Thesis no: BCS-2015-03

Utvärdering av temporala analysmetoder inom brottskategorin bostadsinbrott

Olle Svenhag

Faculty of Computing

This thesis is submitted to the Faculty of Computing at Blekinge Institute of Technology in partial fulfillment of the requirements for the degree of Bachelor in Computer Science. The thesis is equivalent to 10 weeks of full time studies.

Contact Information:
Author:
Olle Svenhag
E-mail: olsv11@student.bth.se

University advisor:
PhD. Martin Boldt
Dept. Computer Science & Engineering

Faculty of Computing
Blekinge Institute of Technology
SE–371 79 Karlskrona, Sweden

Internet : www.bth.se
Phone : +46 455 38 50 00
Fax : +46 455 38 50 57

Abstract

Context. In year 2013 the number of reported residential burglaries in Sweden was 21000, where only 4-5 percent of those actually got solved [1]. The Swedish police is trying to improve their way of working to increase the number of solved burglaries, this by structuring the data collection and analysing with computer science methods. Temporal analysis is the key to figure out when crime actually takes place.

Objectives. This thesis study five different methods for analysing the temporal data of residential burglaries. The temporal analysis is performed on three time spans: time of day, day of the week and day of the month. The objective is to evaluate the five methods in the three time spans and decide which method is the most suitable for each of them.

Methods. This study includes three experiments testing all five methods on the three time spans. The experiments focus on comparing the observed data against the data of burglaries with a known specific time of the crime. In order to test the performance of each method a Chi-squared goodness-of-fit test was used, as well as a visual comparison of the produced plots.

Results. The results showed that the Aoristic-method was the most suitable method to use when analysing temporal data of residential burglars, if looking at the time of day, day of the week and day of the month. Using the methods we also generated plots of the three temporal distributions, with an R script.

Conclusions. We concluded that using the Aoristic-method is the most suitable method to use to generate plots from the temporal data. We also concluded that using this script with the Aoristic-method to generate plots, would make it possible for the police to resource allocation according to when burglaries actually take place.

Keywords: Temporal analysis, Aoristic, analysis methods, residential burglaries.

Contents

Chapter 1

Introduktion

Antalet anmälda brott i Sverige har under de senaste tio åren ökat med ungefär 16 procent, vilket är en ökning på nästan två miljoner [1]. Kategorimässigt ökar antalet brott inom trafikbrott, bedrägeribrott och narkotikabrott under denna period. De kategorier där antalet anmälda brott har minskat är bland annat skadegörelse, brott mot person och stöldbrotten. Inom kategorin stöld tillhör undergrupper så som personrån, butiksstöld, bostadsinbrott med flera. Under året 2013 anmäldes det omkring 21000 bostadsinbrott [1], vilket uppskattas vara en procent av alla brott som begås i Sverige. Bostadsinbrott är enligt Polisen i Sverige ett brott som upplevs mer kränkande än andra stöldbrott, då offren berörs mentalt eftersom att hemmet är en plats att känna sig trygg och har sina ägodelar [2]. Av dessa 21000 brott är uppklarningsfrekvensen för bostadsinbrotten låg, endast 4-5% av alla bostadsinbrotten klaras upp [1]. Detta kan delvis bero på att Polisen år 2013 bestod av 21 olika polismyndigheter, alla med olika sätt att arbete och hantera dessa brott. Lyckligtvis insåg Polisen problemet och håller nu på med en omställning år 2015 för att gå ihop till en gemensam polismyndighet [3].

Bostadsinbrott definieras som en stöld där gärningsmannen olovligen bereder sig tillträde till permanentbostad eller fritidshus, det vill säga när någon brutit sig in, eller på annat sätt tagit sig in utan lov [2]. 2011 påbörjades ett samarbete mellan forskare på BTH (Blekinge Tekniska Högskola) och Polisen i Stockholm, Göteborg och Skåne. Syftet med detta samarbete var att ta fram en standardiserad metod för att samla in uppgifter vid en brottsplatsundersökning för bostadsinbrott. Resultatet blev ett digitalt formulär. Innan samarbetet med BTH påbörjades så förde den individuella Polisen egna noteringar vid brottsplatsundersökning, utan att ha några förutbestämda standarder på dess innehåll. Polisens nya sätt att samla in och lagra information inom bostadsinbrott möjliggör för att med hjälp av ett It-baserat system ta fram och titta på trender bland bostadsinbrott [4]. Man har börjat se möjligheterna och fördelarna i att omfamna den digitala värld vi idag lever i. Idag finns det en sökfunktion där man kan gå in och titta på brott var för sig, man utnyttjar inte den stora mängden data som samlats in för att bygga statistik. Tiden då ett bostadsinbrott begås kan variera från exakt tidpunkt till flera dagar, tidsperioden från det att bostaden lämnades

obevakad tills det att brottet upptäckts.

Detta arbete undersöker hur medels datavetenskapliga metoder kan hjälpa Polisen till ökad kunskap kring fördelningen av bostadsinbrott över kort (dygn), medellång (vecka) och lång (månad) sikt. Givet denna ökade kunskap kommer Polisen bättre kunna allokera sina resurser för att möta de faktiska behov som finns ute i samhället. Viktiga frågor att besvara i projektet var vilken av de tillgängliga analysmetoderna lämpar sig bäst för tidsanalyser på kort, medellång, samt lång sikt inom brottskategorin bostadsinbrott. Även vilka (för Polisen) relevanta tidsanalyser kan automatgenereras med hjälp av en mjukvarukomponent baserad på analysmetoden från föregående fråga. Tidsanalys är ett komplext område, det finns ett antal metoder att välja på för att analysera datan.

1.1 Bakgrund

När ett brott har begåtts och Polisen anländer till brottsplatsen, registreras brottsplatsundersökningen med det tidigare nämnda formuläret en brottsplatsundersökning. Formuläret består av ett hundratal kryss rutor där Polisen fyller i till exempel uppgifter om bostaden, modus-operandi, preventiva åtgärder bostadsägaren vidtagit med mera. Formuläret är digitalt och framtaget av Polisen tillsammans med forskare vid Blekinge Tekniska Högskola (BTH). Vidare skickas det digitala formuläret genom Polisen till en databas som lagrar informationen. Polisen har sedan möjlighet att i ett tillhörande webbaserat system utföra avancerade sökningar baserat på diverse kriterier och tidsspann. Idag kan man alltså med hjälp av detta system titta på ett eller flera brott för att kartlägga, utföra jämförelser av kategorier så som larm fanns eller ej och så vidare.

Vidare vill Polisen få hjälp med att utveckla diverse tillägg som med hjälp av insamlad data kunna hjälpa dem att automatgenerera och analysera metoder för att titta på trender inom bostadsinbrott. Exempelvis för att lokalisera så kallade hotspots. Hotspots är ett eller flera tillfällen då data sticker ut från den normala mängden. Detta kan användas på en karta för att lokalisera ett utsatt område, eller för att titta på tidstrender som när på dygnet majoriteten av inbrotten begås. Polisens intresse och vilja att hitta nya metoder och förbättra sitt arbetssätt är det som motiverat projektet. Sverige måste följa med i den digitala utvecklingen och utforska de nya möjligheterna som kan finnas. Det finns stort intresse hos Polisen att ta del av projektet, då metoden som tas fram för bostadsinbrotten skulle kunna leda till framtida expansion inom andra brottskategorier, till exempel transportstölder, åldringsbrott samt olika typer av bedrägerier.

1.2 Syfte och avgränsning

I arbetet kommer vi att att med hjälp av data för bostadsinbrott ta fram statistik för hur dessa fördelar sig över dygn, vecka och månad. Polisen har som tidigare

nämnts samlat in data på ett standardiserat sätt vilket vidare ledde till ett intresse av att titta på statistik för den lagrade informationen. I arbetet genomförs implementering och utvärdering av fem temporala metoder för att analysera hur brott fördelar sig över ett visst tidsspann. Dessa metoder utvalda baserat på relaterade arbeten och rekommendationer från handledare och hanns kontakter. Syftet är att välja ut den metoden som efter statistiska tester och visuella jämförelser efterliknar den mängden brott som har bestämd exakt tidpunkt för inbrott. Dessa brott används alltså som s.k. "ground truth". Vidare att med hjälp av metoden ta fram viktig relevant statistik önskad av Polisen. Arbetet kommer att hålla sig inom kategorin bostadsinbrott eftersom det är för denna brottskategori vi har tillgång till viss brottsplatsinformation. Med den här informationen och visuella diagram kommer Polisen kunna effektivisera resursfördelning och har möjligheten att förebygga brott på ett helt annat sätt.

Chapter 2

Relaterat arbete

I boken Crime Analysis with Crime Mapping skriven av Rachel B. Santos behandlas bland annat relevanta metoder för spatial och temporalanalys av brott, det vill säga hur dessa fördelar sig såväl geografiskt som i tiden [6]. Boken går inte djupare in på att testa metoderna men beskriver hur dessa metoder fungerar och dess användningsområden. Santos tar även upp något som kallas geografiska hotspots, vilket är ett sätt att titta på utsatta områden med hjälp av ett koordinat baserat system. Den svenska Polisen är förutom att titta på tidpunkter då brott sker, även mycket intresserad av att ta fram dessa hotspots ifrån den insamlade brottsdatan. Santos menar för att kartlägga brott bör både tid och plats tas med i beräkningarna, vilket i detta projekt inte kommer att tas upp. Däremot testas de metoder hon tar upp tillsammans med några andra relevanta metoder för att lösa problemet med tidsanalys. Santos tar även upp problemen som finns med dagens system för att kartlägga brott, att majoriteten fortfarande använder sig av medel-metoden som enligt henne är missvisande. Det krävs nya system som använder metoder som bättre representerar hur brott fördelar sig över tid.

Jerry H. Ratcliffe och Michael J. McCullagh introducerade 1998 en temporal analysmetod som de kallade Aoristic. Denna metod går ut på att hela tidsperioder skall vägas in, det skall inte sättas en punkt som summerar ett helt tidsspann som till exempel ett medelvärde gör. Ratcliffe och McCullagh menar på att med hjälp av dessa metoder kan man övervaka förändringar i brott över tid, inom godtyckliga områden. Dem påpekar att brottsdata ofta saknar temporala definitioner och visar med hjälp av exempel på bilstölder hur man kan utföra detta. Ratcliffe och McCullagh tar fram ramverket för två temporala metoder att använda men dem påpekar ingen slutsats i vilken metod som lämpar sig bäst att använda. Även värt att nämna är deras tidiga upptäckt av att bruka dessa metoder för Hotspot analys för att geografiskt kartlägga trender i brott. I relaterat arbete som utförts inom temporal analys av brott har Matthew PJ. Ashby och Kate J. Bowers dragit slutsatsen att aoristisk och slump-metoden är de mest exakta för tidsscenariot timme per dygn [5]. Detta genom att titta på de olika metoderna som tidigare är nämnda för temporal analys och jämföra dessa mot data för exakta tidpunkter för brott. I deras forskning studerade de cykelstölder vid en station där det fanns övervakningskameror under en period, där dem med hjälp av dessa

kameror kunde få fram de exakta tidpunkterna för brotten att jämföra metoderna mot. Ashby och Bowers utför även statistiska tester på de olika resultaten från metoderna och konstaterar att deras resultat är signifikanta. Cykelstölder är bara en sort av de stöldbrotten som sker, metoderna behöver testas på andra områden så som bostadsinbrott vilket går nu att utföra då information samlats in på ett standardiserat sätt. Dessutom bör tester av metodernas exakthet testas på olika långa tidsspann, det vill säga dag per månad, dag per år och så vidare.

Chapter 3

Temporala analysmetoder

Grundat på litteraturstudien och rekommendationer från handledare valdes fem stycken lämpliga metoder ut. I arbetet finns några aspekter som förklaras nedan för att få förståelse för hur metoderna fungerar.

- Start tid - start tiden från tidsperioden.
- Slut tid - slut tiden från tidsperioden.
- Intervall - skillnaden mellan start tiden och slut tiden.
- Enhet - temporalanalysens upplösning, till exempel timmar eller dagar.

Brottstid definieras som ett tidsspann för möjliga tidpunkter då ett brott kan ha begåtts, exempelvis tiden mellan att målsäganden lämnade bostaden tills dess att denna återvände tillbaka för att upptäcka inbrottet. För att förklara metoderna används tre exempel på brottstider (se nedan) och en enhet på 30 minuter vilket ger 48 tidsspann per dygn.

1. Ett brott med känd exakt brottstid 10 maj klockan 19:38.
2. Ett brott med okänd exakt brottstid 10 maj klockan 08:15 till 17:28.
3. Ett brott med brottstid som löper över olika dagar 10 maj klockan 18:20 till 11 maj klockan 11:45.

3.0.1 Start-metoden

Start-metoden går ut på att använda sig utav den tidpunkt då tidsperioden startade som värde. Se figur 3.1 för exempel, den tidpunkt start-metoden använder markeras med en blå fyrkant.

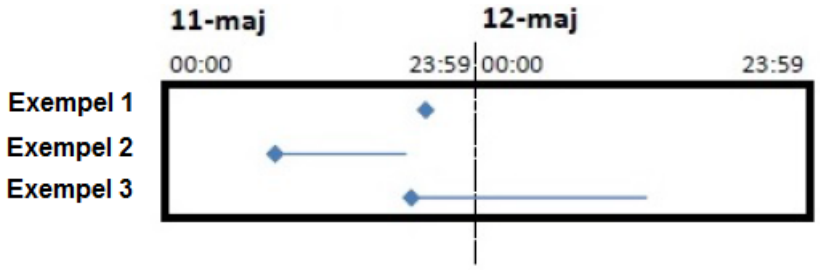

Figure 3.1: Förklaring av Start-metoden.

3.0.2 Slut-metoden

Slutmetoden fungerar på samma sätt som startmetoden med ändringen att tiden då tidsperioden slutade blir värdet. Se figur 3.2 för exempel, den tidpunkt slutmetoden använder markeras med blå fyrkant.

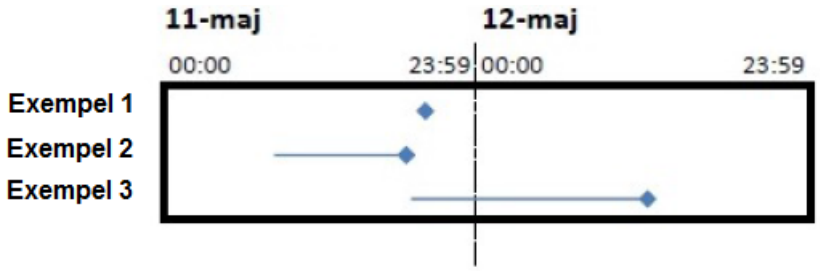

Figure 3.2: Förklaring av Slut-metoden.

3.0.3 Medel-metoden

Metoden där man använder medelvärdet är idag den mest använda metoden för att titta på tidstrender inom brott. Metoden går ut på att ta tiden som ligger mitt i mellan start och slut-tid för tidsperioden. Se figur 3.3 för exempel, den tidpunkt medel-metoden använder markeras med blå fyrkant.

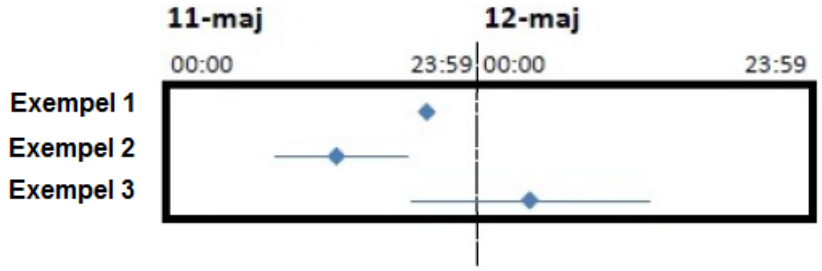

Figure 3.3: Förklaring av Medel-metoden.

3.0.4 Slump-metoden

Slumpmetoden går ut på att titta på hela tidsperioden för att sedan slumpvis välja ut en tidpunkt mellan start och slut-tiden för perioden. Se figur 3.4 för exempel, den tidpunkt slump-metoden använder markeras med en blå fyrkant.

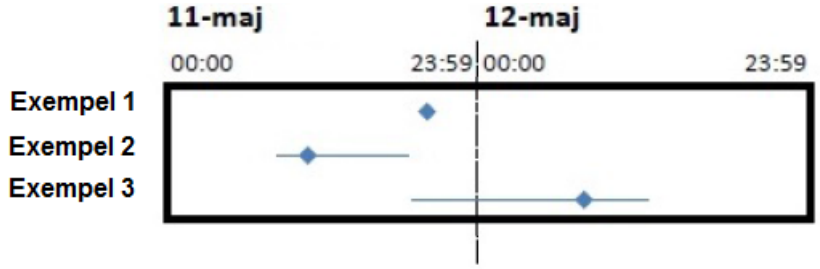

Figure 3.4: Förklaring av Slump-metoden.

3.0.5 Aoristiska-metoden

Metoden går ut på att hela tidsperioden skall tas med och istället för att lägga vikten på en specifik dag eller en specifik tidpunkt, fördelas vikten av tidsperioden i bråkdelar över hela tidsperioden. Till exempel ett brott med start tid klockan åtta och slut tid klockan tio, med en enhet på 30 minuter. Brottet delas då upp i fjärdedelar för att slutligen föras in i statistiken i de fyra enheterna om 30-minuter vardera. Se figur 3.5 för exempel. På exempel två och tre finns ingen exakt tidpunkt, då fördelas vikten över hela tidsspannet beroende på hur många enheter brottet sträcker sig utöver. Detta medför att brott med exakta tids-punkter fördelar sin vikt på endast en enhet, medans andra brott med okänt exakt brottstid delar upp sig över flera enheter. Vilket simuleras med storleken på fyrkanten och linjerna i figuren.

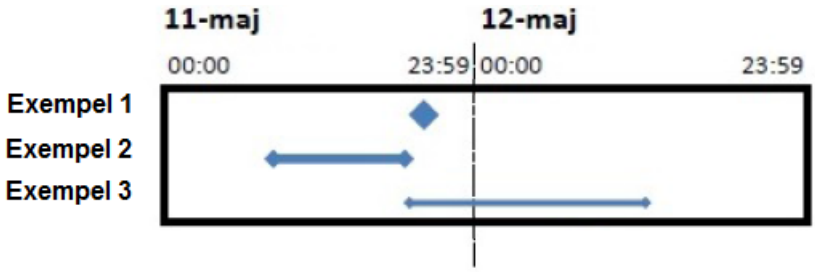

Figure 3.5: Förklaring av Aoristiska-metoden.

Chapter 4

Metod

4.1 Frågeställning

Föregående avsnitt leder fram till följande två forskningsfrågor som kommer besvaras inom detta kandidatarbete:

- FF1. Vilken av de tillgängliga analysmetoderna lämpar sig bäst för tidsanalyser på kort, medellång, samt lång sikt inom brottskategorin bostadsinbrott?
- FF2. Vilka slutsatser kan dras angående bostadsinbrotts fördelning över tiden på kort, medellång, samt lång sikt baserat på analyserna?

4.2 Litteraturstudie

Under litteraturstudien söktes information i böcker och forskningsartiklar utifrån perspektivet spatial och temporal analys av brott. Stöldbrott utgjorde majoriteten av dessa brott. En stor del av litteraturen som användes är beskriven i relaterade arbeten. Det som söktes var vilka analys metoder som fanns och var utvärderade, vilka slutsatser som dragits, samt inom vilka områden de använts. Sedan undersöktes vad för resurser som dessa arbeten och böcker använt sig av för att tillämpa metoderna, även vilka resultat och slutsatser som de resulterat i. Utöver de två arbetena i sektionen relaterade arbeten, söktes information i bland annat forskningsartikeln Temporal Patterns of Danish Residential Burglary, skriven av David W.M Sorensen vid University of Copenhagen. Där diskuterar Sorensen kring olika metoder för spatial och temporal analys och testar dem på olika långa tidsperioder.

4.3 Implementation och Miljö

Den faktiska utvecklingen av metoderna som valdes i litteraturstudien och valet av vilka metoder som skulle prövas gjordes i samråd med handledaren i projektet.

Då systemet i fråga inte fanns sedan tidigare, valdes mjukvara för framtagandet av dessa utifrån rekommendationer från handledare och egna erfarenheter.

4.3.1 Mjukvara och paket

Själva utvecklingen och framtagandet av de utvalda metoderna skedde med ett matematiskt programspråk kallat R och med hjälp av mjukvara från ett användargränssnitt kallat RStudio. R är ett fritt tillgängligt språk och en miljö för statistiska beräkningar och grafik som tillhandahåller en mängd olika statistiska och grafiska tekniker. RStudio är ett kraftfullt och produktivt användargränssnitt för R, som användes för att lättare utföra utvecklingen och experimentet. Det är även möjligt att importera in bibliotek som tillför extra funktionalitet till programmet beroende på vad som efterfrågas. De bibliotek som importerades och användes i detta projekt var följande:

- RMySQL, möjliggör för R att skapa MySQL kommandon för att skicka och hämta data från en MySQL databas.
- Lubridate, ett format att hantera datum och tid.
- QQplot2, möjliggör att skapa fler typer av grafer och diagram med mer avancerade inställningar.
- Circular, möjliggör skapandet av cirkulära diagram.

Alla fyra ovanstående bibliotek även kallade paket finns att ladda ner gratis från Internet.

Experimentets informationshantering skedde med hjälp av en databas med anonymiserad information kring bostadsinbrott i Sverige där bland annat uppgifter som datum, eller datumintervall, för inbrottet och i förekommande fall exakt tidpunkt alternativt ett tidsintervall finns lagrat. Datamängden beskrivs mer detaljerat i sektionen Datamängd. Databasen i fråga var av typen SQL, till den tillkom även ett skript som byggde upp strukturen för tabellen i MySQL. SQL är ett standardiserat programspråk för att hämta och modifiera data i en relationsdatabas. MySQL är en databashanterare som använder språket SQL. För att kunna bygga upp och hantera data och databas på ett smidigare sätt användes utvecklingsmiljön WampServer, som har en hel del funktioner där PhpMyAdmin är det mest nämnvärda för detta projekt.

4.3.2 Utveckling

Utvecklingen och framtagandet av koden som representerade metoderna skedde tillsammans med handledare och med stöd från litteraturen. Målet var att ta fram kod för att möjliggöra analys av de olika metoderna. Koden testades med hjälp av experiment som till exempel att presentera ett diagram över brott som

skett mellan två bestämda datum. Detta för att bland annat säkerhetsställa att koden och metoderna fungerade enligt beskrivning. Till en början skrevs kod för alla metoder över kort sikt, det vill säga tid på dygn. Presentationen av analysen på kort-tidsperiod gjordes med hjälp av en speciell sorts cirkulära diagram som implementerades med hjälp av circular biblioteket, som tidigare nämnts. Alla metoderna använde sig av samma data för att sedan presentera i var sitt cirkulärt diagram att bedöma. På medel-lång och lång period användes ett stapel diagram från biblioteket ggplot2. Även här presenterades alla metoderna vars ett diagram baserat på samma vald data. Det togs alltså fram tre stycken skript för varje metod, en kort, en medel-lång och en lång -tidsperiod. Metoden för slump och Aoristiska krävde speciella funktioner för att fördela ut sig rätt över sin tidsperiod, då dessa skall slumpas ut över en period alternativt fördela i bråkdelar för varje brott.

4.4 Datamängd

En befintlig datamängd med anonymiserad information kring bostadsinbrott i Sverige finns tillgänglig. I denna datamängd finns bland annat uppgifter om start-/slutdatum och start-/sluttid. Dessa start- och sluttider kan därefter användas för att precisera en specifik känd brottstid eller ett intervall. Utöver datum och tidsangivelserna finns även information om:

- Vilken typ av bostad det är, exempelvis villa, radhus, eller lägenhet.
- Specifik beskrivning av bostaden, exempelvis antal våningsplan och så vidare.
- Allmänna uppgifter om tomtens placering, exempelvis. ifall den ligger på landsbygd eller i en tätort.
- Allmänna uppgifter om målsägandens förehavanden vid brottstillfället, exempelvis hemma eller bortrest etc.
- Information om gärningsmannens tillvägagångssätt (modus operandi).

För att bevara anonymiteten hos målsäganden finns inga adressuppgifter med i dataunderlaget. Dock finns uppgifter om vilket landskap respektive postort bostadsinbrotten skett i. För att kunna besvara de båda forskningsfrågorna kommer datamängden delas upp i följande delmängder:

- D1. Bostadsinbrott med okänd brottstidpunkt (ingen tidpunkt angiven, alternativt ett tidsintervall).
- D2. Bostadsinbrott med okänt specifikt brottsdatum (datumintervall som spänner över minst två dagar).

För att besvara frågeställningen kommer datamängd D1 användas för att utvärdera tidsanalysmetoder på kort sikt, det vill säga på dygnsbasis, medan datamängd D2 kommer användas för att utvärdera på medellång och lång sikt, det vill säga på veckobasis respektive månadsbasis. För att kunna utvärdera exaktheten hos de olika metoderna så kommer ytterligare två delmängder skapas utifrån den ursprungliga datamängden som finns tillgänglig. Dessa båda datamängder innehåller bostadsinbrott där ett specifikt brottsdatum är känt (D3) respektive en specifik brottstidpunkt (D4):

- D3. Bostadsinbrott med känt brottsdatum (ett specifikt datum).
- D4. Bostadsinbrott med känd brottstidpunkt (en specifik tidpunkt).

Dessa båda datamängder antas vara representativa urval för samtliga bostadsinbrott och används för att utvärdera exaktheten bland metoderna som utvärderas.

4.4.1 Statistisk jämförelse

Den statistiska metod som används i detta arbete är Chi-square goodness-of-fit testet, framtaget av K. Pearson 1900 [7]. För att få ökad förståelse för Chi-square används även en bok om statistisk analys skriven av Sam K. Kachigan [8]. Syftet med Chi-square testet är att ta reda på skillnaden mellan observerad frekvens och förväntad frekvens. Chi-square används även för att jämföra skillnader mellan två eller flera mängder observerad data. För att undersöka skillnaden i två tabeller med data används ett så kallat Chi-square goodness-of-fit test, vilket är ett test där den observerade datan utvärderas mot den förväntade modellen/fördelningen.

$$\chi^2 = \sum_{i=1}^{k} \frac{(O - E)^2}{E}$$

- O - den observerade frekvensen.
- E - den förväntade frekvensen.
- $\sum$ - Summering, där summan går över i=1 upp till k, där k är antalet ömsesidig uteslutande kategorier.
- χ^2 - Chi-square värde.

Chi-square värdet är storleken på avvikelsen mellan observerade data samt förväntad data, ju större värde desto sämre stämmer observerad data överens med den förväntade fördelningen. Vidare för att utföra ett goodness-of-fit test tas även Degree of Freedoom fram. Degree of Freedoom även kallat DF, står för antalet kategorier som jämförts minus ett, exempelvis är DF lika med sex ifall vi tittar på veckodagar. Vidare om begreppet null-hypotesen, det är antagandet att det inte

finns några relationer mellan observerade värden. P-värdet är sannolikheten (ett värde mellan 0-1) att förkasta null-hypotesen för testet även om null-hypotesen är sann. I detta projekt valdes att förkasta null-hypotesen ifall P-värdet är mindre än fem procent, det vill säga en signifikansnivå på 0,05.

4.5 Experimentdesign

Syftet med att utföra tre experiment med olika tidsperioder är att på flera olika nivåer testa de fem metoderna för att ta reda på om en eller flera metoder är mer lämpliga för de nämna tre tidsperioderna.

4.5.1 Experiment 1 - Kort-tidsperiod

Syftet med det första experimentet är att få svar på frågeställningen om analysmetoderna när man tittar på bostadsinbrott över ett dygn och även hur dessa kan presenteras i diagram för analys. Metoderna som testats nämndes tidigare i sektionen analysmetoder under teknisk bakgrund. Dessa fem metoder testas mot en mängd bostadsinbrott där den totala tidsperioden då inbrottet kan ha skett inte överstiger 30 minuter. Antagandet görs att dessa bostadsinbrott inte påverkas av när på dygnet det är större risk att ett inbrott avslöjas eller upptäcks. Jämförelsen mellan de fem metoder och delmängden D4 sker med hjälp av metoden Chi-Square goodness-of-fit test, förklarat i tidigare sektion metod. För att studera skillnaderna i hur metoderna presenteras visuellt användes Circular paketet nämnt i sektionen om mjukvara och paket. Utöver goodness-of-fit testet användes även visuell bedömning när det gäller bostadsinbrottens fördelning över dygnet eftersom brist på data kan leda till missvisande siffror i statistiska tester (Chi-Square testet i detta fallet). Den data som används är ett utdrag från Polisens databas från och med 2013-05-01 till 2014-05-01, totalt 3241 bostadsinbrott. Anledningen till att just detta tidsintervall valdes var då projektet påbörjades i maj 2014 och att det sedan plockades ut data för ett år tillbaka i tiden.

4.5.2 Experiment 2 - Medellång-tidsperiod

I det andra experimentet söktes svar på samma frågeställning men med fokus på veckodagarna istället för tid på dygnet. Metoderna är desamma. Syftet med att testa på veckodag är att få fler mätvärde i delmängden som metoderna testas mot. I detta fall presenteras en ny tidsperiod som visar tidpunkten då bostadsinbrott verkligen sker under en veckas tid. Precis som experiment 1 kommer jämförelsen att ske med hjälp av Chi-Square goodness-of-fit test, med skillnaden att delmängden som metoderna jämför består av brott som har tidsperiod då inbrott skett just en specifik veckodag. För att presentera metoderna visuellt används på

medellång-tidsperiod histogram. Data som används är densamma som föregående experiment: från och med 2013-05-01 till 2014-05-01, totalt 3241 bostadsinbrott.

4.5.3 Experiment 3 - Lång-tidsperiod

Det tredje experimentet ska precis som de två tidigare, svara på vilken metod som lämpar sig bäst att använda för temporal analys, fast denna gången på lång-tidsperiod. Precis som föregående experiment är syftet att testa på en längre tidsperiod för att få fler exakta värden i delmängden. Delmängden i detta experimentet kommer att vara brott som skedde under just en specifik dag i månaden, även kallad delmängd 3. Detta experiment kommer även det att använda Chi-Square goodness-of-fit test för att testa metoderna mot delmängden. För att presentera metoderna visuellt används på lång-tidsperiod histogram. Data som används är densamma som de två föregående experiment: från och med 2013-05-01 till 2014-05-01, totalt 3241 bostadsinbrott.

Chapter 5

Resultat

5.1 Testresultat Kort-tidsperiod

	χ^2	p-värde
Start-metoden	Inf.	$<$ 2.2e-16
Slut-metoden	Inf.	$<$ 2.2e-16
Medel-metoden	Inf.	$<$ 2.2e-16
Slump-metoden	Inf.	$<$ 2.2e-16
Aoristiska-metoden	Inf.	$<$ 2.2e-16

Table 5.1: Testresultat kort-tidsperiod

Tabellen ovan visar testresultat från Chi-square goodness-of-fit testet, utfört med delmängd 4 som förväntad fördelning. Utifrån resultat utläses ett p-värde som är det minsta värdet R skriver ut, det vill säga att det anses vara noll. Detta tillsammans med ett oändligt χ^2 värde betyder att man med säkerhet kan förkasta null-hypotesen, det vill säga att metodernas data inte visade sig ha något samband med den exakta tidsmängden.

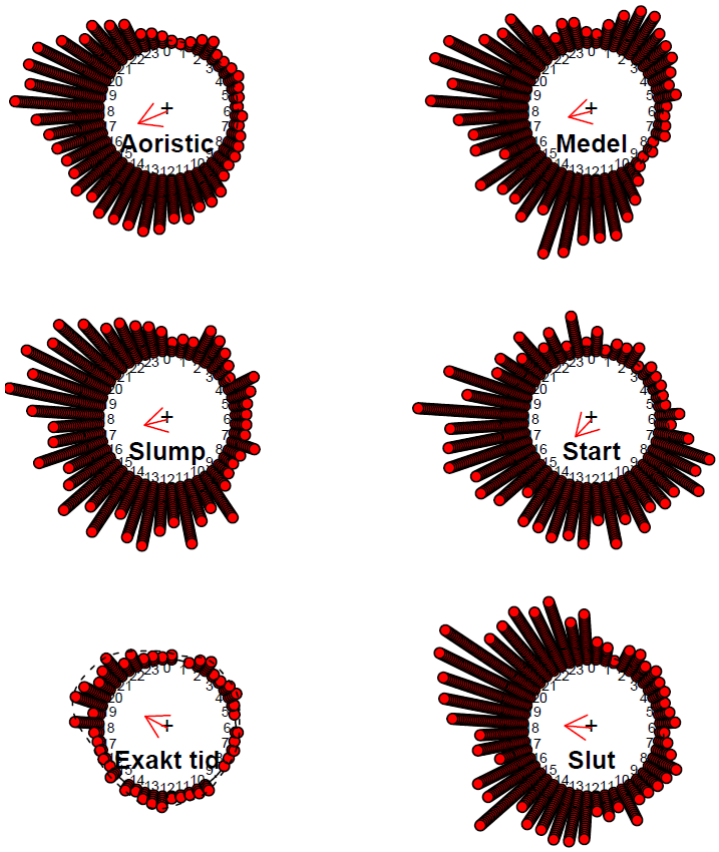

Figure 5.1: Expriment 1, resultat visas som cirkulära diagram

De sex diagrammen som presenteras ovan, är test utfört på alla de metoder som tidigare nämnts samt den förväntade fördelningen av data, här kallad exakt tid. Alla diagram från figur 5.1 ovan utom den för exakt tid (delmängd 4) visar att medelvärdena av metodernas fördelningar är någon gång mellan klockan 15:00 och 18:00. På diagrammen kan det utläsas att den Aoristiska metoden fördelar jämnt ut sig över dygnet utan några större avvikelser. Vidare utläst från tyngdpunkts-pilarna från mitten av diagrammen pekar det exakta tidsdiagrammet mot klockan åtta på kvällen medan resterande diagram pekar mot tidigare kväll och eftermiddag. Pilens längd symboliserar hur mycket tyngdpunkten väger, den Aoristiska har den längsta pilen då den har den jämnaste fördelningen. Dock visar det sig att slut-metoden har den pil som pekar närmst den exakta pil/tyngdpunkt för tidsdiagrammet.

5.2 Testresultat Medel-lång-tidsperiod

	χ^2	p-värde
Start-metoden	11.1425	0.08407
Slut-metoden	33.3337	9.044e-06
Medel-metoden	14.8935	0.0211
Slump-metoden	11.951	0.06307
Aoristiska-metoden	9.215	0.1618

Table 5.2: Testresultat medel-lång-tidsperiod

Tabellen ovan visar testresultat från Chi-square goodness-of-fit testet, utfört med delmängd 3 som förväntad fördelning. I detta experiment kan vi till skillnad från föregående experiment utläsa skillnader mellan resultaten. Det utläses att Slut-metoden har det högsta χ^2 värdet och även det minsta p-värde. Som tidigare nämnts i förklaringen av det statistiska testet kan man förkasta null-hypotesen då p-värdet är mindre än 0.05. Vilket i detta fall betyder att man på slut och medel-metoderna kan förkasta null-hypotesen då data inte anses stämma överens med de förväntade värdena. De resterande tre metoderna start, slump och den Aoristiska har tillräckligt stora p-värden för att inte förkasta null-hypotesen och de anses stämma bättre överens med den förväntade fördelningen. Dock ligger både slump- och start-metodernas p-värden relativt nära vald signifikansnivå.

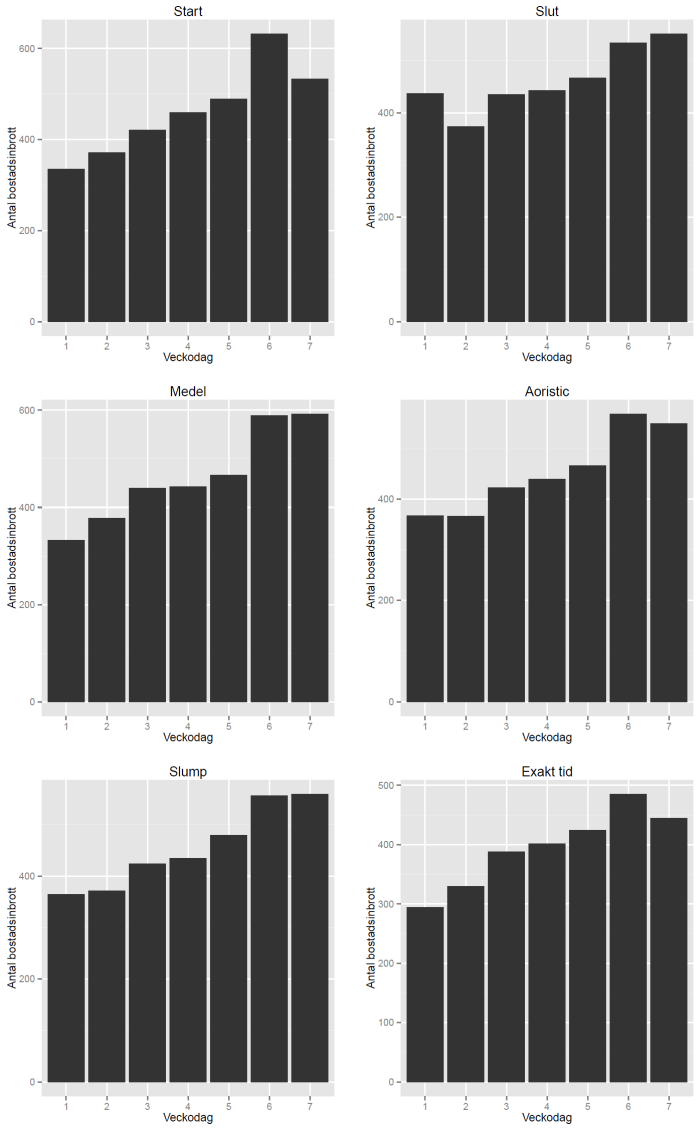

Figure 5.2: Expriment 2, resultat visas som histogram

Diagrammen ovan visar histogram över veckodagarna, med data framtaget med hjälp av de fem metoderna som testats. De sjätte diagrammet är den förväntade fördelningen, delmängd 3 kallat exakt tid, som metoderna statistiskt testades mot. I ovanstående diagram motsvarar 1 på x-axeln söndag, 2 måndag, 3 tisdag och så vidare. Utifrån dessa diagram utläses ett maximum på två veckodagar: fredag och lördag. Det syns även tydligt att färre antal bostadsinbrott sker i början av veckan. Skillnaden mellan början av veckan och slutet av veckan ser ut att vara mellan 100 och 230 inbrott, vilket i vissa fall nästa är en ökning på 100 procent.

5.3 Testresultat Lång-tidsperiod

	χ^2	p-värde
Start-metoden	17.0038	0.9725
Slut-metoden	12.7588	0.9975
Medel-metoden	16.7608	0.9753
Slump-metoden	9.1072	0.9999
Aoristiska-metoden	5.6854	1

Table 5.3: Testresultat lång-tidsperiod

Tabellen ovan visar testresultat från Chi-square goodness-of-fit testet, utfört med delmängd 3 som förväntad fördelning. Här i denna tabell utläses höga p-värden där till och med den Aoristiska metoden gav ett. Alla metoder hade över 0,97 i p-värdet, alltså kan man inte förkasta null-hypotesen för någon metod. Utifrån χ^2 värdet anses den Aoristiska-metoden vara mest lik den förväntade mängden, följt av slump-metoden.

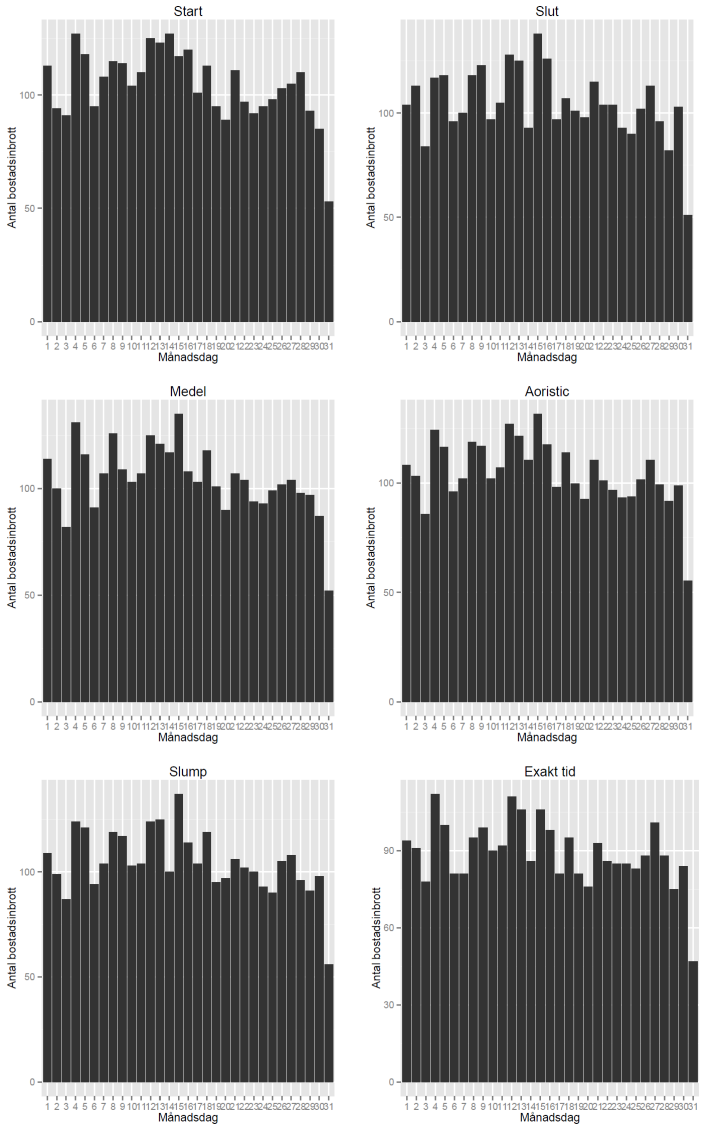

Figure 5.3: Expriment 3, resultat visas som histogram

Figur 5.3 ovan visar histogram över månadsdagarna, med data framtaget med hjälp av de fem metoderna som testats. De sjätte diagrammet är den förväntade fördelningen, delmängd 3 kallat exakt tid, som metoderna statistiskt testades mot. Anledningen till att alla diagram ovan har ett minimum på månadsdag 31 är då dag 31 ej existerar i hälften av månaderna på ett år. Eftersom data i detta projekt är utplockat för ett år syns detta tydligt i alla diagrammen. Det finns ett mindre antal datum som har betydligt fler inbrott än andra, exempelvis månadsdag 15 har i många av diagrammen högst antal.

Chapter 6

Analys och diskussion

Detta arbete syftar till att utvärdera vilka temporal-metoder som enligt resultat från experiment går att använda för att visuellt presentera hur bostadsinbrott fördelar sig över intressanta tidsperioder.

6.1 Experiment 1 - Kort-tidsperiod

Att besvara tesen i fråga genom att titta på kort-tids period visade sig vara svårt, då resultatet från chi-square testet gav identiska svar för samtliga metoder. Att testet visade samma resultat på alla, att det inte alls fanns några samband till delmängden D4 berodde på att den datamängden helt enkelt inte var tillräckligt stor, endast 175 inbrott. Med 48 kategorier och endast 175 värden anser det statistiska testet att det är för lite data och att värdena som finns är för slumpmässiga. I experiment 1 tittar vi alltså med blotta ögat på de diagram som tagits fram för att se ifall det går att utläsa någon skillnad. Det som kan utläsas visuellt från dessa diagram kan vara missvisande då det statistiska testet påpekade att det finns för lite data att jämföra med. För att enkelt kunna jämföra dessa diagram se figur 5.1 i resultatdelen. Efter att visuellt jämfört diagrammen utläses en hel del intressant information. Anledningen varför diagrammet med data om exakt tid visar annorlunda tros bero på flera faktorer, där den största faktorn anses vara att det faktiska tidsdiagrammet egentligen visar när på dygnet gemene man är i sin bostad för att upptäcka ett inbrott. Med antagandet att föregående hypotes stämmer lämpar det sig inte alls att på kort-tidsperiod jämföra mot en delmängd med så kort-tidsperiod, för att ett inbrott skall infinna sig i delmängden måste tidsspannet då brottet kan ha inträffat vara mindre än 30 minuter. Vidare visar medel-metoden, slump-metoden och den Aoristiska metoden att dessa tre har väldigt lika riktningar på den pil som visar tyngdpunkten i diagrammen och även med längden visar hur starkt sambandet är. Utifrån längden på pilarna och även visuellt hur värdena fördelar sig, kan man utläsa att den Aoristiska metoden har en mer jämn fördelning. Detta beror på att den som tidigare nämnt i förklaring av metoder tar hänsyn till alla tidpunkter inom perioden istället för att välja ut bara en tidpunkt då inbrottet kan ha skett. Sammanfattningsvis utifrån experiment 1, men med hjälp av diagrammen och dess funktion kan slutsatsen dras att det

Aoristiska diagrammet visar ett resultat som bäst överensstämmer med utvärderingsdiagrammet för känd tidpunkt. Detta då den sprider ut tidsperioderna och ger en mer realistisk kurva för Polisen att diskutera sin resursfördelning på.

6.2 Experiment 2 - Medellång-tidsperiod

Till skillnad från föregående experiment gav den statistiska jämförelsen med Chi-Square intressantare resultat att diskutera. Från χ^2 värdet från tabellen 5.2, kan utläsas att slut-metoden inte alls stämmer överens mot delmängd 3 (D3) som vi i detta test jämför mot. Överraskande visade det sig start-metoden ge lika bra resultat, om inte bättre, än medel och slump-metoden. Men den som stämde mest överens med vår delmängd med exakt datum för inbrott var den Aoristiska metoden, vilket även påvisas av dess p-värde som ligger dubbelt så nära 1 som den näst bästa metoden, start-metoden. Här noteras åter igen att start-metoden faktiskt ger bättre värden än både slump och medel-metoden. Då de framtagna histogrammen studeras med blotta ögat för att undersöka samband som de statistiska resultatet inte ger, utläses fler antal inbrott i slutet av veckan. Fredag och lördag anses av alla testade metoder vara de två mest utsatta dagarna i veckan. Visuellt ser start-metoden ut att forma en trappa som efterliknar den exakta tidsdiagrammet med delmängd 3. Den statistisk bäst presterande Aoristiska metoden har ungefär samma trapp-effekt. Till skillnad från experiment 1 är det i detta experiment mycket svårare att utläsa skillnaden visuellt, efter noggrann granskning tycks resultatet från de statistiska testet stämma överens med vad histogrammen visar. Även här kan Polisen ta hjälp av dessa diagram som hjälpmedel till att fördela sina resurser över en veckas tid.

6.3 Experiment 3 - Lång-tidsperiod

Resultatet av det tredje och sista experimentet visade värde med signifikantare samband än föregående experiment, vilket kan tydas utifrån värdena i kolumnen för p-värde. Då värdena i p-värde är väldigt nära ett kan man med stor sannolikhet konstatera att det finns likheter mellan den faktiska delmängden och några av de metoder som testats. Vid vidare granskning av värdena i kolumnen X-square syns tydligt att den Aoristiska-metoden är den med minst skillnad mellan den förväntade och den observerade fördelningen. Slump-metoden visade sig vara betydligt mer användbar på en längre tidsperiod jämfört med föregående experiment. Utifrån den statistiska bedömning av metoderna visar experiment 3 det resultat som var förväntat utifrån den bakgrundsstudien som tidigare gjorts i projektet. Något som inte förväntades vara att slut-metoden ger bättre resultat än medel-metoden när de båda jämförs mot exakt tid delmängden (D3). Visuellt sätt kan det i detta fall vara svårt att dra några egna slutsatser, dock när man

studerar diagrammen noga kan man se likheter mellan den exakta tidsdiagrammet och den Aoristiska diagrammet.

6.4 Diskussion

Utifrån experimentet 1, kort-tidsperiod framkom det att Start- och Aoristiska-metoden var de som presterade bäst. Experiment 2, medel-lång-tidsperiod visade att på veckodagarna var den Aoristiska metoden lämpligast. Även på det tredje och sista experimentet lång-tidsperiod preserade den Aoristiska bäst resultat, tätt följt av slump-metoden. Experimentresultaten indikerar att forskningsfråga FF1 kan besvaras. Den Aoristiska-metoden är den lämpligaste metoden av de fem att bruka vid analys av temporal data i kategorin bostadsinbrott. För att besvara den andra forskningsfrågan FF2 anses att Polisen skulle kunna dra en hel mängd slutsatser utifrån de experimentresultat som framtagits. Diagrammen presenterar relevant information som skulle kunna användas som exempelvis mall för resursfördelning. Bakomliggande faktorer kan analyseras för trender där antalet brott ökat eller minskat över tid. Möjligheten finns även att i efterhand utvärdera insatta åtgärder, genom att med hjälp av en mjukvarukomponent, analysera mönster i förekomsten av brott. Så länge som data systematiskt samlas in kan man med hjälp av kod från detta projekt automatiskt generera diagram över valfri tidsperiod på tre olika sätt, dygn, veckodag och månadsdag.

6.4.1 Problem

Problem som uppstod under projektets gång var bland annat vid installationen av RMySQL biblioteket till R, då det var skillnader i programvaruversionen som den utvecklats i och som projektet använde sig av. Problemet löstes med hjälp av en global variabel i operativsystemet Windows8 som pekade på var MySQL servern var installerad på hårddisken. Ett annat problem som uppkom under projektets gång var att komma fram till vad metoderna skulle jämföras med, vad som var den faktiska fördelningen av inbrott, då det vid kort-tids analys inte fanns tillräckligt med brott som hade exakta tidpunkter för tillslag.

Chapter 7

Slutsatser och framtida arbete

Slutsatsen som kan dras utifrån utvecklingen av de R-skript som genom tre experiment utvärderats, är att den lämpligaste metoden att använda för att analysera temporal data av bostadsinbrott var den Aoristiska-metoden. Brist på antal brott med ett spann på mindre än 30 minuter ledde till att experiment 1 inte gav något tillförlitligt resultat. I framtida arbeten inom temporal analys och framtagandet av metoder finns stor möjlighet för Polisen och forskare att med hjälp av ny data ta fram trender och fördela resurser baserat på dessa.

References

[1] Brottsförebyggande Rådet. *Anmälda brott. [Online].* Tillgänglig: https://bra.se/ 2014-01-31.

[2] Polisen. *Bostadsinbrott – Lagar och fakta. [Online].* Tillgänglig: http://polisen.se/Lagar-och-regler/Om-olika-brott/Stold-och-grov-stold/Bostadsinbrott/ 2013-06-27.

[3] Polisen. *Ny polisorganisation 2015. [Online].* Tillgänglig: http://polisen.se/Om-polisen/Organisation/Ny-polisorganisation-2015/ 2013-04-16

[4] A. Borg, M. Boldt, N. Lavesson, U. Melander, V. Boeva. *Detecting serial residential burglaries using clustering.* Elsevier Ltd. 2014.

[5] Jerry H. Ratcliffe och Michael J. McCullagh. *Aoristic Crime Analysis.* Internetational Journal of Geographical Information Science. 1998.

[6] Rachel B. Santos. *Crime Analysis with Crime Mapping, 3 Edition.* Sage publications. 2013.

[7] K. Pearson. *On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Phil. Mag. (5)50, 157-175.* 1900. Reprinted in K. Pearson (1956), pp. 339-357.

[8] Sam K. Kachigan *Statistical Analysis: An Interdisciplinary Introduction to Univariate & Multivariate Methods* Radius Pr. 1986.

Appendix A

Diagram

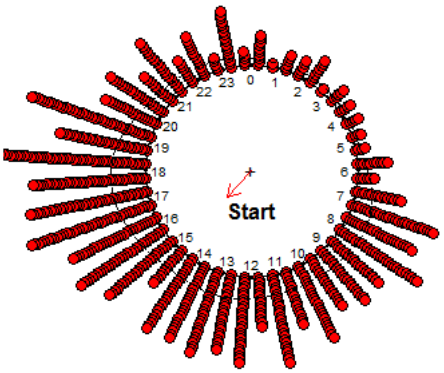

Figure A.1: Resultat experiment 1, start-metoden.

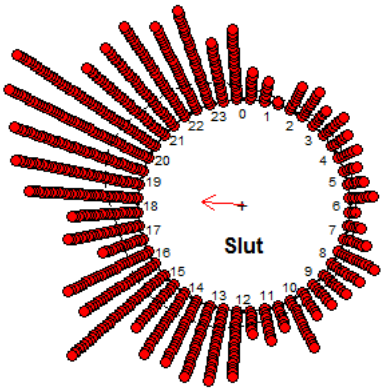

Figure A.2: Resultat experiment 1, slut-metoden.

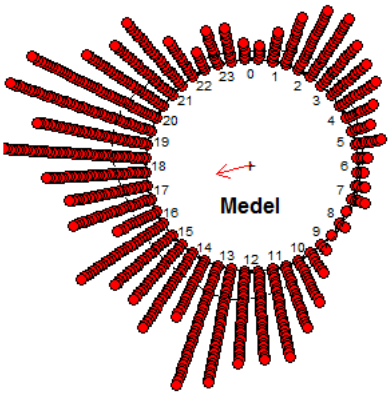

Figure A.3: Resultat experiment 1, medel-metoden.

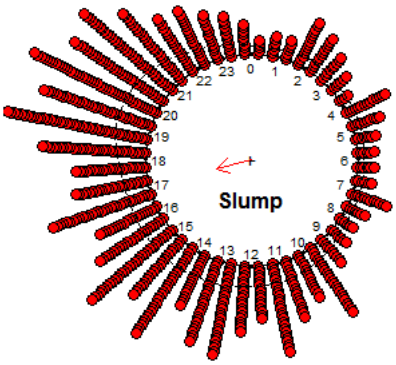

Figure A.4: Resultat experiment 1, slump-metoden.

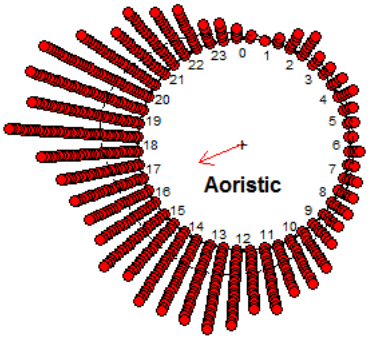

Figure A.5: Resultat experiment 1, Aoristiska-metoden.

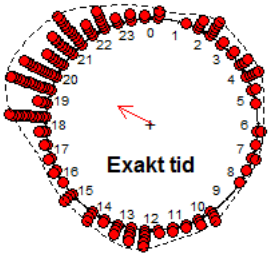

Figure A.6: Resultat experiment 1, delmängd 4 (exakt tid).

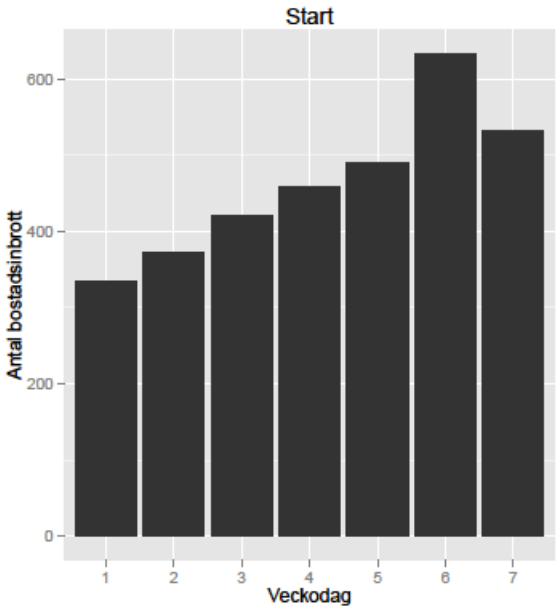

Figure A.7: Resultat experiment 2, start-metoden.

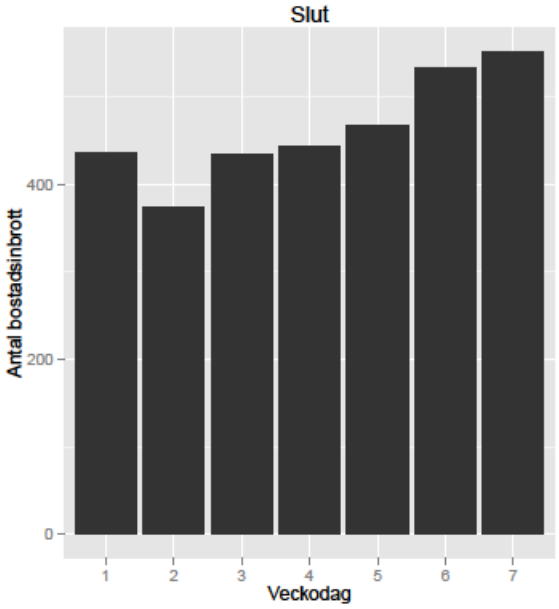

Figure A.8: Resultat experiment 2, slut-metoden.

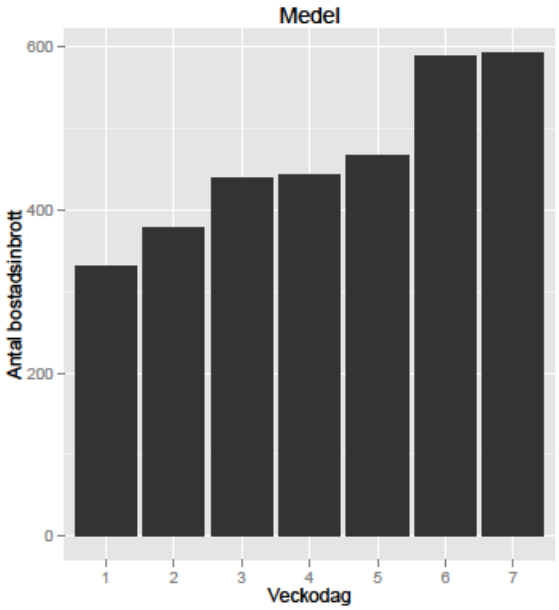

Figure A.9: Resultat experiment 2, medel-metoden.

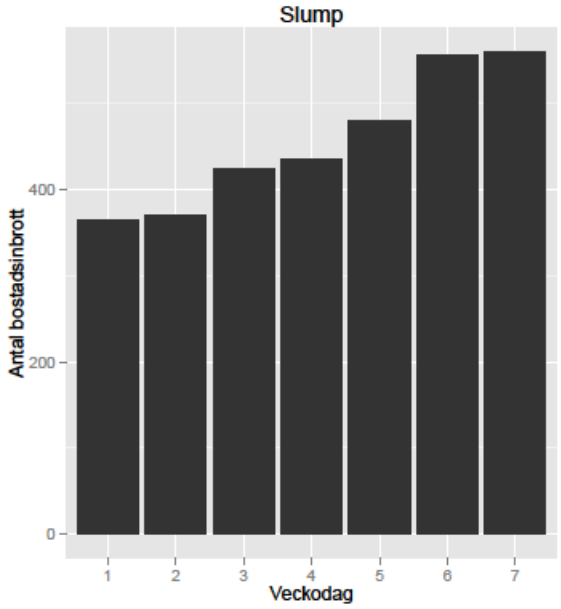

Figure A.10: Resultat experiment 2, slump-metoden.

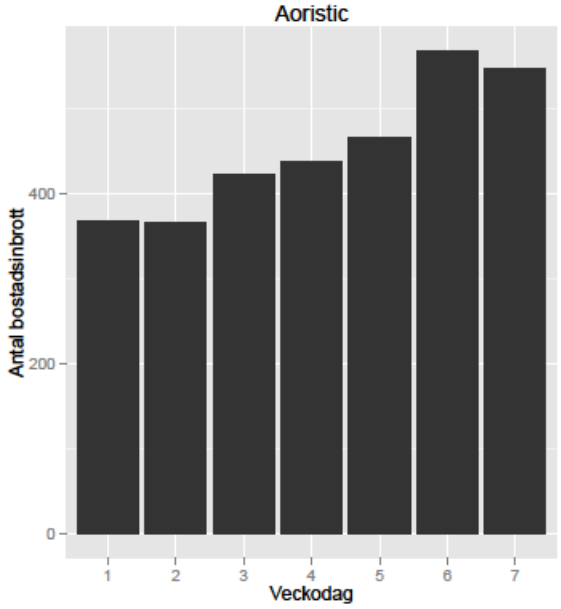

Figure A.11: Resultat experiment 2, Aoristiska-metoden.

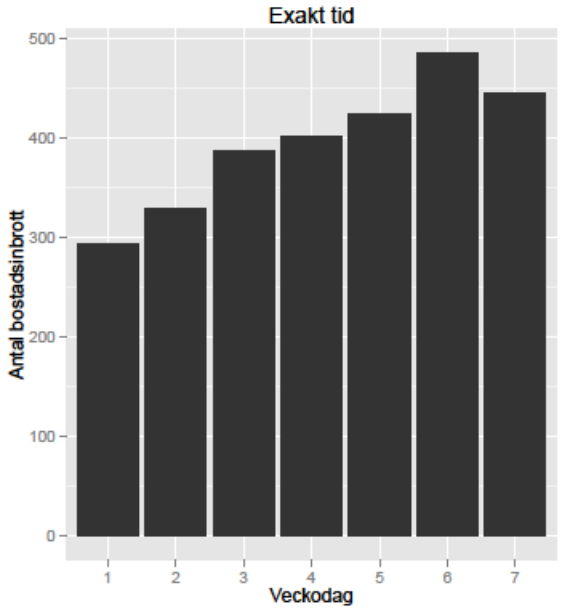

Figure A.12: Resultat experiment 2, delmängd 3 (exakt tid).

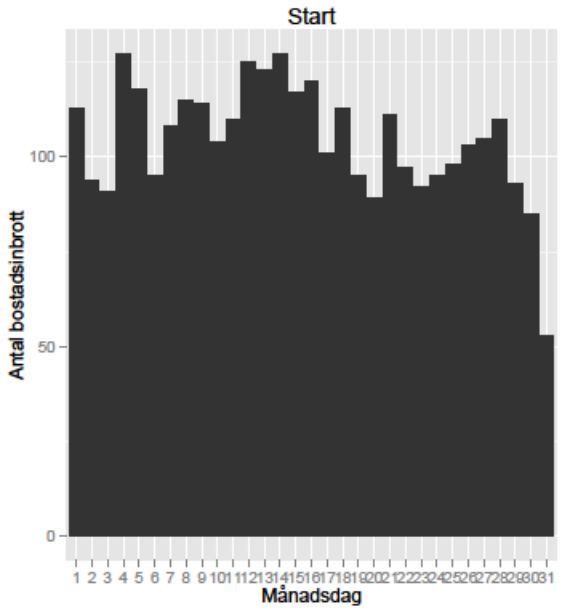

Figure A.13: Resultat experiment 3, start-metoden.

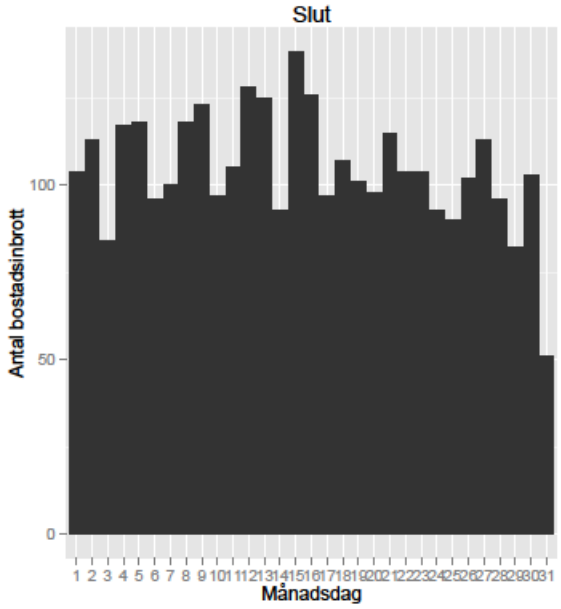

Figure A.14: Resultat experiment 3, slut-metoden.

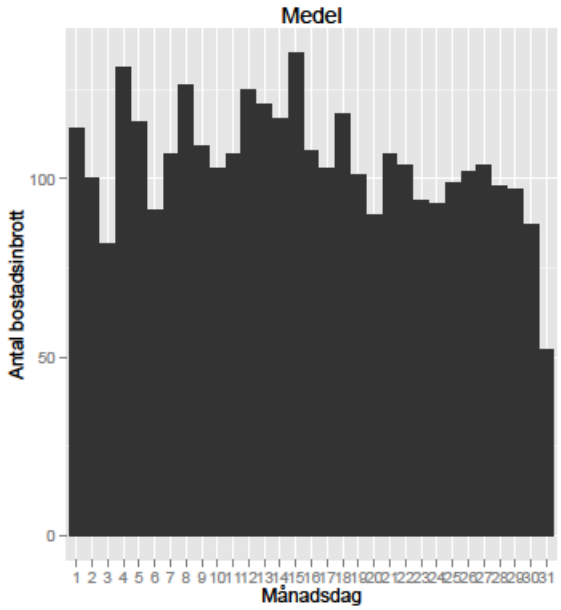

Figure A.15: Resultat experiment 3, medel-metoden.

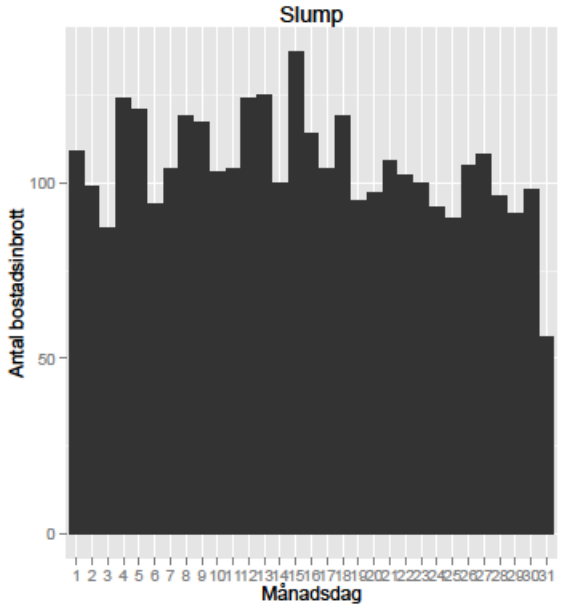

Figure A.16: Resultat experiment 3, slump-metoden.

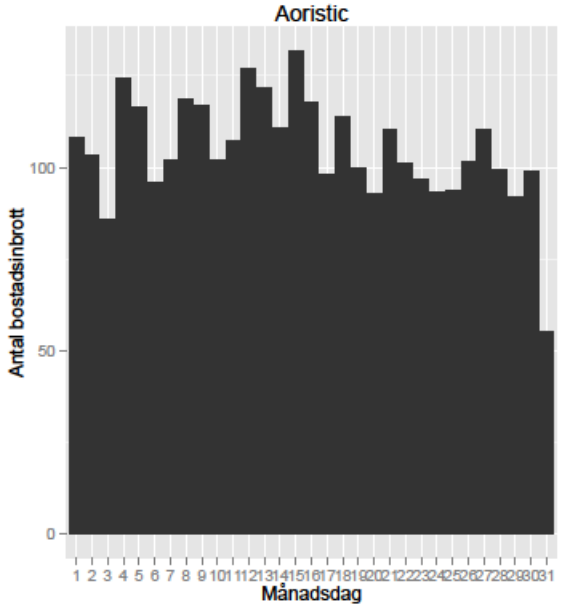

Figure A.17: Resultat experiment 3, Aoristiska-metoden.

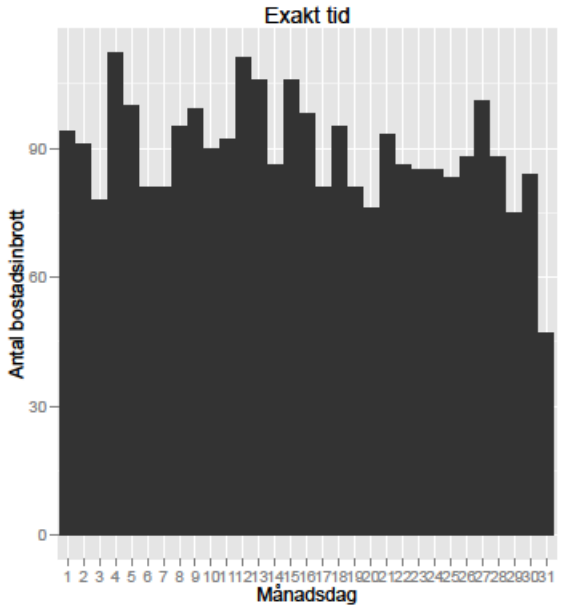

Figure A.18: Resultat experiment 3, delmängd 3 (exakt tid).

Appendix B

Script

B.1 Experiment 1

```
# Autor:          Olle Svenhag, Student & Martin Boldt, PhD
# Date:           2014-04-20
# E-mail:         Olle.Svenhag@hotmail.com
# Affiliation:    Blekinge Institute of Technology, Sweden

# Required libraries
library(RMySQL)
library(lubridate)
library(ggplot2)
library(circular)

#Bin size
unit = 1800

############################## CIRCULAR OBJECT
    ##############################

# plot a vector of 48 bins representing every full and
    half hours bwteen 00:00 - 23:59
circular.cime.plot <- function(p, plot.bins=48, plot.bw
    =30, plot.axis=seq(0, 23, by=1), plot.title=NULL, zero
    =0) {
  # make a circular class object: using native degrees and
      clock24 template that sets zero and direction
   circ.p <- circular(p, units='hours', template="clock24",
      rotation=c("clock"))
  mean.p <- mean(circ.p)
  # compute mean arrow length
  rho.p <- rho.circular(circ.p)
  # setup custom axis
```

```
  axis.p <- circular(plot.axis, units='hour', template="
      clock24")
  # plot circular histogram
  plot(circ.p, axes=FALSE, stack=TRUE, bins=plot.bins,
      shrink=2, cex=1.3, sep=0.024, pch=21, col=1, bg='Red'
      )
  # paint density where bw is the smoothness
  lines(density(circ.p, bw=plot.bw), col='Black', lty=2)
  # draw axes
  axis.circular(at=(axis.p), labels=axis.p, cex=0.6)
  # draw arrow showing mean direction and confidence
  arrows.circular(mean.p, shrink=rho.p, length=0.15, col='
      Red')
  # draw the title in histogram
  text(0, -0.25, plot.title, cex=1, font=2.5)
}

############################## FUNCTIONS
    ##############################

get.bin.number <- function( p.in ) {

  #print( paste( "[", p.in, ", ", abs( p.in - floor( p.in
      ) ), "]", sep="") )

  if ( abs( p.in - floor( p.in ) ) >= 0.5 ) {
    result = 1 + floor( p.in ) * 2
  } else {
    result = floor( p.in ) * 2
  }

  if ( p.in > 48 ) {
    print( paste( "Bin number larger than 48 (", p.in, ").
       Aborting." ) )
  }
  result=result+1
  return( result )
}

calculate.aorisic.fractions <- function( p.startbin, p.
    endbin ) {
  if (p.startbin != p.endbin && p.endbin == 0 ) {
    res = 1
```

```
  } else if ( p.startbin == p.endbin ) {
    res = 1
  } else if ( p.startbin > p.endbin ) {
    bincount = ( 48 - p.startbin + 1) + ( p.endbin )
    res = 1 / bincount
  } else {
    res = 1 / (p.endbin - p.startbin + 1)
  }
  return( res )
}

assign.aorisic.fractions <- function( aoristic.array, p.
   startbin, p.endbin, aoristic.fraction ) {
  if ( p.startbin > p.endbin ) {
    # print( paste( "=>startbin=", p.startbin, " : endbin
       =", p.endbin, " : aoristic.fraction=", aoristic.
       fraction, sep="" ) )
  } else {
    #print( paste( "  startbin=", p.startbin, " : endbin
       =", p.endbin, " : aoristic.fraction=", aoristic.
       fraction, sep="" ) )
    aoristic.array[ p.startbin:p.endbin ] = aoristic.array
       [ p.startbin:p.endbin ] + aoristic.fraction
  }

  return(aoristic.array)
}

randomize.bin.number <- function(p.startbin, p.endbin){
  if(p.startbin > p.endbin)
  {
    nr.of.bins <- ((48 - p.startbin) + p.endbin)
    random.bin.nr <- p.startbin + sample(0:nr.of.bins, 1)
    if(random.bin.nr > 48) {
      res <- (random.bin.nr - 48)
    } else {
      res <- (random.bin.nr)
    }

  } else if(p.startbin < p.endbin) {

    res <- sample(p.startbin:p.endbin, 1)
```

```
  } else if(p.startbin == p.endbin){
    res <- p.startbin

  }else{
    eexxxxxxxxxxxxxxit
  }

  return(res)
}

############################## DATABASE (get data)
    ##############################

# Grab burglary dates and times from burgleform table
#Connect to MySQL
mydb = dbConnect(MySQL(), user='root', password='', dbname
    ='burglary', host='127.0.0.1', port=3306, unix.socket='
    /tmp/mysql.sock')

#Send Query and save result to "data"
rs = dbSendQuery(mydb, "select datestart, timestart,
    dateend, timeend from burgleform where datestart >
    '2013-05-01' && dateend < '2014-05-01'")

data = fetch( rs, n=-1 )

# Close db connection
dbClearResult(rs)
dbDisconnect( mydb )

# Setup crime.time dataframe
from    <- ymd_hms( paste( data$datestart, data$timestart,
     sep=" ") )
to      <- ymd_hms( paste( data$dateend, data$timeend, sep
    =" ") )
range   <- as.double( difftime( to, from, units="secs") )
average <- from + range/2
#random <- runif( length( average ), from, (from+range) )
    %% 86400
crime.time = data.frame( from, to, average, range )
```

```
############################## AORISTIC
    ##############################

aoristic.dataset = subset( crime.time, crime.time$range <=
    unit*47 )
aoristic.array <- array( 0, dim=c(1,48) )

for( counter in 1:nrow( aoristic.dataset ) ) {
  tmp.from <- aoristic.dataset[ counter, 1 ]
  tmp.to <- aoristic.dataset[ counter, 2 ]
  tmp.from.binnumber <- get.bin.number( hour( tmp.from ) +
      (minute( tmp.from ) / 60 ) )
  tmp.to.binnumber <- get.bin.number( hour( tmp.to ) + (
      minute( tmp.to ) / 60 ) )
  tmp.fraction <- calculate.aorisic.fractions( tmp.from.
      binnumber, tmp.to.binnumber )

  aoristic.array <- assign.aorisic.fractions( aoristic.
      array, tmp.from.binnumber, tmp.to.binnumber, tmp.
      fraction )


}

print( aoristic.array )
#print( round( aoristic.array ) )

aoristic.data.array <- array()
#print(aoristic.data.array)

count=1
#print out all the values given from aoristic.arra
for(counterBins in 1:48) {
  while(aoristic.array[counterBins] >= 1){

    aoristic.data.array[count] <-((counterBins/2)-0.25)
    aoristic.array[counterBins] <-(aoristic.array[
        counterBins]-1)
    count<-(count+1)
  }
}
print(aoristic.data.array)
```

```
############################## RANDOM
    ##############################

random.dataset = subset( crime.time, crime.time$range <=
    unit*47 )
random.array <- array( 0, dim=c(1,48) )

for( counter in 1:nrow( random.dataset ) ) {
  tmp.from <- random.dataset[ counter, 1 ]
  tmp.to <- random.dataset[ counter, 2 ]
  tmp.from.binnumber <- get.bin.number( hour( tmp.from ) +
      (minute( tmp.from ) / 60 ) )
  tmp.to.binnumber <- get.bin.number( hour( tmp.to ) + (
      minute( tmp.to ) / 60 ) )
  tmp.random.binnumber <- randomize.bin.number( tmp.from.
      binnumber, tmp.to.binnumber )

  random.array[tmp.random.binnumber] <- (random.array[tmp.
      random.binnumber] + 1)

}

print( random.array )


random.data.array <- array()
#print(random.data.array)

count=1
#print out all the values given from random.array
for(counterBins in 1:48) {
  while(random.array[counterBins] >= 1){

    random.data.array[count] <-((counterBins/2)-0.25)
    random.array[counterBins] <-(random.array[counterBins
        ]-1)
    count<-(count+1)
  }
}
print(random.data.array)
```

```
############################## ACTUAL TIME, AVERAGE, START and
    END ##############################


## Dataset 4 ##
time.within.unit =  subset( crime.time, crime.time$range
   <= unit )
time.within.unit.count = nrow( time.within.unit )

## Dataset 1 ##
time.outside.unit = subset( crime.time, crime.time$range
   <= (unit*47) )

time.outside.unit.count = nrow( time.outside.unit )

## Prepairing Datasets for Plotting ##

#Hour of the day crime took place using Average-method

#Actual time of crime
time.1800.events.within.unit.average = hour( time.within.
   unit$average ) + ( minute( time.within.unit$average )/
   60 )

#Average
time.1800.events.outside.unit.average = hour( time.outside
   .unit$average ) + ( minute( time.outside.unit$average )
   /60 )


#print( time.1800.events.within.unit.average )

#Hour of the day crime took place using Start-method
time.1800.events.outside.unit.start = hour( time.outside.
   unit$from ) + ( minute( time.outside.unit$from )/60 )

#Hour of the day crime took place using End-method
time.1800.events.outside.unit.end = hour( time.outside.
   unit$to ) + ( minute( time.outside.unit$to )/60 )
```

```
############################## PLOTTING
    ##############################
par(mar=c(0,0,0,0), mfcol=c(3,2))


aoristic <- circular( aoristic.array, units='hours',
    template="clock24", rotation=c("clock") )
circular.cime.plot( aoristic.data.array, plot.bw=30, plot.
    title=paste( "\nAoristic", sep="" ) )


random <- circular( random.array, units='hours', template=
    "clock24", rotation=c("clock") )
circular.cime.plot( random.data.array, plot.bw=30, plot.
    title=paste( "\nSlump", sep="" ) )


# Plot[Exact time]
circular.cime.plot( time.1800.events.within.unit.average,
    plot.bw=30, plot.title=paste( "\nExakt_tid", sep="" ) )

# Plot[Average]
circular.cime.plot( time.1800.events.outside.unit.average,
     plot.bw=30, plot.title=paste( "\nMedel", sep="" ) )


# Plot[Start]
circular.cime.plot( time.1800.events.outside.unit.start,
    plot.bw=30, plot.title=paste( "\nStart", sep="" ) )

# Plot[End]
circular.cime.plot( time.1800.events.outside.unit.end,
    plot.bw=30, plot.title=paste( "\nSlut", sep="" ) )



############################## Chi Square (DAY)
    ##############################


#Actual Time (vector)
day.actual.time <- array( 1:48 )
```

```
day.actual.time <- append(day.actual.time, floor(1 + time
    .1800.events.within.unit.average*2))
#Count total number of data
length(day.actual.time) -48
#Make frequency table
day.actual.time.freq <- as.data.frame(table(day.actual.
    time))
#Insert another colom with percentage value (probability)
day.actual.time.freq$probability <- (day.actual.time.freq$
    Freq -1) / (length(day.actual.time)-48)

#Sum of the probability should always be 1 (confirmed
    below)
#sum(day.actual.time.freq$probability)

#Now we have the expected value to use in our Chi square
    test.
day.expected <- c( +day.actual.time.freq$probability)

# - FROM ----------------------------------------#

dfrom.df <- as.data.frame(table(floor(1 + time.1800.events
    .outside.unit.start*2)))

chisq.test.dfrom <- chisq.test(dfrom.df$Freq, p=day.
    expected)
#X-squared = Inf, df = 47, p-value < 2.2e-16

# - TO ------------------------------------------#

dto.df <- as.data.frame(table(floor(1 + time.1800.events.
    outside.unit.end*2)))

chisq.test.dto <- chisq.test(dto.df$Freq, p=day.expected)
#X-squared = Inf, df = 47, p-value < 2.2e-16

# - AVERAGE -------------------------------------#
daverage<- array( 1:48 )
daverage <- append(daverage,(floor(1 + time.1800.events.
    outside.unit.average*2)))
daverage.df <- as.data.frame(table(daverage))
```

```
chisq.test.daverage <- chisq.test(daverage.df$Freq, p=day.
    expected)
#X-squared = Inf, df = 47, p-value < 2.2e-16

# - AORISTIC ----------------------------------#

daoristic.df <- as.data.frame(table(floor(1 + (aoristic.
    data.array)*2)))
chisq.test.daoristic <- chisq.test(daoristic.df$Freq, p=
    day.expected)
#X-squared = Inf, df = 47, p-value < 2.2e-16

# - RANDOM ------------------------------------#

drandom.df <- as.data.frame(table(floor(1 + (random.data.
    array)*2)))
chisq.test.drandom <- chisq.test(drandom.df$Freq, p=day.
    expected)
#X-squared = Inf, df = 47, p-value < 2.2e-16
```

B.2 Experiment 2 och 3

```
# Autor:          Olle Svenhag, Student & Martin Boldt, PhD
# Date:           2014-04-20
# E-mail:         Olle.Svenhag@hotmail.com
# Affiliation:    Blekinge Institute of Technology, Sweden

# Required libraries
library(RMySQL)
library(lubridate)
library(ggplot2)

nr.of.days.in.a.week <- 7
nr.of.days.in.a.month <- 31


############################## DATABASE (get data)
    ##############################

# Grab burglary dates and times from burgleform table
#Connect to MySQL
mmdb = dbConnect(MySQL(), user='root', password='', dbname
    ='burglary', host='127.0.0.1', port=3306, unix.socket='
    /tmp/mysql.sock')

#Send Query and save result to "data"
rs = dbSendQuery(mmdb, "select datestart, timestart,
    dateend, timeend from burgleform where datestart >
    '2013-05-01' && dateend < '2014-05-01'")

data = fetch( rs, n=-1 )

# Close db connection
dbClearResult(rs)
dbDisconnect( mmdb )

############################## DATAFRAME
    ##############################

# Setup crime.time dataframe
from     <- ymd( paste( data$datestart) )  # yyyy-mm-dd
```

```
mdfrom  <- mday( paste( data$datestart) ) # day of the
    month
wdfrom  <- wday( paste( data$datestart) ) # day of the
    week (start on sunday)

to      <- ymd( paste( data$dateend) )  # yyyy-mm-dd
mdto    <- mday( paste( data$dateend) ) # day of the month
wdto    <- wday( paste( data$dateend) ) # day of the week

range   <- as.double( difftime( to, from, units="days") )

mdaverage <- mday(from + ((range*86400)/2)) # day of the
    month
wdaverage <- wday(from + ((range*86400)/2)) # day of the
    week

# Build it
crime.time = data.frame( from, mdfrom, wdfrom, to, mdto,
    wdto, mdaverage, wdaverage, range )


############################## FUNCTIONS
    ##############################

## Randomize function ##
randomize.bin.number <- function(p.startbin, p.endbin, p.
    number){
  if(p.startbin > p.endbin)
  {
    nr.of.bins <- ((p.number - p.startbin) + p.endbin)
    random.bin.nr <- p.startbin + sample(0:nr.of.bins, 1)
    if(random.bin.nr > p.number) {
      res <- (random.bin.nr - p.number)
    } else {
      res <- (random.bin.nr)
    }

  } else if(p.startbin < p.endbin) {

    res <- sample(p.startbin:p.endbin, 1)

  } else if(p.startbin == p.endbin){
    res <- p.startbin
```

```
  }else{
    eexxxxxxxxxxxxxxit
  }

  return(res)
}

############################ ACTUAL TIME
    ############################

actual.time.dataset = subset( crime.time, crime.time$range
    == 0 ) #Only leave crimes with range less than 24hrs


############################## RANDOM
    ##############################

## Day of month
    ################################################################

mday.random.array <- array( 0, dim=c(nr.of.days.in.a.month
    ) )

for( counter in 1:nrow( crime.time ) ) {
  tmp.from <- crime.time[ counter, 2 ]
  tmp.to <- crime.time[ counter, 5 ]

  tmp.random.binnumber <- randomize.bin.number( tmp.from,
      tmp.to, nr.of.days.in.a.month )

  mday.random.array[tmp.random.binnumber] <- (mday.random.
      array[tmp.random.binnumber] + 1)

}
#print( mday.random.array )
# TEST PLOTTING Random: Day of month # ------------------
#qplot(1:nr.of.days.in.a.month, mday.random.array, geom="
    bar", width=1, stat="identity")
# ------------------------------------------------------
```

```
## Day of week
    ##################################################################

random.dataset = subset( crime.time, crime.time$range <=
    nr.of.days.in.a.week) #Remove all with range > a week

wday.random.array <- array( 0, dim=c(nr.of.days.in.a.week)
    )

for( counter in 1:nrow( random.dataset ) ) {
  tmp.from <- random.dataset[ counter, 3 ]
  tmp.to <- random.dataset[ counter, 6 ]

  tmp.random.binnumber <- randomize.bin.number( tmp.from,
      tmp.to, nr.of.days.in.a.week )

  wday.random.array[tmp.random.binnumber] <- (wday.random.
      array[tmp.random.binnumber] + 1)

}
#print( wday.random.array )
# TEST PLOTTING Random: Day of Week # ----------------------
#qplot(1:nr.of.days.in.a.week, wday.random.array, geom="
    bar", stat="identity")
# ----------------------------------------------------------

############################## AORISTIC
    ##############################

## Day of month
    ##################################################################

mday.aoristic.array <- array( 0, dim=c(nr.of.days.in.a.
    month) )

for( counter in 1:nrow( crime.time ) ) {
  tmp.fraction <- 1 / (crime.time[ counter, 9 ] + 1)
  tmp.from <- crime.time[ counter, 2 ]
  tmp.to <- crime.time[ counter, 5 ]

  if( tmp.from > tmp.to) {
    mday.aoristic.array[tmp.from:nr.of.days.in.a.month] <-
        mday.aoristic.array[tmp.from:nr.of.days.in.a.month
```

```
        ] + tmp.fraction
      mday.aoristic.array[1:tmp.to] <- mday.aoristic.array
        [1:tmp.to] + tmp.fraction
    }
    else {
      mday.aoristic.array[tmp.from:tmp.to] <- mday.aoristic.
        array[tmp.from:tmp.to] + tmp.fraction
    }

}
#print( mday.aoristic.array )
# TEST PLOTTING Aoristic: Day of month # ----------------
#qplot(1:nr.of.days.in.a.month, mday.aoristic.array, geom
    ="bar", width=1, stat="identity")
# ----------------------------------------------------------

## Day of week
    ##################################################################

aoristic.dataset = subset( crime.time, crime.time$range <=
    nr.of.days.in.a.week) #Remove all with range > a week

wday.aoristic.array <- array( 0, dim=c(nr.of.days.in.a.
    week) )

for( counter in 1:nrow( aoristic.dataset ) ) {
  tmp.fraction <- 1 / (aoristic.dataset[ counter, 9 ] + 1)
  tmp.from <- aoristic.dataset[ counter, 3 ]
  tmp.to <- aoristic.dataset[ counter, 6 ]

  if( tmp.from > tmp.to) {
    wday.aoristic.array[tmp.from:nr.of.days.in.a.week] <-
      wday.aoristic.array[tmp.from:nr.of.days.in.a.week]
      + tmp.fraction
    wday.aoristic.array[1:tmp.to] <- wday.aoristic.array
      [1:tmp.to] + tmp.fraction
  }
  else {
    wday.aoristic.array[tmp.from:tmp.to] <- wday.aoristic.
      array[tmp.from:tmp.to] + tmp.fraction
  }

}
```

```
#print( wday.aoristic.array )
# TEST PLOTTING Aoristic: Day of Week # ------------------
#qplot(1:nr.of.days.in.a.week, wday.aoristic.array, geom="
    bar", stat="identity")
# ------------------------------------------------------------


############################## PLOTTING
    ##############################

## --[[ DAY OF MONTH ]]-- ## ////////////////////////////
    / ##

## FROM ##
ggplot(crime.time, aes(factor(x=mdfrom))) + geom_histogram
    (binwidth = 1) + xlab("manadsdag") + ylab("Antal␣
    bostadsinbrott") + ggtitle("Start")

## TO ##
ggplot(crime.time, aes(factor(x=mdto))) + geom_histogram(
    binwidth = 1) + xlab("manadsdag") + ylab("Antal␣
    bostadsinbrott") + ggtitle("Slut")

## AVERAGE ##
ggplot(crime.time, aes(factor(x=mdaverage))) + geom_
    histogram(binwidth = 1) + xlab("manadsdag") + ylab("
    Antal␣bostadsinbrott") + ggtitle("Medel")

## AORISTIC ##
qplot(factor(1:nr.of.days.in.a.month), mday.aoristic.array
    , geom="bar", binwidth=1, stat="identity", xlab="
    manadsdag", ylab="Antal␣bostadsinbrott") + ggtitle("
    Aoristic")

## RANDOM ##
qplot(factor(1:nr.of.days.in.a.month), mday.random.array,
    geom="bar", binwidth=1, stat="identity", xlab="
    manadsdag", ylab="Antal␣bostadsinbrott") + ggtitle("
    Slump")

## ACTUAL TIME ##
ggplot(actual.time.dataset, aes(factor(x=mdfrom))) + geom_
    histogram(binwidth = 1) + xlab("manadsdag") + ylab("
```

```
    Antal_bostadsinbrott") + ggtitle("Exakt_tid")


## --[[ DAY OF WEEK ]]-- ## ////////////////////////////
    ##

## FROM ##
ggplot(crime.time, aes(factor(x=wdfrom))) + geom_histogram
    (binwidth = 1) + xlab("Veckodag") + ylab("Antal_
    bostadsinbrott") + ggtitle("Start")

## TO ##
ggplot(crime.time, aes(factor(x=wdto))) + geom_histogram(
    binwidth = 1) + xlab("Veckodag") + ylab("Antal_
    bostadsinbrott") + ggtitle("Slut")

## AVERAGE ##
ggplot(crime.time, aes(factor(x=wdaverage))) + geom_
    histogram(binwidth = 1) + xlab("Veckodag") + ylab("
    Antal_bostadsinbrott")  + ggtitle("Medel")

## AORISTIC ##
qplot(factor(1:nr.of.days.in.a.week), wday.aoristic.array,
     geom="bar", binwidth=1, stat="identity", xlab="
    Veckodag", ylab="Antal_bostadsinbrott") + ggtitle("
    Aoristic")

## RANDOM ##
qplot(factor(1:nr.of.days.in.a.week), wday.random.array,
    geom="bar", binwidth=1, stat="identity", xlab="Veckodag
    ", ylab="Antal_bostadsinbrott") + ggtitle("Slump")

## ACTUAL TIME ##
ggplot(actual.time.dataset, aes(factor(x=wdfrom))) + geom_
    histogram(binwidth = 1) + xlab("Veckodag") + ylab("
    Antal_bostadsinbrott") + ggtitle("Exakt_tid")


############################## Chi Square (WEEK)
    ##############################


#Actual Time (vector)
```

```
week.actual.time <- actual.time.dataset$wdfrom
#Count total number of data
length(week.actual.time)
#Make frequency table
week.actual.time.freq <- as.data.frame(table(week.actual.
    time))
#Insert another colom with percentage value (probability)
week.actual.time.freq$probability <- week.actual.time.freq
    $Freq / length(week.actual.time)

#Sum of the probability should always be 1 (confirmed
    below)
#sum(week.actual.time.freq$probability)

#Now we have the expected value to use in our Chi square
    test.
week.expected <- c(week.actual.time.freq$probability)

# - FROM ----------------------------------------#

wdfrom.df <- as.data.frame(table(crime.time$wdfrom))

chisq.test.wdfrom <- chisq.test(wdfrom.df$Freq, p=week.
    expected)
#X-squared = 11.1425, df = 6, p-value = 0.08407

# - TO ------------------------------------------#

wdto.df <- as.data.frame(table(crime.time$wdto))

chisq.test.wdto <- chisq.test(wdto.df$Freq, p=week.
    expected)
#X-squared = 33.3337, df = 6, p-value = 9.044e-06

# - AVERAGE -------------------------------------#

wdaverage.df <- as.data.frame(table(crime.time$wdaverage))

chisq.test.wdaverage <- chisq.test(wdaverage.df$Freq, p=
    week.expected)
#X-squared = 14.8935, df = 6, p-value = 0.0211

# - RANDOM --------------------------------------#
```

```
chisq.test.wdrandom <- chisq.test(wday.random.array, p=
    week.expected)
#X-squared = 11.951, df = 6, p-value = 0.06307

# - AORISTIC ----------------------------------#

chisq.test.wdaoristic <- chisq.test(wday.aoristic.array, p
    =week.expected)
#X-squared = 9.215, df = 6, p-value = 0.1618




############################## Chi Square (MONTH)
    ##############################

#Actual Time (vector)
month.actual.time <- actual.time.dataset$mdfrom
#Count total number of data
length(month.actual.time)
#Make frequency table (note: day 31 only exists 7 times
    per year)
month.actual.time.freq <- as.data.frame(table(month.actual
    .time))
#Insert another colom with percentage value (probability)
month.actual.time.freq$probability <- month.actual.time.
    freq$Freq / length(month.actual.time)

#Sum of the probability should always be 1 (confirmed
    below)
#sum(month.actual.time.freq$probability)

#Now we have the expected value to use in our Chi square
    test.
month.expected <- c(month.actual.time.freq$probability)

# - FROM --------------------------------------#

mdfrom.df <- as.data.frame(table(crime.time$mdfrom))
```

```
chisq.test.mdfrom <- chisq.test(mdfrom.df$Freq, p=month.
    expected)
#X-squared = 17.0038, df = 30, p-value = 0.9725

# - TO ----------------------------------------#

mdto.df <- as.data.frame(table(crime.time$mdto))

chisq.test.mdto <- chisq.test(mdto.df$Freq, p=month.
    expected)
#X-squared = 12.7588, df = 30, p-value = 0.9975

# - AVERAGE -----------------------------------#

mdaverage.df <- as.data.frame(table(crime.time$mdaverage))

chisq.test.mdaverage <- chisq.test(mdaverage.df$Freq, p=
    month.expected)
#X-squared = 16.7608, df = 30, p-value = 0.9753

# - RANDOM ------------------------------------#

chisq.test.mdrandom <- chisq.test(mday.random.array, p=
    month.expected)
#X-squared = 9.1072, df = 30, p-value = 0.9999

# - AORISTIC ----------------------------------#

chisq.test.mdaoristic <- chisq.test(mday.aoristic.array, p
    =month.expected)
#X-squared = 5.6854, df = 30, p-value = 1
```