även att residualerna kunde antas vara approximativt normalfördelade, varför det ansågs att observationer minst 3 standardavvikelse från medelvärdet kunde betraktas som outliers (Figur 15).

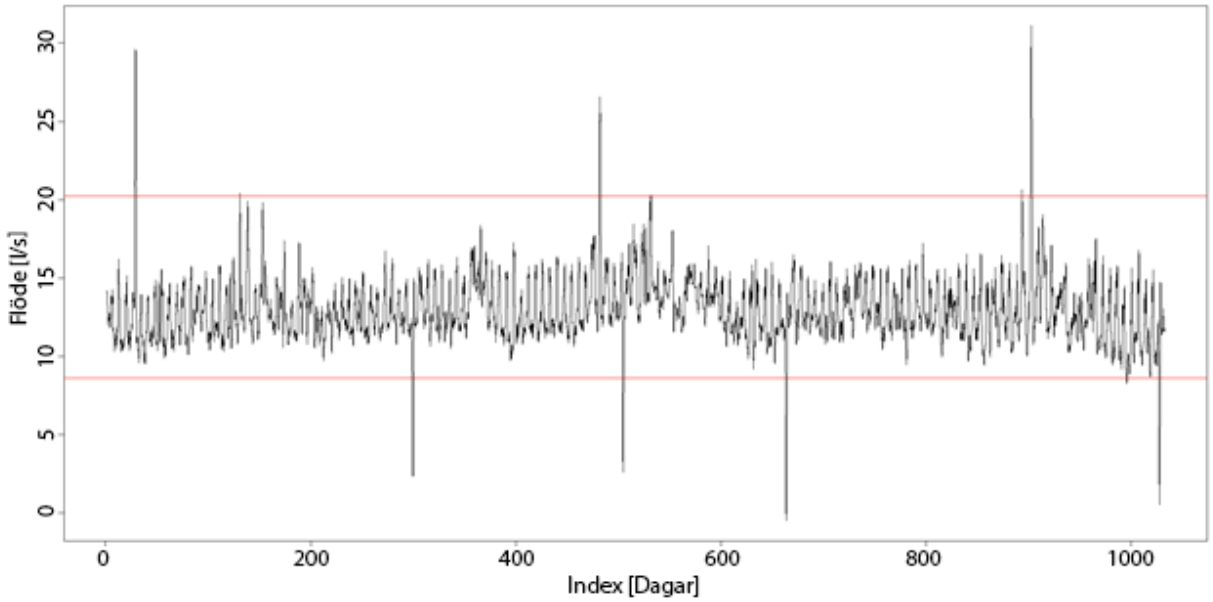


Figur 14 - Residualerna för säsongsresning med TBATS.

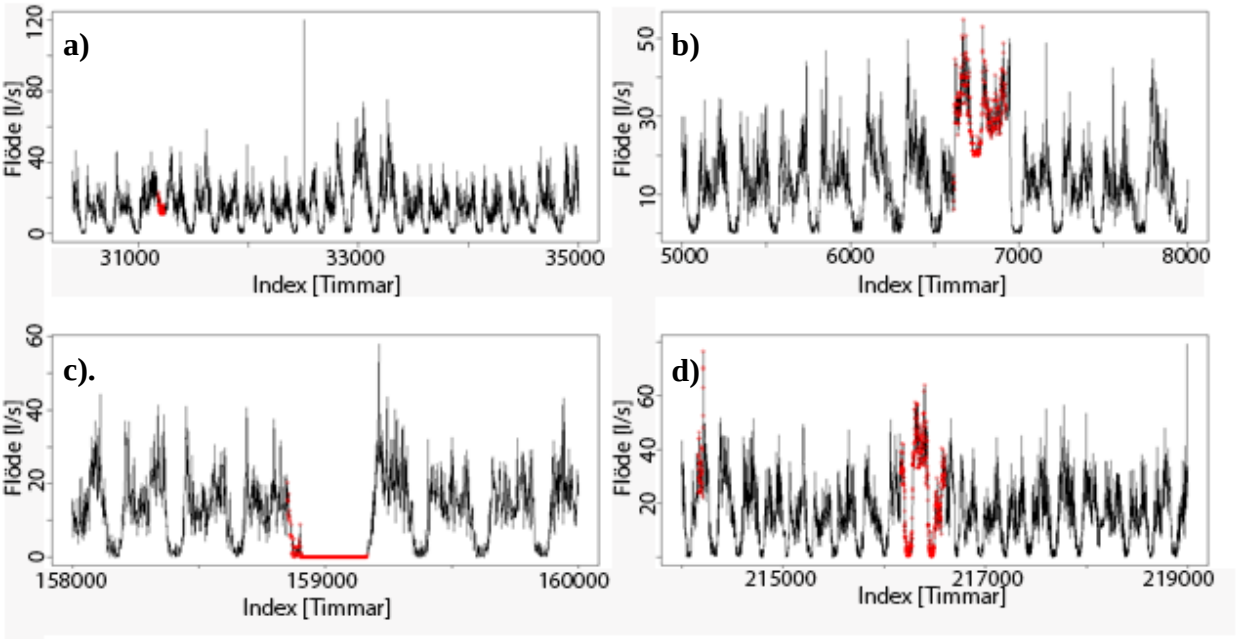


Figur 15 – Normalitetsritning från Residualerna av TBATS-modellen. Den teoretiska kvartilen plottas mot den samplade. Den raka linjen är den förväntade lösningen vid normalfördelning.

En upptäckt som gjordes i projektet var att den resulterande trendkomponenten från modelleringen med STL och flytande medelvärde kunde användas för att finna anomalier. Detta beror antagligen på att algoritmen för dessa felskattar, förutom kvarvarande periodicitet, även anomalier som en trend i modellen. Därför sattes ett konfidensintervall på 99,5 % av datat (Figur 16). Värden som översteg detta studerades okulärt. Det visade sig att höga pulser i trendkomponenten oftast berodde på en oregelbunden sekvens i mätserien (Figur 17). Anomalier upptäcka med denna metod har flaggats som suspekta då endast en okulär bedömning om dess riktighet har gjorts.



Figur 16. Trendkomponenten från säsongsrensningsprocessen med STL för mätserien Tosseryd. Höga pulser kan ses i trenden. Ett tröskelvärde sattes vid 0,5 % respektive 99,5 % för att sortera ut topparna.

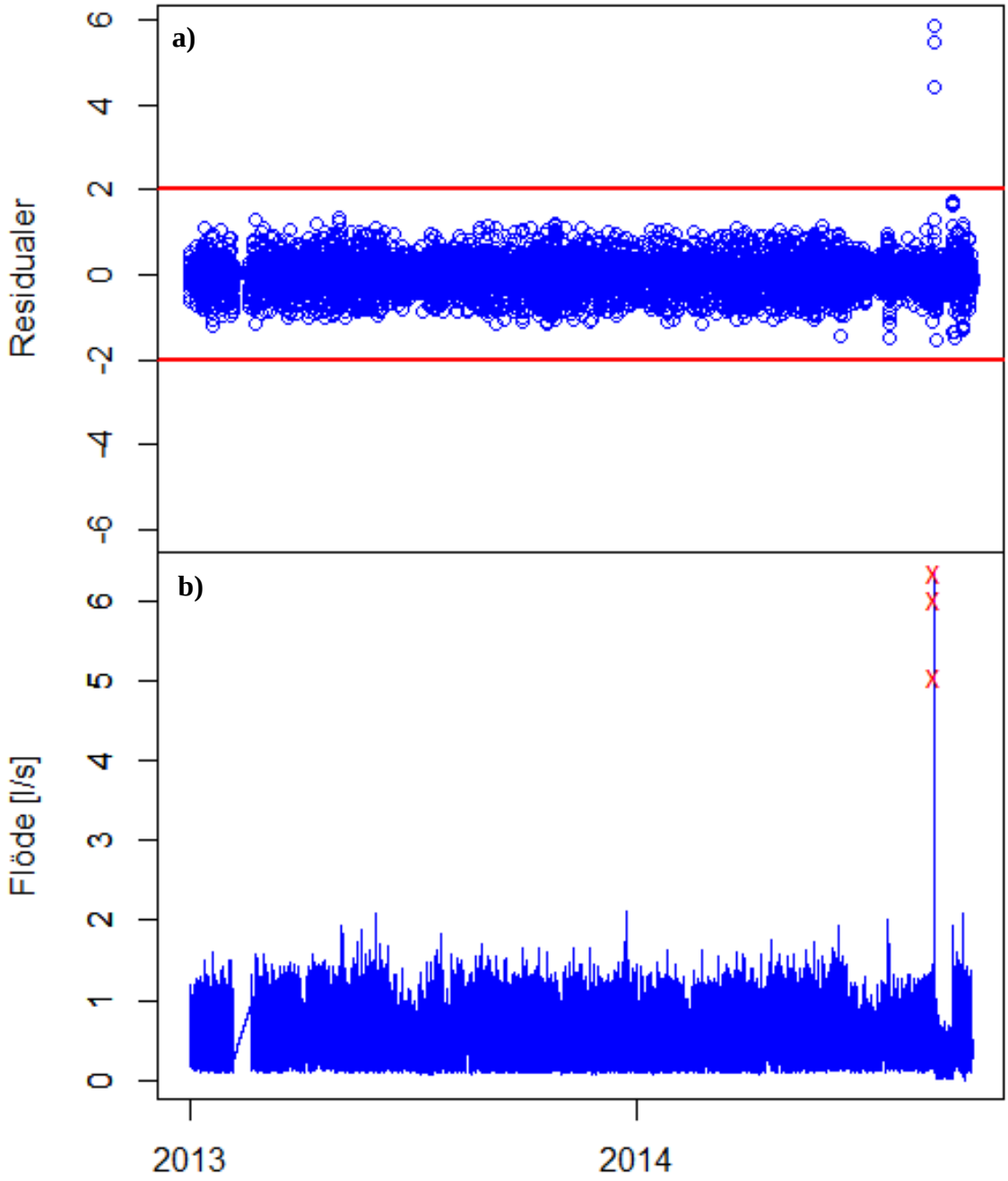


Figur 17 - Sekvenser av rådatat där pulser i trenden har uppstått vid säsongsrensningen med STL. a) Markerar observationer där nattliga förbrukningen var något högre än normalt. b) markerar en sekvens där ett dygns medelvärde var förhöjt mot normalfallet. c) Sekvenser där förbrukningen var noll markerades även som pulser i trenden. d) Två sekvenser markerade, den första berodde på ett maximumvärde. I den andra sekvensen markerades observationer som tydde på en förhöjd nivå under dagtid mot normaldygnet.

4.4 STATISTISKA ANALYSER

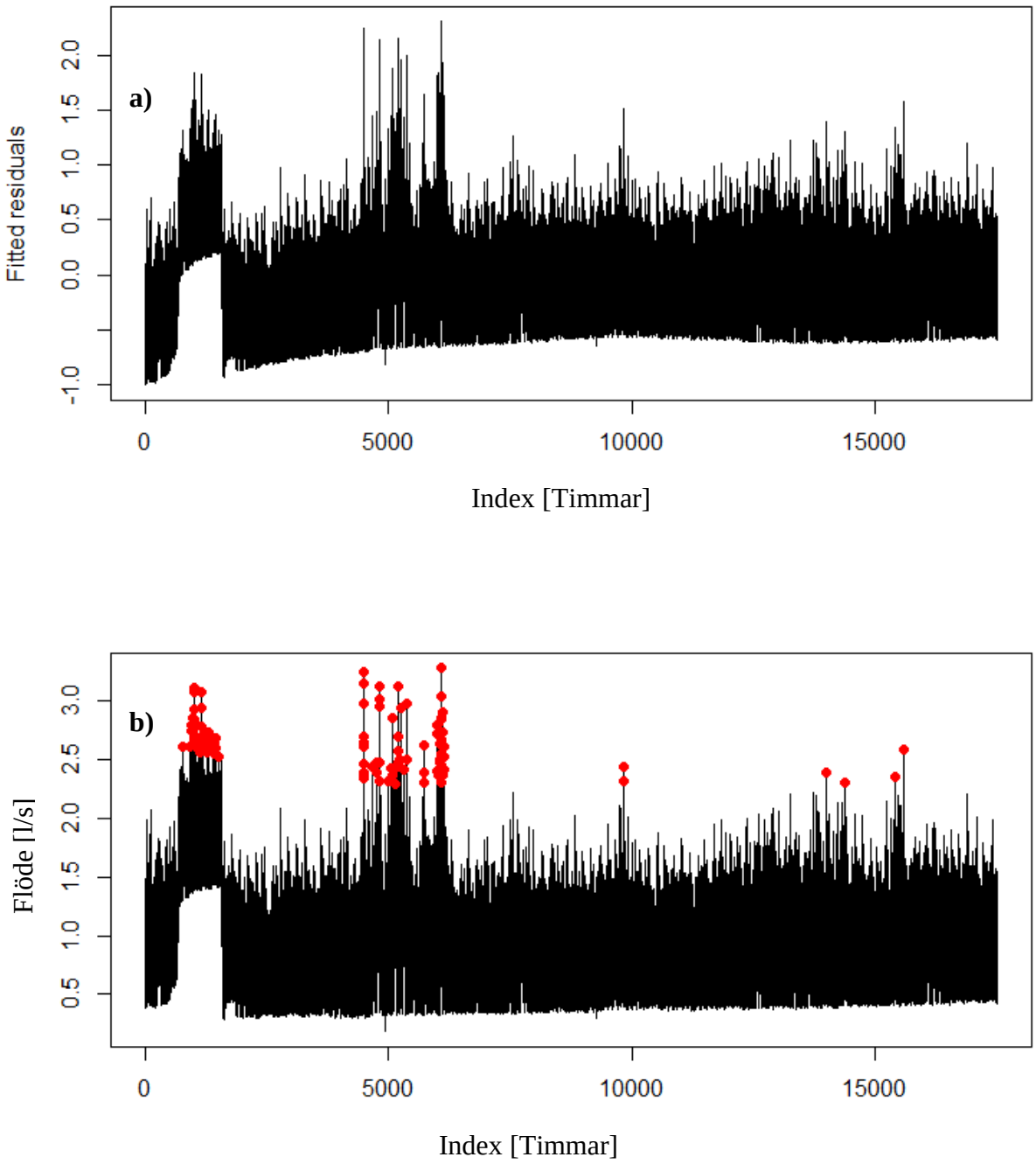
4.4.1 Outliers

Från residualerna av säsongsrensad data kunde additiva outliers flaggas. Detta gjordes genom att sätta ett tröskelvärde på residualerna. Den övre figuren (Figur 18**Error! Reference source not found.**) visar säsongsrensad serie för Mårdgatan. Den nedre visar de outliers som överstiger tröskelvärdet för den serien.



Figur 18. Residualanalys för Mårdagatan. a) Residualerna, där den röda linjen är tröskelvärdet för var outliers skall markeras. b)Rådatat med tre röda kors som tillhör de observationer som översteg tröskelvärdet i residualanalysen.

Försök gjordes även att detektera outliers med en LOESS modell över dataserien, det vill säga ingen hänsyn togs till periodicitet. Den övre figuren (Figur 19) visar LOESS modellen över mätserien för Alster och den undre figuren visar rådatat över Alster med flaggade outliers.

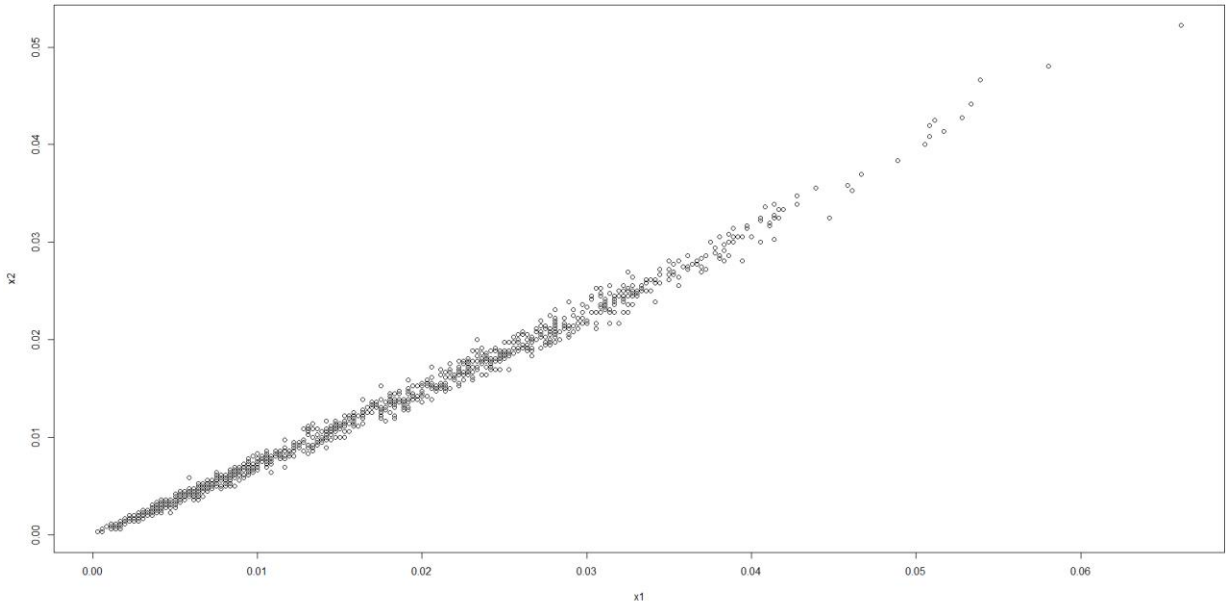


Figur 19. Outliers detektion från en LOESS kurva (Alster). Observationerna markeras som outliers då residualerna överstiger tröskelvärdet.

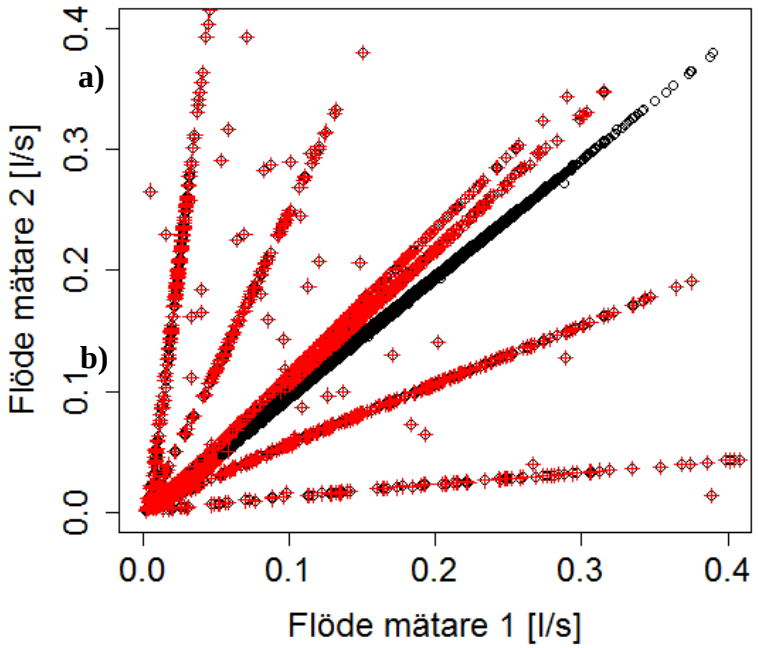
De outliers som upptäckts med dessa metoder markerades som suspekta.

4.4.2 Regressionsanalys

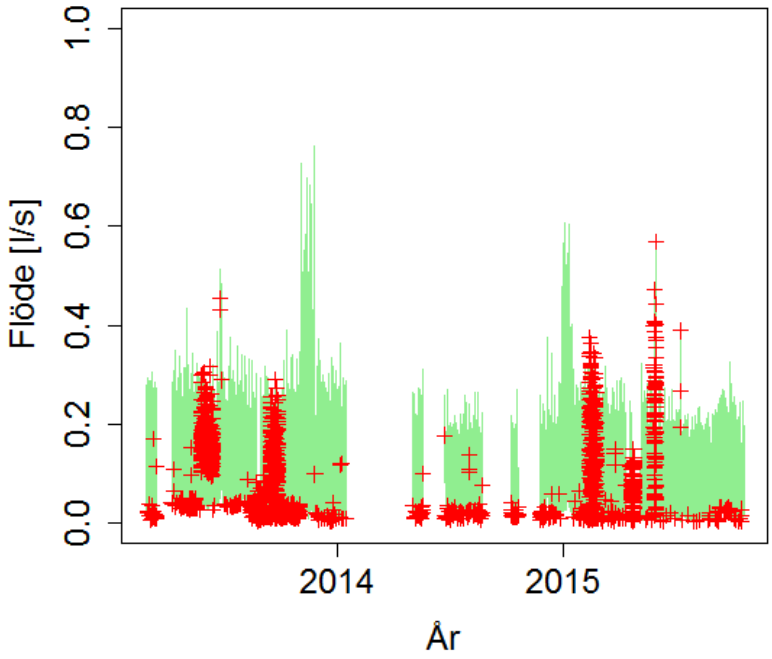
Regressionsanalysen av de parallellkopplade serierna förväntades vara proportionella, då mätningarna skedde i samma mätpunkt (Figur 20). I några fall uppstod emellertid flera olika linjära samband (Figur 21). Då detta fenomen studerades närmare visade det sig att det härrörde från särskilda perioder i mätserier som var sammanhängande i tiden och där den ena mätaren visade linjärt mer eller mindre än den andra mätaren (Figur 22). Värden som härstammade från dessa perioder ansågs bero på en felaktighet i någon eller båda mätarna, varför observationerna flaggades som suspekta.



Figur 20. Två parallellkopplade mätare för Klintesväng plottade mot varande. Exempel där ett linjärt samband visade sig för de parallellkopplade mätarna.

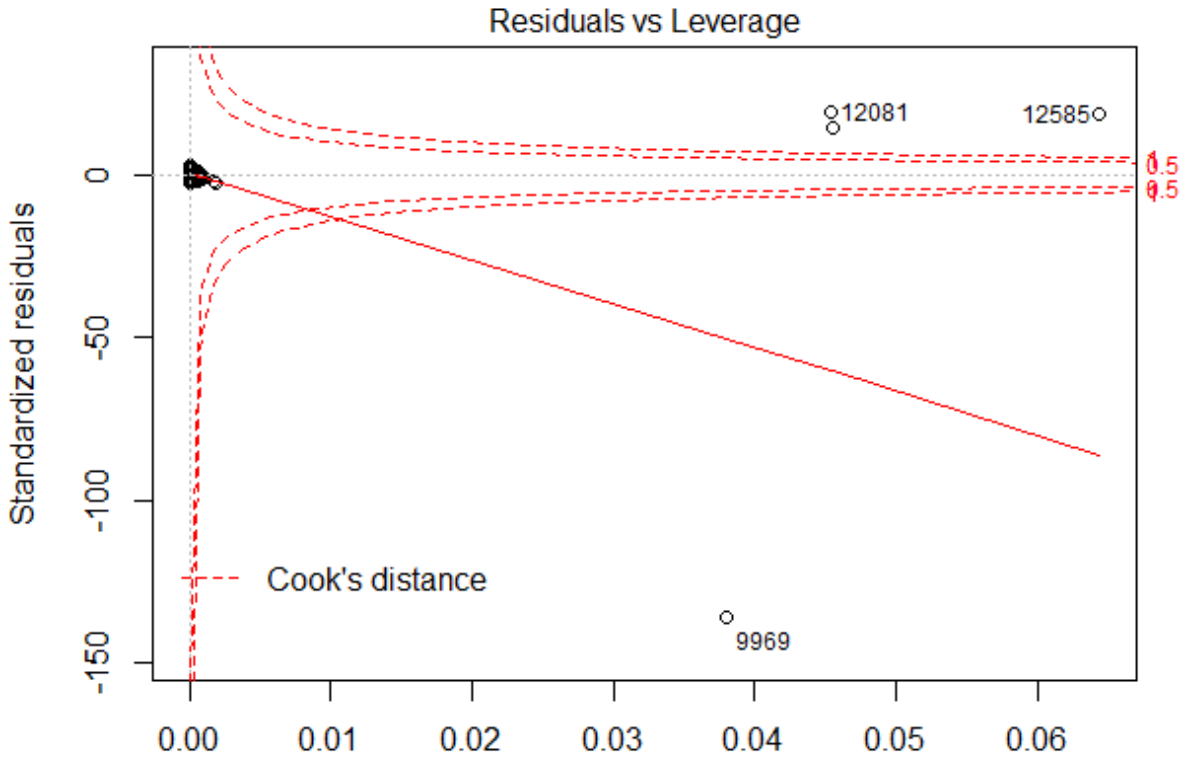


Figur 21. Billdalsgatans två mätare plottade mot varandra. I figuren kan ses att det finns flera linjära samband. De röda punkterna är observationer som flaggats som varningar på grund av att kvoten mellan mätarna antogs felaktiga med avseende på normalfallet inom tidsserien.

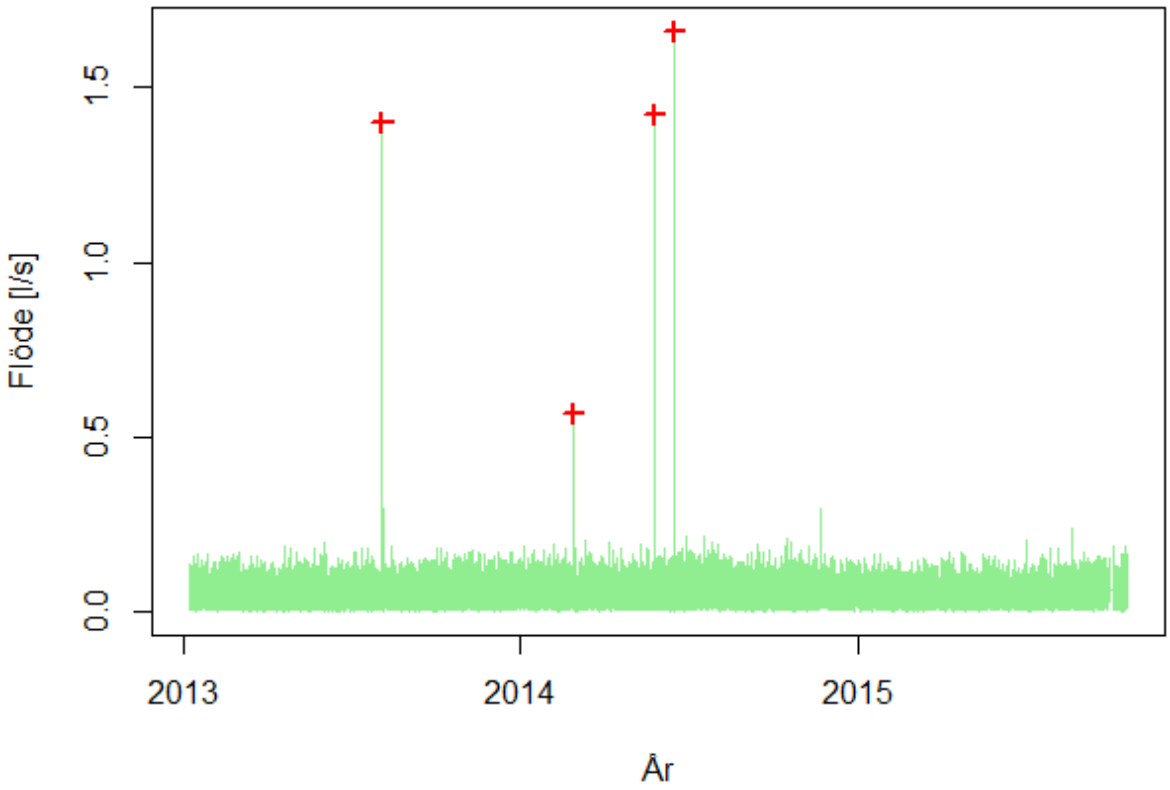


Figur 22. Totala Flödet för Billdalsgatan. Röda punkter visar de observationer där de två parallellkopplade mätarna har olika proportioner mot normalfallet hos tidserien.

Vidare visade analyser från regressionsmodellen att outliers kunde identifieras med Cooks avstånd. Detta gjordes med utgångspunkten att Cook's avstånd var större än 1 (Figur 23). Cook's avstånd är större än 1 då residularna för den linjära modellen är stora eller då observationen påverkan på skattningen är stor. Detta visualiseras som den röda streckade linjen i figuren. Notera att det finns två streckade linjer varav den inre av dem är då Cook's avstånd är lika med 0,5. Observationer identifierade med Cook's avstånd betraktades som outliers och markerades som suspekta (Figur 24).



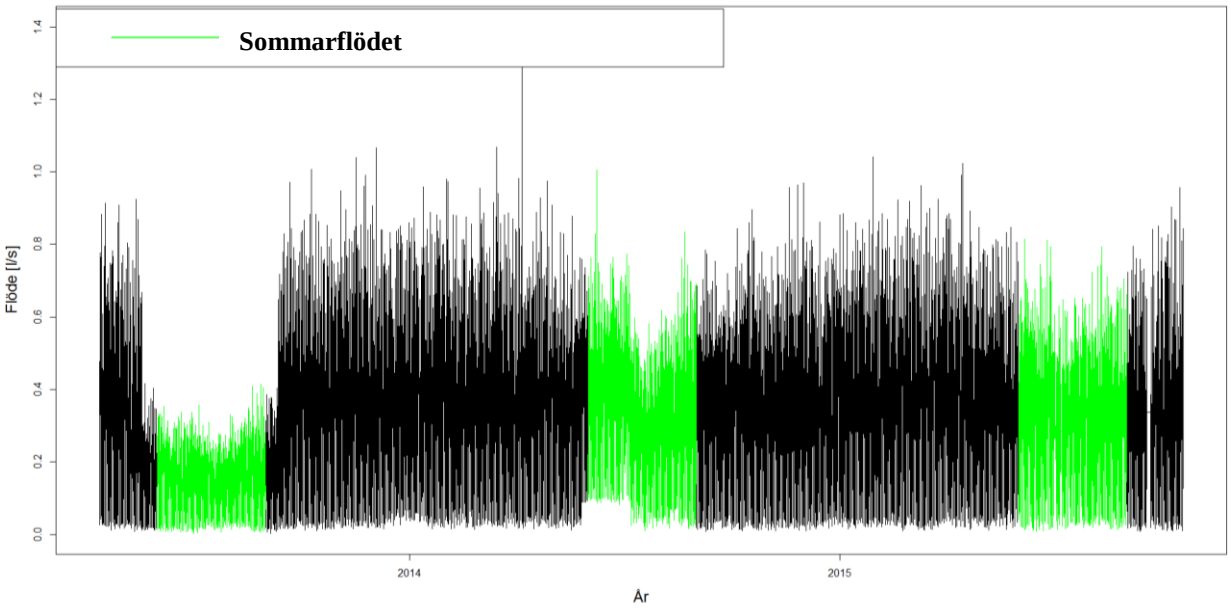
Figur 23 - Residualerna mot effekten de har på skattningen av modellen för Hällegatan 22. Cooks avstånd = 1 visas som den yttre röda streckade linjen i grafen.



Figur 24. Flödet för Hällegatan 22. Röda punkter är outliers som detekterats med hjälp av Cooks avstånd.

4.4.3 Förändringar i variansen

För sommarmånaderna förväntades att flödet skulle förändras, antingen genom att antalet brukare ökar eller minskar eller att någon form av trädgårdsbevattningen. För mätserier på enskilda fastigheter kunde det okulärt avgöras att förbrukningen i många fall minskade (Figur 25).

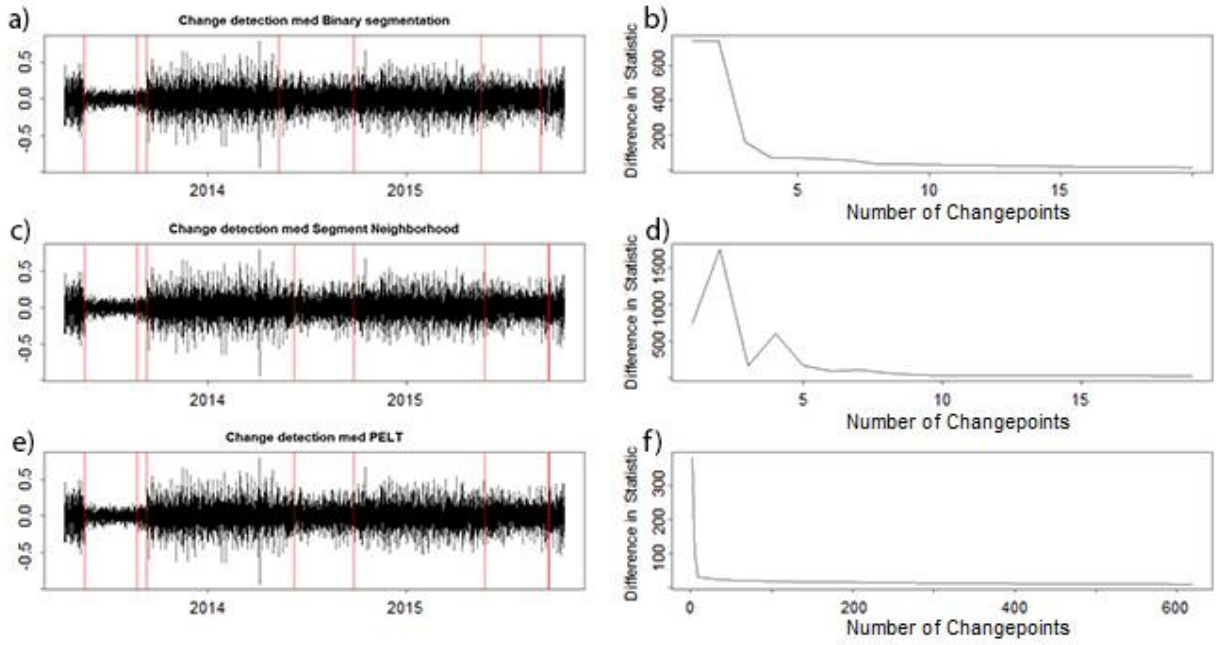


Figur 25. Flödesdata för Vejlegatan 6. De gröna områdena visar perioden maj-augusti för åren 2013-2015.

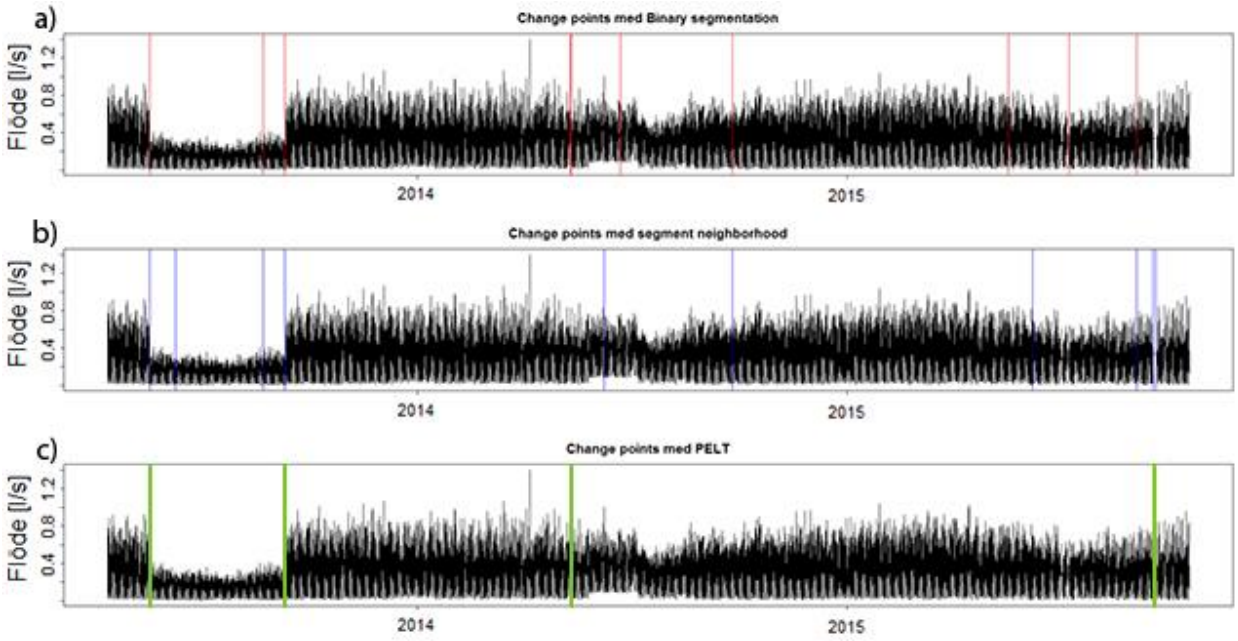
För att identifiera förändringar i variationen för vattenförbrukningen undersöktes plötsliga förändringar i variansen på säsongsrensad data. För detektion av förändringar i variansen krävdes att hela serien rensades från periodicitet och trend. I denna sektion presenteras resultat som har rensats med hjälp av den differentierade mätserien. Denna metod var enkel att implementera och krävde inte långa beräkningstider.

Med paketet *changepoint* i R testades binär segmentering, Segment neighborhood samt PELT (Figur 26). Där visar figurerna till vänster visar resulterade förändringspunkter för de tre metoderna och figurerna till höger visar skillnaden i statistik mellan varje förändringspunkt. Den senare kan användas som ett argument för hur många förändringspunkter som ska illustreras.

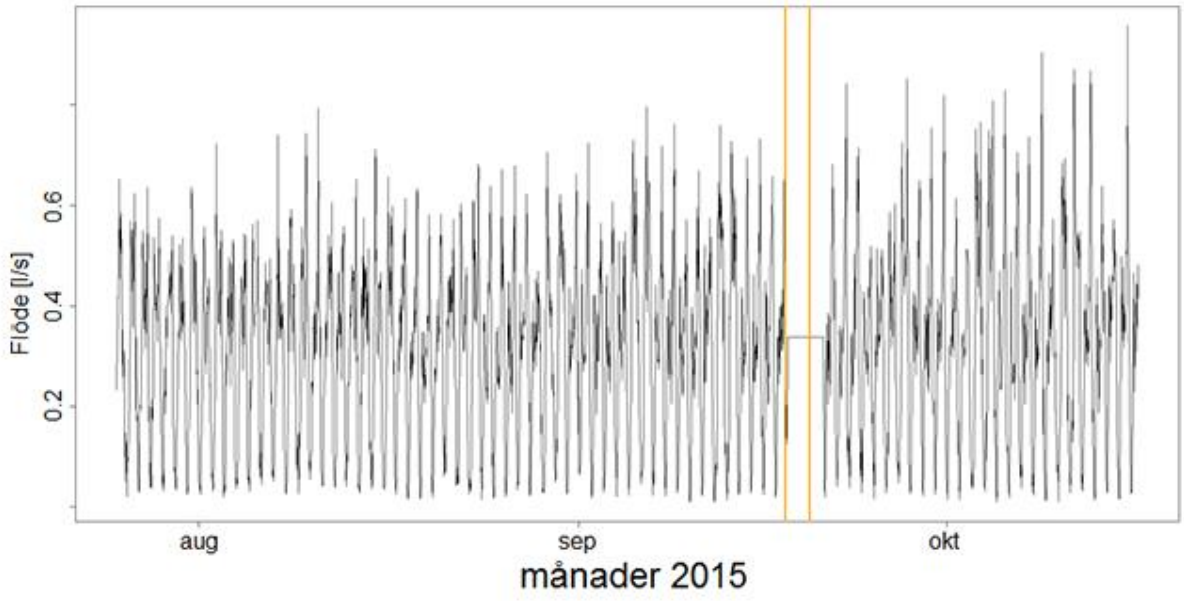
Alla tre metoderna lyckades fånga de förändringspunkter som inträffade under sommaren (Figur 27). Dock var det vid de långa mätserierna (över 10000 observationer) väldigt lång beräkningstid för segment neighborhood, ca 20 minuter, medan PELT hade en beräkningstid på ca 20 sekunder och binär segmentering endast hade en beräkningstid på ca 0,2 sekunder. Det gick dock att se att binär segmentering inte lyckades att fånga vissa anomalier i mätserierna vilket både PELT och segment neighborhood lyckades med. Binär segmentering upptäckte inte heller sekvenser med vissa typer av anomalier, såsom upprepade värden med variansen noll, vilket både PELT och segment neighborhood klarade av (Figur 28).



Figur 26. Identifierade förändringspunkter för den differentierade mätserien av Vejlegatan 6 med metoderna binär segmentering (a), segment neighborhood (c) samt PELT (e). Respektive figur till höger (b, d och f) visar skillnaden i statistiken för förändringspunkterna, vilken hjälpte som ett argument för hur många förändringspunkter som skulle visas.

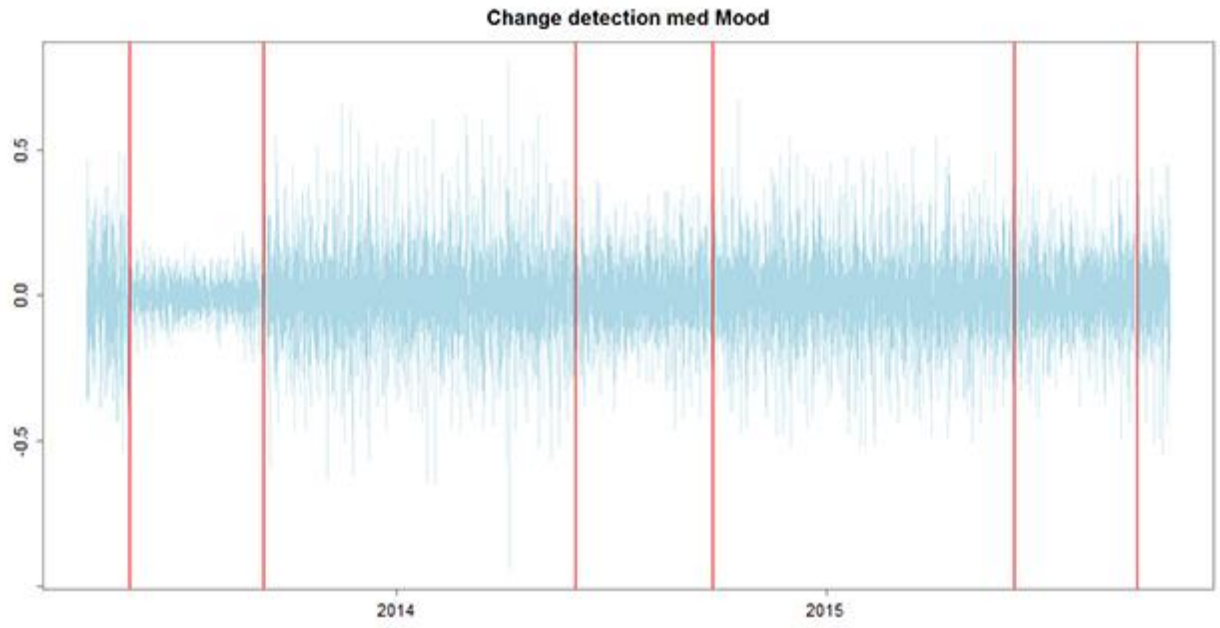


Figur 27. Upptäckta abrupta förändringar plottade i rådatat för Vejlegatan 6. Visar förändringspunkter i variansen med binär segmentering (a), segment neighborhood (b) och PELT (c).



Figur 28. Tidserie med en sekvens av konstanta värden. Både PELT och segment neighborhood identifierade denna sekvens som en sekvens med annorlunda fördelning.

Med paketet *cpm* detekterades förändringar i variansen med hjälp av statistiken för Mood (Figur 29). Resultatet av denna visade på att de förändringar i vattenförbrukningen som skedde under sommaren (se Figur 25) detekterades i algoritmen.



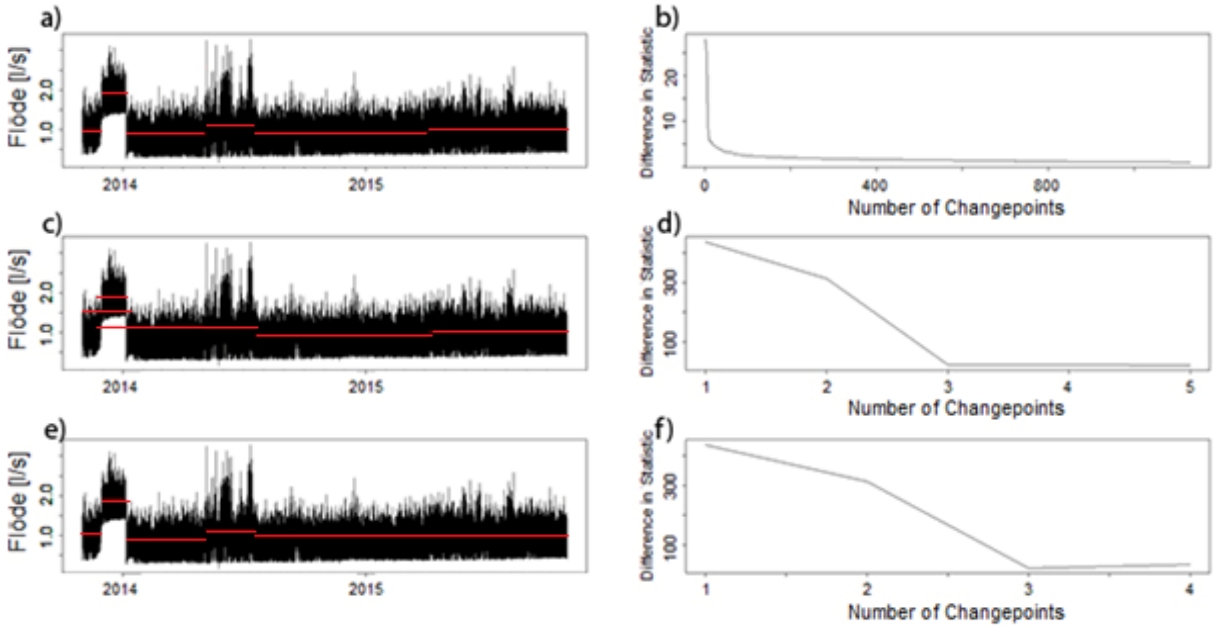
Figur 29 - Resultatet från detektering av förändringar i variansen på den differentierade serien för Vejlegatan 6. Identifierade förändringspunkter är markerade som röda linjer i grafen.

Upptäckta sekvenser med abrupta förändringar i variansen har markerats som suspekta och behöver analyseras vidare då variationsförändringar eventuellt har naturliga orsaker.

4.4.4 Förändringar i medelvärde

I paketet *changepoint* testades förändringar i medelvärde med binär segmentering, segment neighborhood och PELT (Figur 30). Det visade sig att binär segmentering lyckades väldigt bra med att identifiera tydliga förändringar i medelvärdet men att den också angav att en sekvens tillhörde flera olika överlappande fördelningar. Segment neighborhood och PELT klarade bättre av att beskriva en större andel olika förändringar. PELT lyckades dessutom med att fånga att hela mätseriens medelvärde ökar svagt med tiden. En ytterligare fördel var även att beräkningstiden

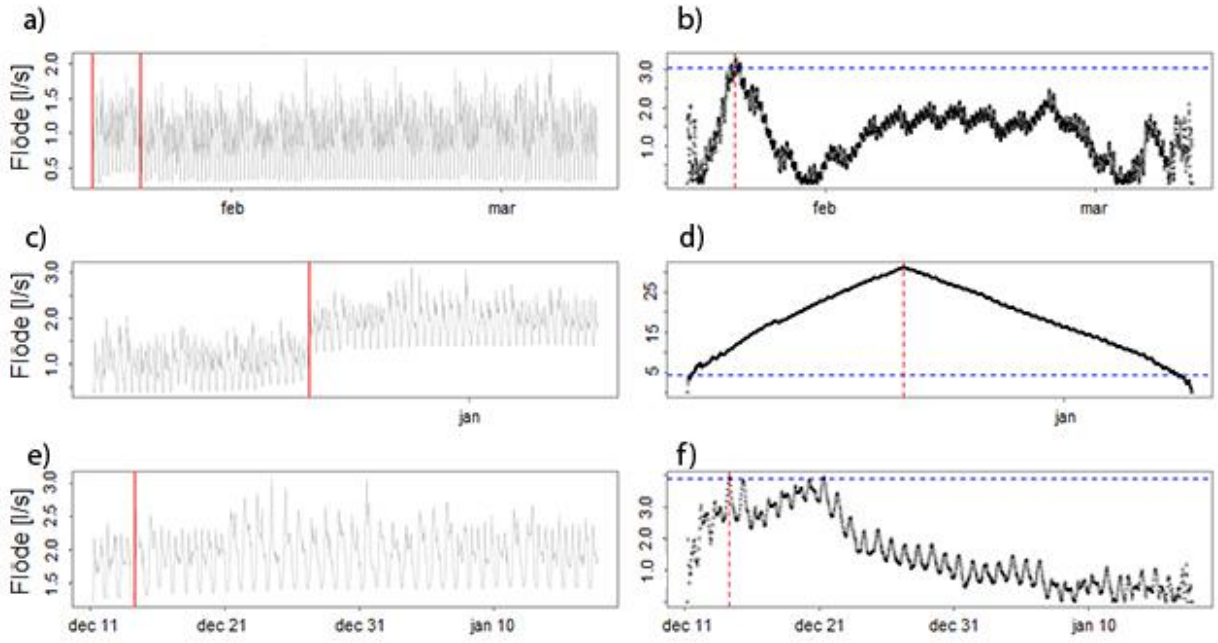
var betydligt snabbare än för segment neighborhood (Figur 30).



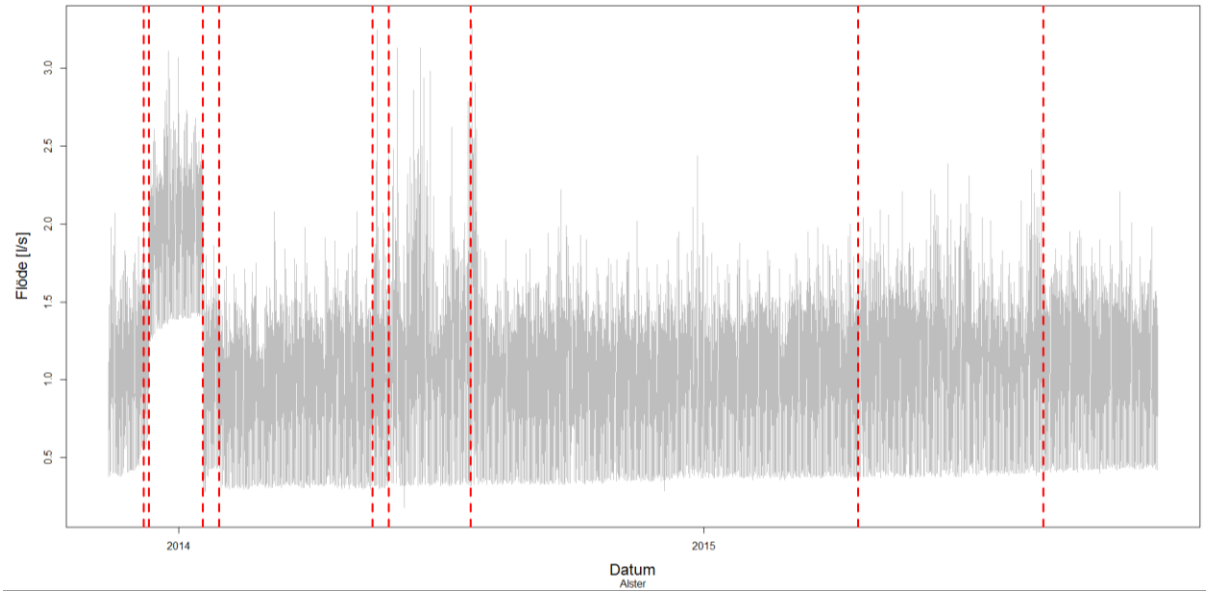
Figur 30 - Förändringspunkter för binär segmentering, segment neighborhood och PELT. Figurer till höger visar motsvarande metods skillnad i statistik vilken hjälper som argument för hur många förändringspunkts som bör anges.

Med paketet *cpm* och statistiken för Mann-Whitney, där förändringar i medelvärdet identifierades (Figur 31). Det gick att se i figuren att de tydliga anomalierna fångades väldigt bra. Graferna till vänster visar några sekvenser där en abrupt förändring har identifierats. Bilderna till höger om dessa sekvenser visar motsvarande sekvens statistiska värde enligt Mann-Whitney metoden. Den abrupta förändringen detekteras för den observation som får högts poäng enligt testet om det överstiger tröskelvärdet. Det sammanlagda resultatet för detektionen av abrupta förändringar för Alster, visas i Figur 32. Som jämförelse presenteras resultatet med samma test för Vejlegatan 6 (

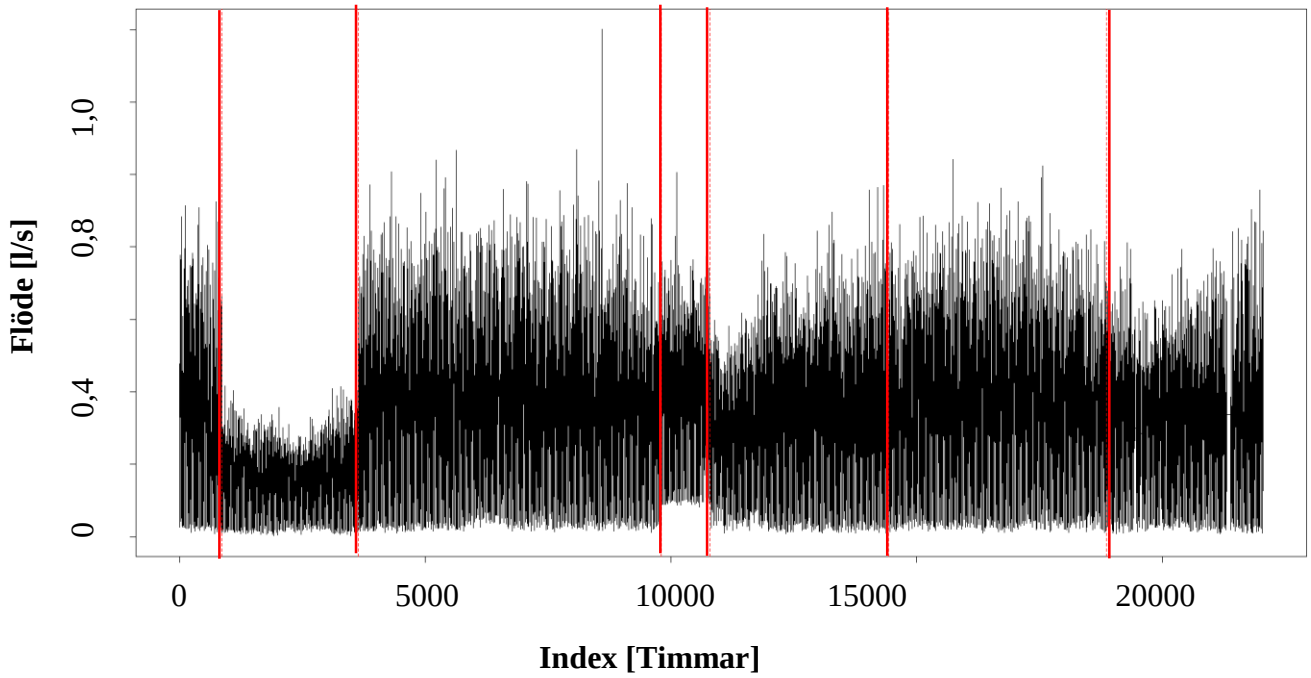
Figur 33). Även där verkade metoden fånga abrupta förändringar bra.



Figur 31. Resultatet för detektion med statistiken för Mann-Whitney för Alster. Figurerna till höger visar sekvenser där en abrupt förändring har detekterats. Figurerna till höger är motsvarande sekvens statistik för varje mätpunkt. De röda linjerna motsvarar en detekterad förändring och de blåa linjerna motsvarar ansatt tröskelvärde.



Figur 32 - Visar samtliga abrupta förändringar som kunde upptäckas med statistiken för Mann-Whitney för mätområdet Alster. De röda linjerna indikerar detekterad förändring.

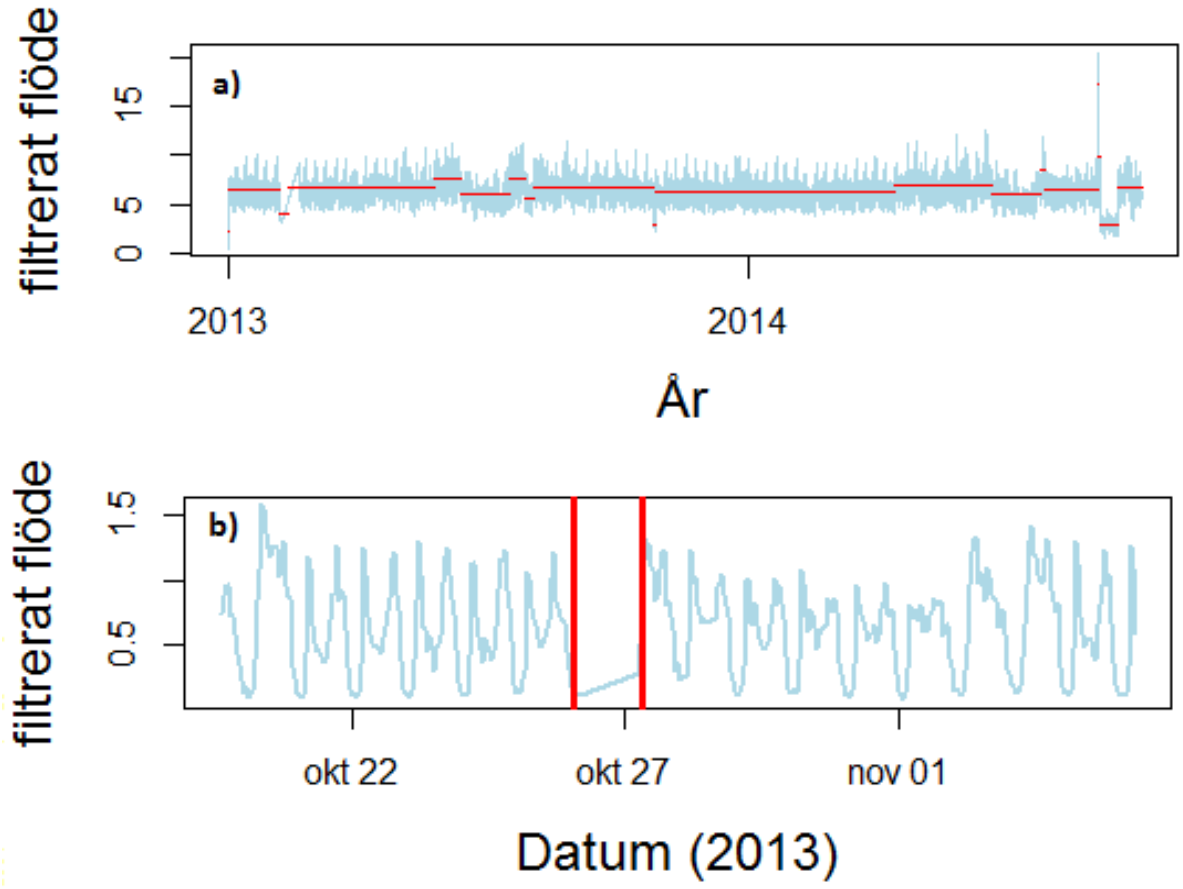


Figur 33 – Markerade förändringspunkter för Vejlegatan 6 beräknad med statistiken för Mann-Whitney. Vilken söker efter förändringar i medelvärdet.

Abrupta förändringar i medelvärdet har markerats som suspekta och behöver genomgå ytterligare analyser.

4.4.5 Rekursiva filter

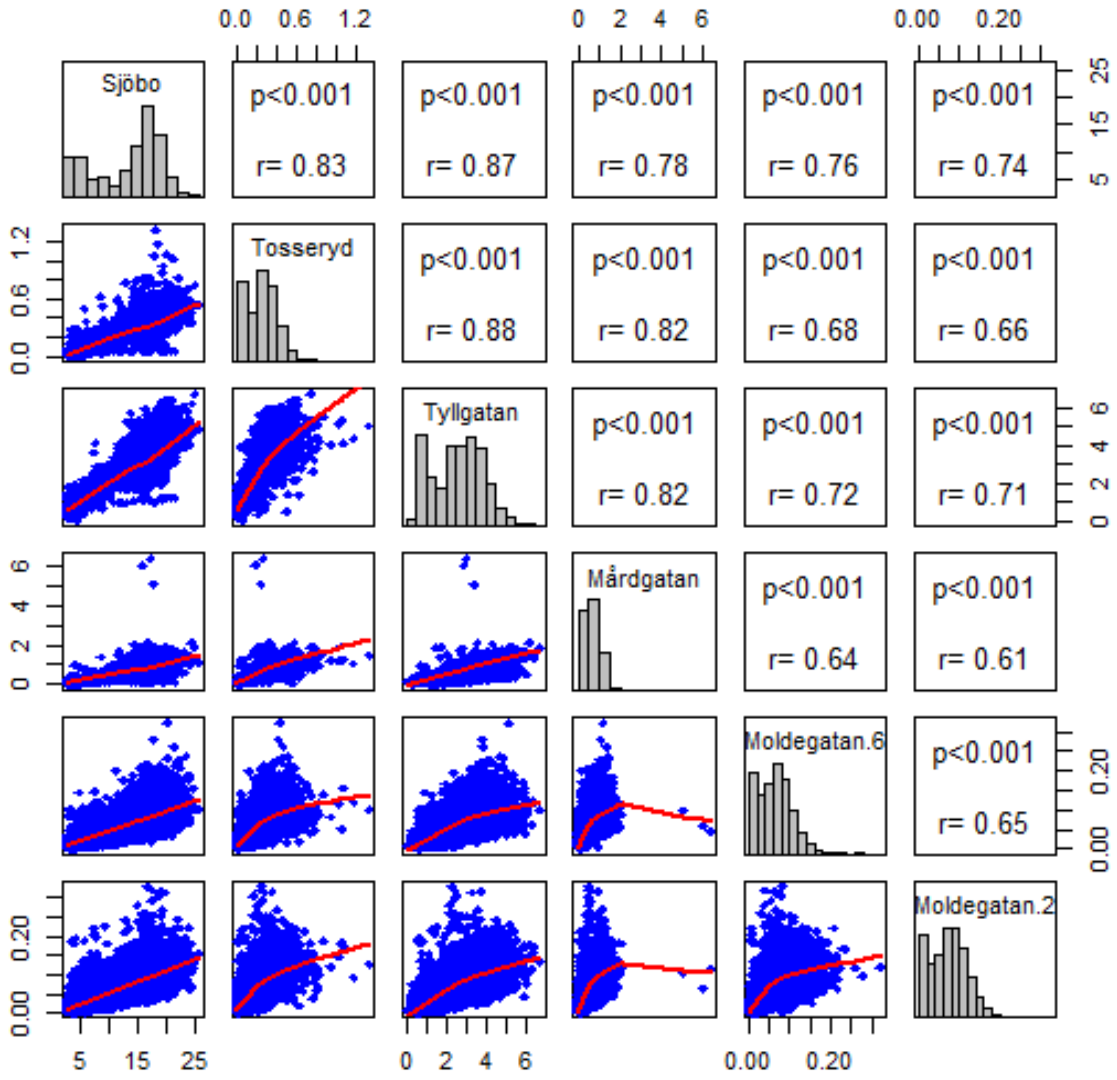
Rekursivt filter tillsammans med detektering av förändringspunkter visade sig vara användbart vid identifiering av anomalier. Genom att lägga till ett rekursivt filter till mätserien kunde anomalier förstärkas för att underlätta identifieringen av dem. Genom att sedan studera förändringspunkter på den filtrerade mätserien hittades flera anomalier (Figur 34). Vid användandet av rekursiva filter väljs ett värde på filtret mellan 0 och 1 i R. För värden nära 1 på filtret visade det sig att mätserierna påverkades som mest av filtreringen, vilket även ledde till en förenklad detektering av anomalier. Med anledningen av detta valdes värdet på filtret i R till 0,9 för alla tester med rekursiva filter.



Figur 34. a) Förändringspunkter detekterade med PELT för Mårdgatan filtrerat med ett rekursivt filter med värdet 0,9. b) Ett exempel på en anomali från rådatat som kunde detekteras tack vare att mätserien filtrerats

4.5 KORRELATION

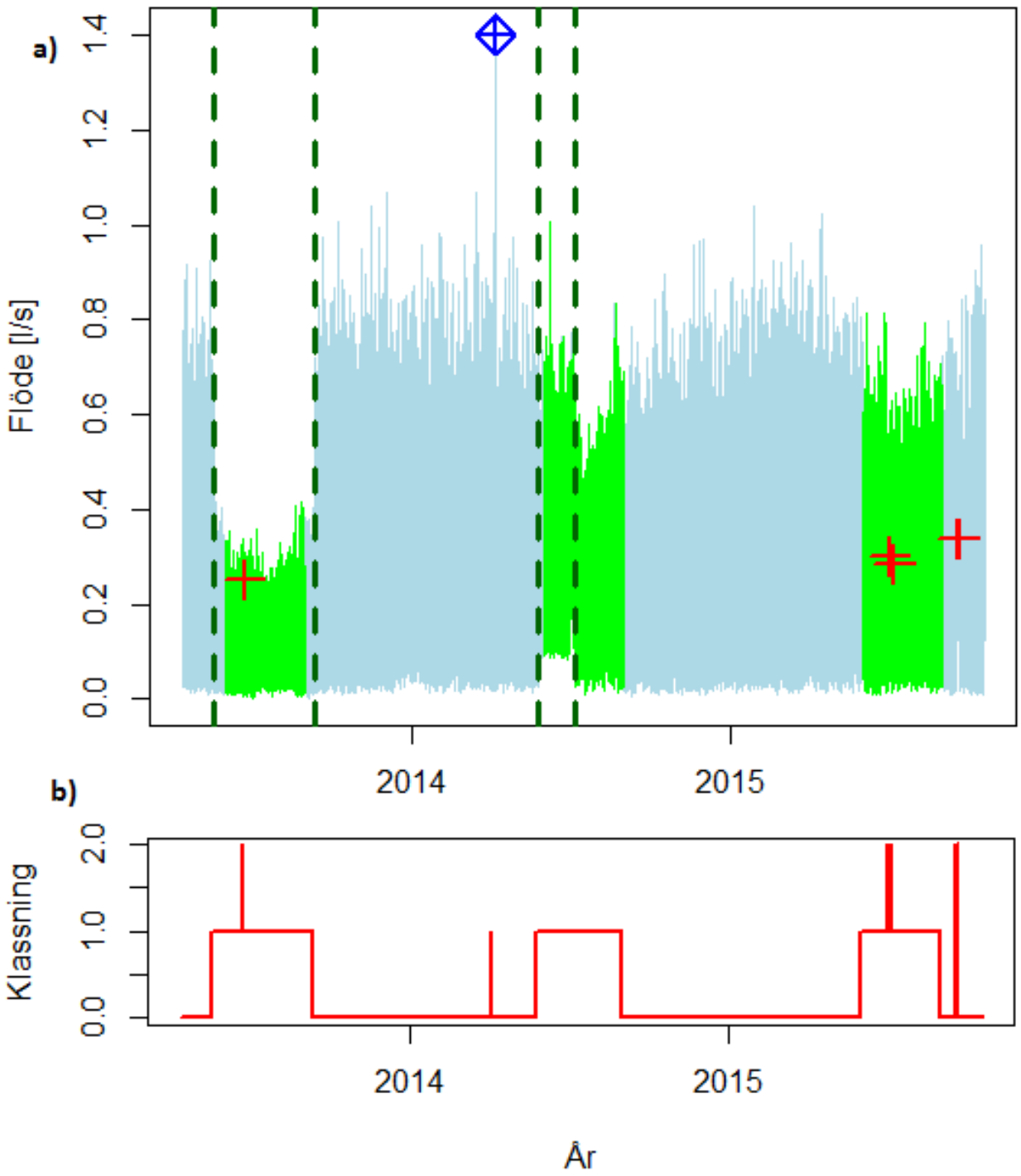
Korrelationsgraf visar på större korrelation mellan större områden än med små (Figur 35). Hypotesen var att de större områdena skulle ha mindre skillnader sinsemellan än de mätserierna med enskilda brukare, på grund av att individuellt beteendemönster får mindre effekt i större mätområden. Vilket alltså verkar stämma utifrån korrelationsgrafen. I figuren går det att se att två av de större områdena exempelvis Sjöbo och Tosseryd visade ett R^2-värde på 0,83 och från ett korrelationstest enligt Pearson visades ett p-värde<0,001, vilket innebar att nollhypotesen att serierna korrelerar ej kan förkastas. För två områden med enskilda brukare, exempelvis Moldegatan 6 och Moldegatan var motsvarande R^2-värde 0,65 och p-värde<0,001.



Figur 35. Korrelationen mellan några studerade mätserier i projektet. Figuren visar på en tydligare korrelation mellan de större distributionsområdena än med de små. Siffrorna i den övre delen av grafen visar R2-värdet mellan varje mätserie samt korrelationstestets p-värde, där små värden bejakar nollhypotesen, att det finns korrelation.

4.6 KVALITETSGRANSKAD DATA

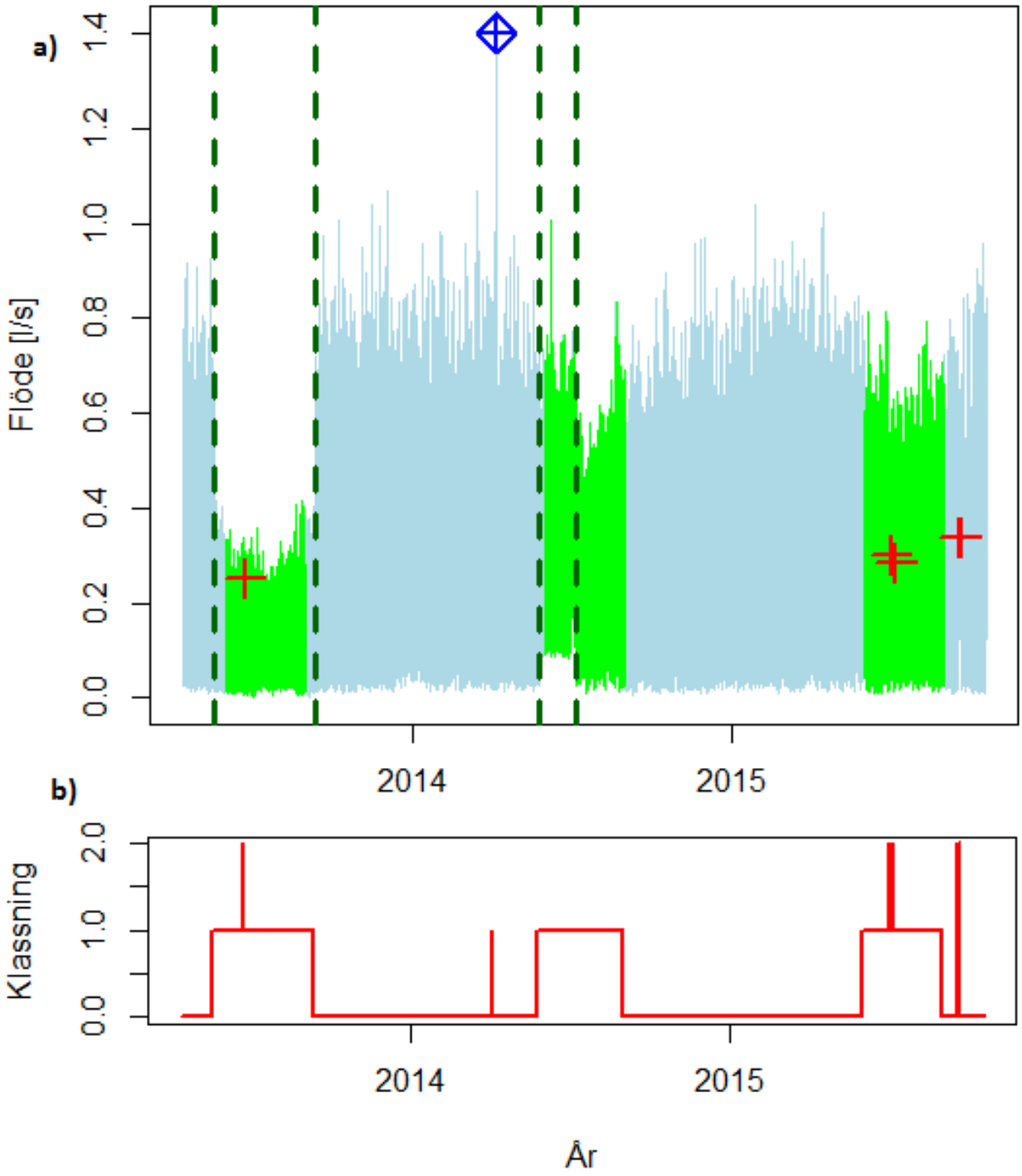
Det slutgiltiga resultatet från kvalitetsgranskningen presenteras som en figur där samtliga flaggade varningar visas. Utifrån de flaggade observationerna konstruerades en ny tidsvektor som anger klassningen för rådatat med vilka observationer vilka observationer som klarade sig godkända genom valideringsprocessen samt vilka som flaggades som suspekta eller felaktiga (



Figur 36).

Klassningen var enligt följande:

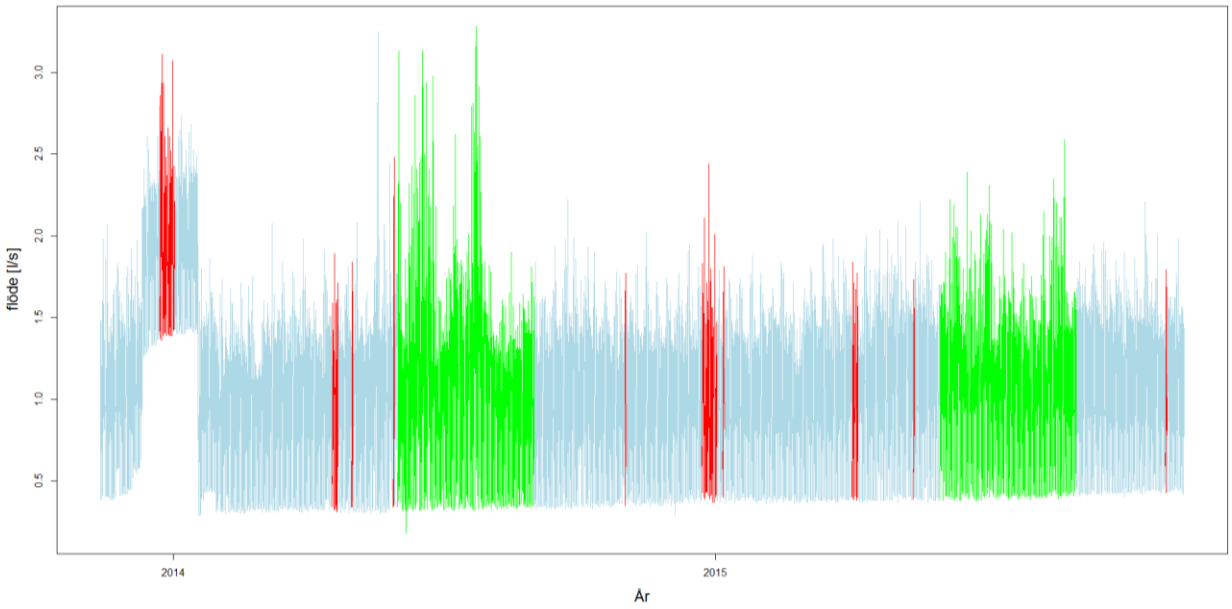
- Godkänd observation, tilldelades värdet 0
- Suspekt observation, tilldelades värdet 1
- Felaktig observation, tilldelades värdet



Figur 36 – a) Samtliga identifierade varningar för Vejlegatan 6. Streckade gröna linjer anger abrupta förändringar. Ljusgröna områden anger förändrat flöde under sommaren. Röda kryss anger upprepade värden och den blåa punkten är en detekterad outlier. Orangea linjerna är observationer som helt saknades. b) Kvalitetsklassningen för de tester som har gjorts. Värdet 0 innebär att observationen är godkänd, 1 att den suspekt och 2 att den felaktig.

4.6.1 Jämförelse med säsongstillhörighet

En enkel analys gjordes även för att studera om observationer tillhörde en storhelg eller en sommarmånad (Figur 37). I figuren kan ses att lokala extrempunkter uppstår kring julafton samt på sommaren. Detta är därför något som behöver tas i beaktande vid analys av identifierade outliers.



Figur 37. Mätserien för Alster. Gröna linjer visar observationer som har inträffat under sommaren och röda linjer observationer som påträffats under storhelger.

5 DISKUSSION

I detta projekt har data som härstammar från flödesmätare kopplat till vattenledningsnätet studerats. De metoder som har utnyttjats i detta projekt för att finna outliers har baserats på analys av residualerna från säsongskorrigerad data, där ett tröskelvärde sattes vid tillräckligt stor avvikelse från medelvärdet. Detta är en metod som diskuteras av Kiware (2009), som menar att det är en bristfällig metod för data som inte kan antas vara normalfördelad och därmed riskerar att felaktigt markera outliers. I detta projekt användes emellertid denna metod ändå med argumentet att residualerna från säsongs- och trendrensad data var approximativt normalfördelade. Andra studier (Quevedo m.fl., 2010) har använt ARIMA-modellering för att skapa avancerade modeller av flödesdatat och funnit outliers och anomalier från modellfelet. Denna metod föreföll ge tillförlitliga resultat men användes i denna studie då kalibreringen av modellen krävde goda kunskaper och erfarenhet för att välja rätt parametrar till ARIMA-modellen.

Vidare har det i detta examensarbete endast studerats mätserier bestående av inflödet till mätområdena. Om utflöde funnits tillgängligt hade en massbalans kunnat utnyttjas i valideringsprocessen. Om ytterligare data såsom pumpdata för ledningsnätet till distributionsområdet gjorts tillgänglig hade en klusteranalys kunnat göras för att finna både outliers och anomalier (Kiware, 2009).

Syftet med säsong- och trendrensning av mätdata som har gjorts i detta projekt har gjorts för att finna outliers från residualerna samt detektera förändringspunkter i variansen. Andra studier som gjort har utnyttjat resulterade komponenter för att med lyckade resultat skapa statistiska modeller, där vattenförbrukning jämförs med annan data såsom nederbörd, temperatur, ekonomiska aspekter samt befolkningstillväxt (Zhang m.fl., 2014). Syftet med dessa studier var att kartlägga trender för framtida dimensionering av vattenledningsnät i Kina. Detta kan vara högst intressant för det pågående SVU-projektet.

För att med säkerhet kunna säga att en observation markerad som outlier var ett felaktigt värde som behövde rensas från datasetet behövdes ytterligare information kopplad till den observationen. I vissa fall fanns sådan information tillgänglig, såsom data på inträffade rörbrott eller iakttagelser när flödesmätarna summerat flöden i enskilda observationer. Dessutom kunde det i vissa fall ses att observationer kopplade till sommarmånaderna samt julafton kunde visa lokala extrempunkter. För att kunna förklara extrempunkters riktighet hade ytterligare information kunnat vara användbar. Denna typ av information kan vara i form av iakttagelser eller erfarenhet vad extremvärdena kan bero på. Exempelvis bakspolning av ledningar och anordningar som kan ge upphov till lokala extremvärden. (Referensgruppmöte 1, 2016). Även information om bränder, då stora uttag av vatten sker hade varit intressant för att förklara outliers. Den information som fanns att tillgå i detta projekt var endast, i vissa fall, data på rörbrott samt tiden för varje observation, vilket möjliggjorde analys för vilken säsongstillhörighet en observation hade. Julafton visade sig exempelvis ofta ge relativt stora extremvärden. Sammanfattningsvis kan sägas att de observationer som identifierats som outliers behöver genomgå ytterligare analyser av en expert för att med säkerhet kunna förkastas som felaktiga mätvärden.

Att identifiera förändringar av medelvärdet i mätserierna visade sig vara ett praktiskt verktyg för att automatiskt kunna detektera anomalier i mätserierna. Av dessa gav metoden segment neighborhood från R paketet changepoint bra resultat, men att beräkningstiden var väldigt lång. Därför ansågs funktionen PELT vara att rekommendera för den här typen av analyser, då den gav likvärdiga resultat men till en betydligt kortare beräkningstid. Det har tidigare gjorts studier som konstaterat att PELT är en snabbare algoritm än segment neighborhood, då beslutsträdet som används i programmering kan förkortas avsevärt i PELT (Killick m.fl., 2012). Även de analyser som gjordes från paketet cpm gav tillförlitliga resultat. För de metoder som utnyttjades i detta projekt gjordes dock endast sökningar efter en förändringpunkt åt gången, vilket innebar att varje ny sekvens var en sökning ska ske måste definieras manuellt. Noterbart är att det i paketet cpm även finns metoder för att göra så kallade fas 2 sökningar efter abrupta förändringar i mätserien. Detta är en typ av kontinuerlig mätning, vilket innebär att nya förändringspunkter söks baserat på statistik från äldre observationer. Flera förändringspunkter kan därmed detekteras genom endast en initial formulering av parametervalet i algoritmen (Ross, 2015). Denna metod användes emellertid inte i detta projekt.

Flera förändringar i variansen kunde också detekteras, framförallt under sommaren. Detta beror troligtvis inte på mätfel. Att vattenförbrukningen förändras under sommaren har oftast naturliga orsaker såsom ökad trädgårdsbevattning och in- och utpendling från distributionsområdet (Referensgruppmöte 1, 2016). Det ansågs ändå vara intressant att definiera de perioder där vattenförbrukningen genomgår stora förändringar.

Anomalier kunde även detekteras genom att studera den resulterade trendkomponenten från säsongsrensning med STL. Problemet med denna lösning var att detektion behövde studeras okulärt för avgöra om det var ett faktiskt fel. Det var också uppenbart att säsongsrensningen med LOESS inte var den ultimata lösningen då det återstod en tydlig periodicitet i mätserien efter rensningen.

Analysen av de parallellkopplade mätarna visade att då flödesdata från mätarna plottades mot varandra framkom ibland flera linjära trender. Vid närmare analys av detta fenomen framgick att det härstammar från vissa sammanhängande sekvenser i mätserien, ofta endast något dygn, där den ena mätaren registrerat än den andra en viss konstant andel av mätarens flödesdata. Då detta härstammar från att något är onormalt, exempelvis genom ett mätfel eller genom mänsklig inblandning (exempelvis att avdelningsröret stängs av helt eller delvis eller att mätaren kalibrerats om), beslutades att alla observationer kopplade till detta fenomen flaggades. Dessa observationer behöver genomgå ytterligare analys, antingen genom att undersöka varför mätningen ser ut som den gör eller ytterligare analyser för att avgöra om den andra mätarens värden är korrekta. I så fall finns möjligheten att interpolera de felaktiga värdena utifrån värdena på den mätare som visar korrekt för att erhålla en mer komplett tidsserie, eftersom mätarna rimligtvis bör ge proportionella värden.

Av metoder som användes i det här projektet för att reglera för periodicitet kan det kort sägas att alla hade brister för denna typ av data. Detta beror antagligen på att det i mätserierna finns en komplex periodicitet. Nyttan för just det här projektet att göra någon säsongsrensning kan diskuteras, eftersom både STL och säsongsrensning med hjälp av glidande medelvärde såg ut att

ha kvarvarande periodicitet och den mer komplexa metoden TBATS krävde mer kunskap för kalibreringen av modellparametrarna samt hade väldigt lång beräkningstid för de mätserier som användes i projektet. Den enkla differentierade lösningen som användes visade sig fungera tillräckligt bra för de analyser som gjordes, nämligen att finna förändringar i variansen. Dessutom kunde outliers detekteras från en enkel LOESS-modell, vilken inte krävde något parameterval för säsongsrensning överhuvudtaget. För framtida analyser och prognostisering av framtida vattenförbrukning kan det emellertid vara av intresse att känna till mer komplexa metoder för reglering av mätseriernas periodicitet (De Livera m.fl., 2011).

6 SLUTSATS

På grund av det stora antalet mätserier som ämnas ligga till grund för SVU-projektets analyser och därför behöver kvalitetsgranskas, har det i detta examensarbete prioriterats att finna metoder som detekterar anomalier och outliers automatiskt. De flesta av de metoder som har utnyttjats har lyckats väl i denna ambition då endast ett initialt parameterval har behövts göras. Vissa metoder har emellertid krävt en del manuell kalibrering och granskning för att nå tillförlitliga resultat. En annan aspekt som tagits hänsyn till är den beräkningstid olika algoritmer har tagit till anspråk. Detta är viktigt därför att även om en metod söker igenom mätserierna automatiskt utan tidskrävande inblandning i kalibreringen och därtill levererar exakta resultat, så är det ändå inte lönsamt om det sker till priset av långa beräkningstider. Den totala tidsåtgången att använda sagda metod på alla mätserier hade helt enkelt blivit för stor.

Det här examensarbetet lyckades finna metoder som med tillförlitliga resultat kunde identifiera sekvenser i mätserierna där anomalier förekom. Det visade sig att genom att studera när förändringar i fördelningen sker kunde flera anomalier detekteras. Vidare kan nämnas att vid detekteringen av förändringspunkter var slutsatsen inte alltid att det berodde på effekten av ett mätfel. Men det ansågs att det ändå var av intresse att bryta ut dessa sekvenser från mätserien, då de statistiska egenskaperna för dessa sekvenser var annorlunda

Skurlängdskodningen var ett tillförlitligt verktyg för markera sekvenser där observationer upprepades. Dessa beror mest sannolikt på ett mätarfel, men eftersom vissa av mätvärdena i tidsserierna hade för få värdesiffror finns det en möjlighet att det exempelvis under natten, där den lokala variationen är som lägst, uppstår korrekta upprepade värden. För att reglera för detta söktes emellertid i det här projektet endast efter upprepade värden som var minst tre i följd. För att minimera risken att korrekta mätvärden flaggas som oriktiga hade emellertid ett ännu högre tröskelvärde kunnat väljas.

För många av de observationer som erhöll flaggningen att de var suspekta behöver ytterligare analyser göras för att konstatera huruvida den flaggade observationen berodde på ett mätfel eller ej. Detta behöver göras av någon på plats med bättre vetskap om mätarnas inställningar samt platsspecifika händelser eller av en expert med erfarenhet inom ämnesområdet. För analysen hade det underlättat om ytterligare information kring specifika händelser knutna till varje mätserie hade funnits. Detta skulle kunna vara i form av utökad data angående rörbrott eller liknande händelser som kan påverka en mätseries utseende kraftigt. Det var således ofta ett val som fick tas i den här studien om en flaggad observation berodde på en händelse som inte ska styra dimensioneringen av ledningsnätet och därmed skulle tas bort från mätserien eller inte. Detta gör att de tröskelvärden som ansätts för att flagga extremvärden blir ganska godtyckliga och många riktiga värden riskerar att flaggas. Det hjälper emellertid till att upptäcka observationer som behöver genomgå särskilda analyser av en expert.

För fortsatta analyser som görs på dataseten rekommenderas att dessa analyser görs både med avseende på rensad data och rådata, för att kunna se hur extremvärdena påverkade. Därtill kan det som sagt vara bra att detaljstudera orsaken till de abrupta förändringar som detekterats. Även om dessa inte beror på ett mätfel finns det antagligen en anledning till att fördelningen plötsligt genomgår en förändring, vilken kan vara intresse att känna till.

7 REFERENSER

7.1 LITERATURHÄNVISNINGAR

Acevedo, F. (2012). *Data analysis and statistics for geography, enviromental science, and engineering.* London: Taylor & Francis.

Auger, I. E., & Lawrence, C. E. (1989). Algorithms for the Optimal Identification of Segment Neighborhoods. *Bulletin of Mathematical Biology, 51*(1), 39-54.

Brockwell, P. J., & Davis, R. A. (1991). *Time series : Theory and methods* (2 uppl.). Springer series in statistics.

Cleveland, R. B., Cleveland, W. S., Mcrae, J. E., & Terpenning, I. (1990). STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *Journal of Official Statistics, 6*(1), 3-73.

De Livera, A. M., Hyndman, R. J., & Snyder, R. D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association, 106*(496), 1513-1527.

De Nadai, M., & van Someren, M. (2015). Short-term anomaly detection in gas consumption through ARIMA and Artificial Neural Network forecast. *Environmental, Energy and Structural Monitoring Systems (EESMS)* (ss. 250-255). IEEE Workshop on, Trento, 2015.

Ekwall, J. (2016). *Kvalitetssäkring av vattenförbrukningsdata - Ursprung till Fel och Osäkerheter i mätdata – en fallstudie av Göteborgs kommun.* Opublicerat manuskript. Stockholm: KTH.

Grubbs, F. E. (1969). Procedures for Detecting Outlying Observations in Samples. *Technometrics, 11*(1), 1-21.

Gustavsson, A. (2009). *Elpriserna på den nordiska elbörsen - Prognosmodellering med hjälp av ARIMA-modeller.* Umeå universitet, Samhällsvetenskapliga fakulteten, Umeå. Diva: diva2:325840

Haynes, K., Eckley, I. A., & Fearnhead, P. (2014). *Efficient penalty search for multiple changepoint.* Lancaster University, Department of Mathematics and Statistics. arXiv:1412.3617v1

IPCC. (2000). *Good Practice Guidance and Uncertainty Management in National Greenhouse Gas Inventories.* Intergovernmental Panel on Climate Change.

Kendall, M., & Stuart, A. (1983). *The Advanced Theory of Statistics* (Vol. 3). Griffin.

Killick, R., & Eckley, I. A. (2014). changepoint: An R Package for Changepoint Analysis. *Journal of Statistical Software, 58*(3).

Killick, R., Eckley, I. A., & Fearnhead, P. (2012). Optimal detection of changepoints with a linear computational cost. *JASA, 107*(500), 1590-1598.

Kiware, S. S. (2010). *Detection of Outliers in Time Series Data.* Marquette University. Hämtat från http://epuplications.marquette.edu/thesis_open/48

Lidström, V. (2013). *Vårt Vatten.* Svenskt Vatten AB.

Metcalfe, A. V., & Cowpertwait, P. (2009). *Introductory Time Series With R.*

Morandat, F., Hill, B., Osvald, L., & Vitek, J. (2012). *Evaluating the design of the R language objects and functions for data analysis.* Springer.

Muenchen, R. A. (2015). *The Popularity of Data Analysis Software.* Hämtat från http://r4stats.com/articles/popularity/ den 21 03 2016

NIST/SEMATECH. (den 30/10-2013). *e-Handbook of Statistical Methods*. Hämtat från http://www.itl.nist.gov/div898/handbook [22/04-2016]

Näsman Melander, E. (2012). *Dimensionering av åtgärder i kombinerade ledningssystem vid ökad spillvattenbelastning.* Uppsala Universitet, Institutionen för geovetenskaper. UPTEC W12 016

Quevedo, J., Puig, V., Cembrano, G., Blanch, J., Aguilar, J., Saporta, D., . . . Molina, A. (2010). Validation and reconstruction of flow meter data in the Barcelona water distribution network. *Control Engineering Practice, 18*(6), 640-651.

Ross, G. J. (2015). Parametric and Nonparametric Sequential change detection in R: The cpm Package. *Journal of Statistical Software, 66*(3), 1-20.

Ross, G. J., & Adams, N. M. (2012). Two Nonparametric Control Charts for Detecting Arbitrary Distribution Changes. *Journal of Quality Technology, 44*(2).

Scott, A. J., & Knott, M. (1974). A Cluster Analysis Method for Grouping Means in the Analysis of Variance. *Biometrics, 30*(3), 507-512.

Smith, S. W. (2001). *The Scientist and Engineer's Guide to Digital Signal Processing.* Hämtat från http://www.dspguide.com/ den 05 02 2016

SVU projektförslag. (2015). *Studie om dimensioneringstal för vattenförbrukning.* Stockholm: Svenskt Vatten.

Tiberg, C., Back, P. E., Ohlsson, Y., Carling, M., & Berggren Kleja, D. (2014). *Kvalitetssäkring av ämnesdata för beräkning av hälsoriskbaserade riktvärden för förorenad mark.* Linköping: tatens geotekniska institut, SGI.

Ullén, P., & Abdu, M. (2014). *Dimensionerande vattenförbrukning och dess variationer.* KTH.

VAV P100. (2009). *Kallvattenmätare.* Stockholm: Svenska Vatten- och Avloppsverksföreningen, VAV.

VAV P83. (2001). *Allmänna vattenledningsnät: Anvisningar för utformning, förnyelse och beräkning.* Stockholm: Svenska Vatten- och Avloppsverksföreningen, VAV.

Zainol, M. S., Zahari, S. M., Ibrahim, K., Zaharim, A., & Sopian, K. (2010). Additive Outliers (AO) and Innovative Outliers (IO) in GARCH (1, 1). *RECENT ADVANCES in APPLIED MATHEMATICS*, 471-479.

Zhang, D., Guangheng, N., Zhentao, C., Tie, C., & Tong, Z. (2014). Statistical interpretation of the daily variation of urban water consumption in Beijing, China. *Hydrological Science Journal*, 181-192.

7.2 DATORPROGRAM

Microsoft Excel (2013)

R Core Team, (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

7.2.1 R-paket

Dowle, M., Srinivasan, A., Short, T., Lianoglou S., med bidrag från Saporta, R. och Antonyan, E. (2015). data.table: Extension of Data.frame. R package version 1.9.6.

Hyndman R., J., (2015). forecast: Forecasting functions for time series and linear models. R package version 6.2

Killick R., Haynes K. and Eckley I.A. (2015). changepoint: An R package for changepoint analysis. R package version 2.2

López-de-Lacalle, J., (2015). tsoutliers: Detection of Outliers in Time Series. R package version 0.6.

Rmetrics Core Team, Wuertz, D., Setz, T., Chalabi, Y., Maechler, M. and Byers, J.W., (2015). timeDate: Rmetrics - Chronological and Calendar Objects. R package version 3012.100

Wickham H., (2009). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York

7.3 PERSONLIG KOMMUNIKATION

Referensgruppsmöte 1, Tyréns AB, Stockholm. Möte 18 mars 2016

Törneke, K.,VA-utredare, Tyréns Stockholm, Möte 17 december 2015

BILAGA

Här presenteras delar av den R-kod som användes i projektet

```
######### Examensarbete ############
#Kvalitetsgranskning av mätdata från ledningsnätet
#Av: Victor Eliasson, Mars 2016
# Civilingenjör i Miljö- och vattenteknik, Uppsala Universitet
#####Set working directory#####
setwd("D://skola/exjobb/r/All_data2")
#####Laddar data#####
library(timeDate)
temp <- list.files(pattern = "*.csv")
ladda <- function(temp){datum <- read.csv2(temp, stringsAsFactors = FALSE)
datum$Time <- as.timeDate(datum$Time)
datum
datum$Flow <- as.numeric(datum$Flow)
datum
}

alldata <- lapply(temp, ladda)
names(alldata) <- temp
##### Skurlängdskodning #####

library(data.table) #Laddar paket "data table"
y <- alldata[[1]] #Väljer mätserie
n <- 1:length(y[,2])
x<- data.frame(flow=y[,2],index=n) #Skapar tillfällig data

x$Run_seq <- rleidv(x[,1])
x$run_ol <- NA;
for (i in 1:(length(x[,2])-1)) {
  if (x[i,3]== x[(i+1),3] && x[i,3]== x[(i-1),3]){
    x[i,4] <- x[i,1]
    x[(i+1),4] <- x[(i+1),1]
    x[(i-1),4] <- x[(i-1),1]
  }
}
plot(x[,2],x[,1], type = "l")
points(x[,2], x[,4], col="red")

##### Spektralanalys #####
x <- na.omit(y[,2])
k <- kernel("daniell", 4)  #kernel-smoothing
```

```
spec.pgram(x,k,taper = 0,log="no", main="Peridogram", cex.main=2.5, cex.lab=1.5)
#periodogram
##### Säsongsrensning med MA #####
serie.ts <- ts(x, frequency=24) ### Skapar tidsserie med frekvensen 24h
x.dec <- decompose(serie.ts,type ="additive" )
plot(x.dec, main="MA")
##### Säsongsrensningning med STL
serie.ts <- ts(x, frequency=24) ### Skapar tidsserie med frekvensen 24h
x.stl <- stl(serie.ts, "periodic", robust = TRUE)
outliers <- which(x.stl$weights<1e-8)
op <- par(mar=c(0,4,0,3), oma=c(5,0,4,0), mfcol=c(4,1))
plot(x.stl, set.pars =NULL,main="STL")
temp <- x.stl$time.series
points(time(temp)[outliers], 1*temp[,"remainder"][outliers],pch="x",col="red")
par(op)
##### Hittar anomalier från resulterande komponenter av STL #####
tr <- x.stl$time.series
tr <- tr[,2]
plot(tr, main="Trenden från STL")
quantile(tr, probs = seq(0,1,.01)) ## Ger värden för 0,05% och 95,5% av datat
abline(h=c(9.45,18.6), col="red") ## Lägger in tröskelvärde
abline(h=c(8.59,20.23), col="red")

n=1:length(tr)
tr2 <- data.frame(tr)
tr2$nr <- NA
tr2$p <- NA
for(i in 1:length(n)){
  if (tr[i]>20.2315087 | tr[i]< 8.5932609){
    tr2[i,2]<-n[i]
    tr2[i,3] <- x[i]
  }
}
plot(tr)
abline(v=(tr2[,2]/24), col="blue")
##### Säsongsrensning med TBATS #####
library(forecast)
serie.ts <- msts(x, seasonal.periods=c(12,24,168,8766)) ### Skapar tidsserie med sfrekvenserna
12h,24h veckovis samt årligen
x.tbats <- tbats(serie.ts, use.parallel=FALSE)
plot(x.tbats, main="TBATS")
resid <- residuals(x.bats) ## Residualerna för TBATS-modellen
qqnorm(resid) ## Skapar QQ-plott
qqline(resid)
```

```
##### Differentierar mätserien #####
x.diff <- diff(x,lag=24)
plot(x.diff, type="l", main="Differentierad mätserie map lagg 24")
##### plottar alla områden #####
###################### FOR-Loop, plot för alla gator, för jämförelse med säsong, storhelger,
tid på dygnet osv ########################################
timdat2 <- read.csv2("timdat_korr.csv")

###################### Lägger till kolumner för datum, veckodag, veckonummer etc.
################
timdat2$tiddate <- format(as.timeDate(timdat2$TimeDate), "%Y-%m-%d %H:%M")
timdat2$wday <- as.POSIXlt(timdat2$tiddate)$wday
timdat2$wnum <- format(as.Date(timdat2$tiddate), "%U")
timdat2$dat <- format(as.Date(timdat2$TimeDate), "%Y-%m-%d")
timdat2$year <- format(as.Date(timdat2$TimeDate), "%Y")
timdat2$tid <- format(as.timeDate(timdat2$TimeDate), "%H:%M")
timdat2$datum <- format(as.timeDate(timdat2$TimeDate), "%m-%d")

for (i in 5:25) {
  plot(timdat2[,i], main=colnames(timdat2[i]), ylab="Flow", col="blue", type="l")
  ### Scatterplot ###
  x <- data.frame(year = as.vector(timdat2$year), flow = as.vector(timdat2[,i]), datum
=as.vector(timdat2$datum), season = as.vector(timdat2$season_index), holidays=
as.vector(timdat2$Holiday_index), timeperiod = as.vector(timdat2$time_index), vecka =
as.vector(timdat2$wnum), wday=as.vector(timdat2$wday))
  x <- as.data.frame(na.omit(x))
  p <- qplot(timeperiod, flow, data = x, col="flow", facets = year~season,
main=colnames(timdat2[i]), xlab = "datum",ylab = "flowrate")
  plot(p+theme_bw())
  ### Boxplot ###
  p <- ggplot(x, aes(holidays,flow,group=holidays))+geom_boxplot()
  p <- p+labs(main=colnames(timdat2[i]))
  plot(p + theme(axis.text.x=element_text(angle=-90)))
  ### Denistyplot ###
  p <- qplot(flow, data = x, geom = "density", fill= as.factor(timeperiod),
alpha=I(.5),main=colnames(timdat2[i]), xlab = "flowrate",ylab = "Density")
  plot(p+theme_bw())
  ### Jämförelse, scatterplot ###
  p2 <- qplot(datum, flow, data = x, col=timeperiod, facets = year~season,
main=colnames(timdat2[i]), xlab = "vecka",ylab = "flowrate")
  plot(p2+theme_bw()+geom_point(data = x, aes(x=datum,y=flow, shape=factor(holidays))))
}
##### Change detection i medelvärde med "Changepoint-paketet #####
library(changepoint)
```

```
x.PELT.mean=cpt.mean(x,pen.value=c(log(length(x)),100*log(length(x))),penalty =
"CROPS",method = "PELT") ## change points med Pelt
cpts.full(x.PELT.mean)  # Visar beräknade segment enligt angivet pen.value
pen.value.full(x.PELT.mean)
plot(x.PELT.mean,diagnostic=TRUE)
plot(x.PELT.mean, ncpts=8) #PLott med antalet change points angivet

x.SN.mean=cpt.mean(x,penalty = "AIC",method = "SegNeigh") ## change points med Segment
neighborhood och 20 segment
plot(x.SN.mean,diagnostic=TRUE)
plot(x.SN.mean, ncpts=10) #PLott med antalet change points angivet

x.BS.mean=cpt.mean(x,penalty = "AIC",method = "BinSeg", Q=20) ## change points med binär
segmentering och 20 segment
plot(x.BS.mean,diagnostic=TRUE)
plot(x.BS.mean, ncpts=10) #Plott med antalet change points angivet
##### Skapar rekursivt filter, delta måste vara mellan 0-1 #####
f1 <- filter(x, filter = 0, method = "recursive")
f2 <- filter(x, filter = 0.3, method = "recursive")
f3 <- filter(x, filter = .7, method = "recursive")
f4 <- filter(x, filter = .9, method = "recursive")
op <- par(mfrow = c(2,2))
plot(f1, main = "f delta = 0")
plot(f2, main = "f delta = 0.3")
plot(f3, main = "f delta = 0.7")
plot(f4, main = "f delta = 0.9")
par(op)
filter.pelt=cpt.mean(f4,pen.value=c(log(length(f4)),100*log(length(f4))),penalty =
"CROPS",method = "PELT") ## change points med Pelt
plot(filter.pelt,diagnostic=TRUE)
plot(filter.pelt, ncpts=8) #Plott med antalet change points angivet

##### Change detection, location shift med cpm #####
library(cpm)

t <- detectChangePointBatch(x, cpmType = "Mann-Whitney",
                      alpha = 0.001, lambda = 0.3)
plot(x, type="l")

if (t$changeDetected)
  abline(v = t$changePoint,col="red", lty = 2)
plot(t$Ds)
abline(h=t$threshold, lty=2,col="red")
n <- length(x)
```

```
seg1 <- x[1:t$changePoint]
seg2 <- x[t$changePoint:n]

t2 <- detectChangePointBatch(seg1, cpmType = "Mann-Whitney",
                      alpha = 0.001, lambda = 0.3)
plot(seg1, type="l")
if (t2$changeDetected)
  abline(v = t2$changePoint,col="red", lty = 2)
plot(t2$Ds)
abline(h=t2$threshold, lty=2,col="red")

seg1.2 <- seg1[1:t2$changePoint]
seg1.3 <- seg1[t2$changePoint:length(seg1)]

t2.1 <- detectChangePointBatch(seg1.2, cpmType = "Mann-Whitney",
                       alpha = 0.005, lambda = 1)
plot(seg1.2, type="l")
if (t2.1$changeDetected)
  abline(v = t2.1$changePoint,col="red", lty = 2)
plot(t2.1$Ds)
abline(h=t2.1$threshold, lty=2,col="red")

t2.2 <- detectChangePointBatch(seg1.3, cpmType = "Mann-Whitney",
                       alpha = 0.005, lambda = 1)
plot(seg1.3, type="l")
if (t2.2$changeDetected)
  abline(v = t2.2$changePoint,col="red", lty = 2)
plot(t2.2$Ds, type="l")
abline(h=t2.2$threshold, lty=2,col="red")

t3 <- detectChangePointBatch(seg2, cpmType = "Mann-Whitney",
                      alpha = 0.001, lambda = 0.3)
plot(seg2, type="l")
if (t3$changeDetected)
  abline(v = t3$changePoint,col="red", lty = 2)
plot(t3$Ds)
abline(h=t3$threshold, lty=2,col="red")

seg2.1 <- seg2[1:t3$changePoint]
seg2.2 <- seg2[t3$changePoint:length(seg2)]

t3.1 <- detectChangePointBatch(seg2.1, cpmType = "Mann-Whitney",
                       alpha = 0.005, lambda = 1)
plot(seg2.1, type="l")
```

```
if (t3.1$changeDetected)
  abline(v = t3.1$changePoint,col="red", lty = 2)
plot(t3.1$Ds)
abline(h=t3.1$threshold, lty=2,col="red")

seg2.1.1 <- seg2[1:t3.1$changePoint]
seg2.1.2 <- seg2[t3.1$changePoint:length(seg2.1)]
t3.1.1 <- detectChangePointBatch(seg2.1.1, cpmType = "Mann-Whitney",
                          alpha = 0.1, lambda = 1)
plot(seg2.1.1, type="l")
if (t3.1.1$changeDetected)
  abline(v = t3.1.1$changePoint,col="red", lty = 2)
plot(t3.1.1$Ds)
abline(h=t3.1.1$threshold, lty=2,col="red")

t3.1.2 <- detectChangePointBatch(seg2.1.2, cpmType = "Mann-Whitney",
                          alpha = 0.1, lambda = 1)
plot(seg2.1.2, type="l")
if (t3.1.2$changeDetected)
  abline(v = t3.1.1$changePoint,col="red", lty = 2)
plot(t3.1.2$Ds)
abline(h=t3.1.2$threshold, lty=2,col="red")

seg2.1.2.1 <- seg2.1.2[1:t3.1.1$changePoint]
seg2.1.2.2 <- seg2.1.2[t3.1.1$changePoint:length(seg2.1.2)]
t3.1.3 <- detectChangePointBatch(seg2.1.2.2, cpmType = "Mann-Whitney",
                          alpha = 0.1, lambda = 1)
plot(seg2.1.2.2, type="l")
if (t3.1.3$changeDetected)
  abline(v = t3.1.3$changePoint,col="red", lty = 2)
plot(t3.1.3$Ds)
abline(h=t3.1.3$threshold, lty=2,col="red")

t3.1.4 <- detectChangePointBatch(seg2.1.2.1, cpmType = "Mann-Whitney",
                          alpha = 0.1, lambda = 1)
plot(seg2.1.2.1, type="l")
if (t3.1.4$changeDetected)
  abline(v = t3.1.4$changePoint,col="red", lty = 2)
plot(t3.1.4$Ds)
abline(h=t3.1.4$threshold, lty=2,col="red")

seg2.1.1.1 <- seg2.1.1[1:t3.1.1$changePoint]
t3.1.1.1 <- detectChangePointBatch(seg2.1.1.1, cpmType = "Mann-Whitney",
                           alpha = 0.1, lambda = 1)
```

```
plot(seg2.1.1.1, type="l")
if (t3.1.1.1$changeDetected)
  abline(v = t3.1.1.1$changePoint,col="red", lty = 2)
plot(t3.1.1.1$Ds)
abline(h=t3.1.1.1$threshold, lty=2,col="red")

t3.2 <- detectChangePointBatch(seg2.2, cpmType = "Mann-Whitney",
                        alpha = 0.1, lambda = 1)
plot(seg2.2, type="l")
if (t3.2$changeDetected)
  abline(v = t3.2$changePoint,col="red", lty = 2)
plot(t3.2$Ds)
abline(h=t3.2$threshold, lty=2,col="red")
###### Chnage detection, varians med cpm #####
k <- detectChangePointBatch(x.diff, "Mood", alpha = 0.05)
plot(x.diff, type="l", col="lightblue", xlab="",ylab="",xaxt="n")
abline(v=k$changePoint, col="red", lwd=2)
n <- length(x.diff)
k2 <- detectChangePointBatch(ii[1:k$changePoint], "Mood", alpha = 0.05)
abline(v=k2$changePoint, col="red", lwd=2)
k3 <- detectChangePointBatch(ii[k$changePoint:n], "Mood", alpha = 0.05)
abline(v=(k$changePoint+k3$changePoint), col="red", lwd=2)
k4 <- detectChangePointBatch(ii[(k$changePoint+k3$changePoint):n], "Mood", alpha = 0.05)
abline(v=(k$changePoint+k3$changePoint+k4$changePoint), col="red", lwd=2)
k5 <- detectChangePointBatch(ii[(k$changePoint+k3$changePoint+k4$changePoint):n],
"Mood", alpha = 0.05)
abline(v=(k$changePoint+k3$changePoint+k4$changePoint+k5$changePoint), col="red",
lwd=2)
k6 <-
detectChangePointBatch(ii[(k$changePoint+k3$changePoint+k4$changePoint+k5$changePoint)
:n], "Mood", alpha = 0.05)
abline(v=(k$changePoint+k3$changePoint+k4$changePoint+k5$changePoint+k6$changePoint),
col="red", lwd=2)
##### Change detection, varians med changepoint #####
## Använder funktionen cpt.var enligt samma metod som för cpt.mean
x.PELT.var=cpt.var(x.diff,pen.value=c(log(length(x.diff)),100*log(length(x.diff))),penalty =
"CROPS",method = "PELT") ## change points med Pelt
cpts.full(x.PELT.var)  # Visar beräknade segment enligt angivet pen.value
pen.value.full(x.PELT.var)
plot(x.PELT.var,diagnostic=TRUE)
plot(x.PELT.var, ncpts=8) #PLott med antalet change points angivet

x.SN.var=cpt.mean(x.diff,penalty = "AIC",method = "SegNeigh") ## change points med Segment
neighborhood och 20 segment
```

```
plot(x.SN.var,diagnostic=TRUE)
plot(x.SN.var, ncpts=10) #PLott med antalet change points angivet

## change points med binär segmentering och 20 segment
x.BS.mean=cpt.mean(x.diff,penalty = "AIC", method = "BinSeg", Q=20)
plot(x.BS.var,diagnostic=TRUE)
plot(x.BS.var, ncpts=10) #Plott med antalet change points angivet
##### Outlier detektion för parallellkopplade mätare #####
## laddar data
setwd("D://skola/exjobb/r/mult_meters")
library(timeDate)

temp <- list.files(pattern = "*.csv")

ladda <- function(temp){datum <- read.csv2(temp, stringsAsFactors = FALSE)
datum$Time <- as.timeDate(datum$Time)
datum
}

data3 <- lapply(temp, ladda)
names(data3) <- temp
##### Lägger till konfidensintervall
x <- data3[[1]]
x1 <- as.numeric(x$Meter1)
x2 <- as.numeric(x$Meter2)

x_kvot <- x1/x2
n <- 1:length(x_kvot)
x3 <- data.frame(kvot=x_kvot,index=n)
for (i in 1:length(n)) {
  if (is.na(x3[i,1])){}
  else if (x3[i,1]>1.1 | x3[i,1]<0.9){
    x3[i,3] <- x1[i]
  }
}

plot(x1,x2,ylim=c(0,.4),xlim=c(0,.4))
points(x3[,3],x2,col="red")

### Linjär Regresionsmodell
lrm <- lm(formula = x1~x2)
summary(lrm)
par(mfrow=c(2,2))
plot(lrm) ##markerat outliers med identify(x1,x2)
```