



UPPSALA  
UNIVERSITET

UPTEC-STS-14003

Examensarbete 30 hp  
Februari 2014

# Identifikation av icke-representativa svar i frågeundersökningar genom detektion av multivariata avvikare

---

Hugo Galvenius



UPPSALA  
UNIVERSITET

Teknisk- naturvetenskaplig fakultet  
UTH-enheten

Besöksadress:  
Ångströmlaboratoriet  
Lägerhyddsvägen 1  
Hus 4, Plan 0

Postadress:  
Box 536  
751 21 Uppsala

Telefon:  
018 – 471 30 03

Telefax:  
018 – 471 30 00

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

## Abstract

Identifikation av icke-representativa svar i frågeundersökningar genom detektion av multivariata avvikare

### **Identification of non-representative survey responses through multivariate outlier detection**

*Hugo Galvenius*

To United Minds, large-scale surveys are an important offering to clients, not least the public opinion poll Väljarbarometern. A risk associated with surveys is satisficing – sub-optimal response behaviour impairing the possibility of correctly describing the sampled population through its results. The purpose of this study is to – through the use of multivariate outlier detection methods - identify those observations assumed to be non-representative of the population. The possibility of categorizing responses generated through satisficing as outliers is investigated. With regards to the character of the Väljarbarometern dataset, three existing algorithms are adapted to detect these outliers. Also, a number of randomly generated observations are added to the data, by all algorithms correctly labelled as outliers. The resulting anomaly scores generated by each algorithm are compared, concluding the Otey algorithm as the most effective for the purpose, above all since it takes into account correlation between variables. A plausible cut-off value for outliers and separation between non-representative and representative outliers are discussed. The resulting recommendation is to handle observations labelled as outliers through respondent follow-up or if not possible, through downweighting, inversely proportional to the anomaly scores.

Handledare: Markus Larsson  
Ämnesgranskare: Jesper Rydén  
Examinator: Elisabet Andrésdóttir  
ISSN: 1650-8319, UPTEC STS14 003

## Sammanfattning

För United Minds är storskaliga frågeundersökningar, inte minst opinionsundersökningen Väljarbarometern, en viktig del av kunderbjudandet. En risk med frågeundersökningar är så kallad *satisficing* – suboptimalt svarsbeteende som försämrar möjligheten att rättvist beskriva populationen med hjälp av dess resultat. Undersökningens syfte är att med hjälp av metoder för multivariat detektion av avvikare identifiera de observationer som kan antas vara icke-representativa för populationen. Möjligheten att kategorisera svar som utsatts för *satisficing* som avvikare undersöks, och utifrån de förutsättningar som gäller för datamaterialet för Väljarbarometern anpassas tre existerande algoritmer för detektion av dessa. Till datamaterialet adderas även slumpmässigt genererade observationer, som av samtliga algoritmer korrekt flaggas som avvikare. De resulterande anomalipoäng som varje algoritm genererar jämförs sinsemellan och den så kallade Oteyalgoritmen anses vara den mest lämpliga för ändamålet, framförallt då den även tar i beräkning korrelation mellan variabler. Diskussion förs rörande avskärningsvärde för avvikare och särskiljning mellan icke-representativa och representativa avvikare vilket leder till rekommendationen att de av algoritmen flaggade avvikarna bör hanteras genom uppföljning av respondenter, eller om detta ej är möjligt genom nedviktning, omvänt proportionerligt mot dess anomalipoäng.

# Innehållsförteckning

<b>1. Inledning</b>	<b>2</b>
1.1. Bakgrund	3
1.2. Syfte	3
1.3. Avgränsningar	4
<b>2. Relaterad forskning</b>	<b>5</b>
2.1. Satisficing	5
2.2. Avvikare	6
2.2.1. Definition	6
2.2.2. Applikationer för detektion av avvikare	6
2.2.3. Typ av data	7
2.2.4. Metoder för detektion av avvikare	8
2.2.5. Särskiljning av avvikare	9
2.2.6. Hantering av avvikare	9
<b>3. Algoritmer</b>	<b>11</b>
3.1. Otey	11
3.2. Greedy	12
3.3. AVF	13
<b>4. Experiment</b>	<b>15</b>
4.1. Datamaterial	15
4.2. Resultat och diskussion	15
4.2.1. Med slumpade observationer	16
4.2.2. Otey	17
4.2.2. Greedy	21
4.2.3. AVF	23
4.2.4. Jämförelse	27
4.2.5. Representativitet	31
4.2.6. Hantering	31
<b>5. Slutsatser</b>	<b>32</b>
<b>Referenser</b>	<b>34</b>
Böcker	34
Vetenskapliga artiklar	34
Tidningsartiklar	36

# 1. Inledning

En uppföljande mätning av en opinionsundersökning bland drygt tvåtusen svenskar gav uppseendeväckande resultat. Arton månader efter den första undersökningen uppgav 0,7 % av de som ursprungligen sagt sig vara kvinnor, att de nu var av manligt kön! Motsvarande andel för könsbyte från man till kvinna var 0,2 %. Media rapporterar visserligen om en ökning i antalet könsbyten i landet – men då talas det snarare om att det nationellt genomförs ett sextiotial årligen, medan opinionsundersökningen föreslår att antalet skulle vara uppåt femtiotusen. En möjlig orsak till opinionsundersökningens överdrivna resultat skulle kunna vara att det tiotal personer som i undersökningarna uppgett olika kön hör till de knappt hundra personer som uppskattas byta kön under en 18-månadersperiod. Det skulle dock betyda att undersökningen lyckats fånga ca tio procent av denna grupp. Med tanke på att undersökningens totala representativitet av ett urval på tvåtusen respondenter motsvarar endast drygt 0,2 promille av Sveriges befolkning vore detta ytterst förvånande. En sådan snedvridning i urvalet är givetvis möjlig, t.ex. om överrepresentationen orsakats av att respondenter rekryterats från en webbpanel med övervägande transsexuella medlemmar - men skulle det inte kunna vara så att opinionsundersökningarna vid något eller båda tillfällena fyllts i slarvigt eller på något annat sätt felaktigt?

Tidigare studier stöder tesen att vissa respondenter slarvigt svarar på olika typer av frågeundersökningar, och har försökt kvantifiera i vilken utsträckning detta sker. Meade et al. (2012) identifierade 10-12 % av studenter som fyllde i en anonym webbenkät mot ersättning av kurspoäng som slarviga respondenter. Kurtz et al. (2001) fann motsvarande inkonsistens hos 10,6 % av respondenterna i ett urval av studenter, medan Johnson (2005) bedömde att 3,5 % av respondenterna återupprepat valde samma svarsalternativ i den psykometriska NEO-PI-R-undersökningen.

Anta att ett antal respondenter slarvigt eller felaktigt svarar på frågorna i en opinionsundersökning. För enkelhets skull kan vi föreställa oss en situation där deras svar är helt slumpmässiga och saknar korrelation med deras egentliga åsikter. Anta sedan att en av frågorna uppmanar respondenterna att svara vilket politiskt parti de sympatiserar mest med, med syftet att prognostisera ett framtida valresultat. Det är då lätt att inse att ett slumpmässigt svar gynnar resultatet för små partier, medan det missgynnar de stora. Sannolikheten är ju mindre att den egentliga sympatin var med det lilla partiet. Om ettusen respondenter tillfrågas är det också lätt att inse att varje respondent bidrar med en tiondels procentenhet till totalresultatet, givet att lika stor vikt läggs vid varje respondents svar. Detta hypotetiska resonemang används för att illustrera några konsekvenser av suboptimalt svarande och varför det är intressant att undersöka för forskare och analytiker som i sitt arbete använder sig av kvantitativa frågeundersökningar.

## 1.1. Bakgrund

För trend-, analys- och strategiföretaget United Minds är storskaliga frågeundersökningar ett viktigt kunderbjudande. Med hjälp av underleverantörers webbpanelsystem sänds frågebatterier ut via e-post och besvaras av tusentals personer. Det tar bara några dagar att samla in tusen svar från en undersökning riktad till svensk allmänhet, representativ med avseende på kön och åldersgrupp. Webbenkäter är en förhållandevis billig metod för att ställa samma frågor till ett stort antal svarande, och med hjälp av resultaten aspireras på att dra generella slutsatser avseende hela populationen. En låg kostnad samt enkel och snabb insamling av svar har gjort att metoden används flitigt i traditionella opinionsundersökningar, men även i studier beställda av företag eller offentlig förvaltning som söker insikt i sina kunders eller användares beteendemönster och preferenser.

Tillförlitlighet är en avgörande del av kunderbjudandet. Avseende opinionsundersökningar har en metoddiskussion förts mellan aktörer i branschen. Å ena sidan har leverantörer av undersökningar som samlar in svar via webbpaneler, däribland United Minds, fått utstå kritik från konkurrenter som använder sig av telefonintervjuer. Man menar att urvalet blir skevt då rekryteringen till webbpaneler inte sker slumpmässigt (DN Debatt 2012; Lund Business Review 2013). Å andra sidan påverkas respondenter i telefonintervjuer mer av den så kallade intervjuareffekten, där man i större utsträckning ger socialt önskvärda svar jämfört med när man fyller i en webbenkät (Chang et al. 2009). Det finns alltså en pågående debatt om tillförlitlighet i opinionsundersökningar, och anledning att ständigt arbeta för att förbättra dess kvalitet.

En faktor som påverkar undersökningars resultat och tillförlitlighet är förekomsten av *satisficing*, då respondenter inte noggrant överväger frågor och svarsalternativ, utan slarvar och därmed ger suboptimala svar. Barge et al. (2012) konkluderade efter en undersökning av *satisficing* att beteendet har en mätbar påverkan på enkätresultat, och påpekade att insamlingsmetoder via onlineverktyg är mer utsatta för beteendet. Däremot ansågs dessa verktyg ge bättre möjligheter att utvärdera respondenters engagemang genom att inledningsvis identifiera och isolera respondenter som hänger sig till suboptimalt svarsbeteende. Detta för att sedan kunna utvärdera den påverkan beteendet har på datamaterialet.

## 1.2. Syfte

Denna undersökning tar avstamp i nyligen nämnda insikter och har syftet att identifiera och isolera icke-representativa svar i frågeundersökningar. För att uppnå detta ämnas följande frågeställningar besvaras:

- Vilka svar kan antas vara icke-representativa för den undersökta populationen och hur skiljer de sig från representativa svar?
- Vilka metoder existerar för detektion av avvikande observationer i stora datamaterial, och vilka passar bäst för ändamålet?
- Vilka algoritmer är användbara för ändamålet och vad genererar de för resultat om anpassade för existerande data från frågeundersökningar?
- Hur tolkas resultatet för att särskilja icke-representativa svar från representativa?
- Hur och varför skiljer sig resultat åt från olika algoritmer och vilken algoritm passar bäst för ändamålet?
- Hur hanteras avvikande observationer som antas vara icke-representativa?

- Vilka implikationer har resultatet för distribution och analys av frågeundersökningar?

### 1.3. Avgränsningar

Vissa nödvändiga avgränsningar har gjorts för att uppnå syftet och begränsa undersökningens omfattning.

De algoritmer som anpassas och utvärderas är sådana som lämpar sig för karaktären på testdata och för den typ av data som resultatet förväntas tillämpas på. Därför utvärderas och anpassas endast algoritmer som hanterar kvalitativ (kategorisk) data. Består datamaterialet av en övervägande andel kvantitativ data, t.ex. påståenden om beteende som respondenter besvarat på en Likertskala, skulle andra metoder troligtvis bättre detektera avvikande svar. Vidare bygger algoritmerna på multivariat analys av datamaterial och kräver därför ett stort antal observationer,  $N$ , och flera variabler,  $M$ , för att vara användbara. De algoritmer som jämförs är ett urval av de algoritmer som tillämpas i Koufakou et al.'s (2007) undersökning. Fler algoritmer existerar och jämförelsen hade kunnat göras än mer utförlig.

Vid tolkningen av resultaten från de olika algoritmerna, där observationer tilldelats en poäng med avseende på hur mycket de avviker från övrig data, krävs vetenskapen om förväntat antal avvikande observationer eller att något typ av avskärningsvärde fastställs. Många tillämpningar (t.ex. Koufakou et al. 2007) förutsätter att antalet avvikelser är känt, men för tillämpning på resultat från frågeundersökningar skulle endast en vag uppskattning av detta kunna göras. Istället fastställs ett avgränsningsvärde med avseende på karaktären på observationernas spridning. Detta tillvägagångssätt anses vara nödvändigt för att kunna dra användbara slutsatser av resultatet.

Undersökningen studerar övergripande mönster i datamaterialet och gör anspråk på att särskilja de observationer som avviker från andra. Företrädevis eftersöks någon generisk metod för detektion av avvikelser, som skulle kunna appliceras på likartade datamaterial. Varför observationer avviker, dvs. vilka svar respondenten gett som ger upphov till avvikelser, vore självklart intressant att detaljstudera. Det ligger dock utanför ramen för denna undersökning. Någon känslighetsanalys av enskilda observationers påverkan på resultatet genomförs inte heller. Detta skulle kunna vara mycket användbart i diskussionen hur avvikande observationer bör hanteras, då observationer med betydande påverkan torde vara mest kritiska att behandla. Dessa mer detaljerade analyser av enskilda observationer ligger utanför undersökningens omfattning.

Endast en översiktlig jämförelse görs med avseende på komplexitet och tidsåtgång för de olika algoritmerna, men någon noggrann utvärdering av detta undviks. Avsikten är inte att ranka de snabbaste algoritmerna, och i anpassningen av dessa har säkerligen suboptimala programmeringslösningar resulterat i avsevärt lägre effektivitet än vad som är möjligt. De uppmätta tiderna ger endast en fingervisning om tidsåtgången för respektive algoritm.

## 2. Relaterad forskning

Här följer en genomgång av de teoretiska koncept vilka undersökningen bygger på. Inledningsvis redogörs för suboptimalt svarande i frågeundersökningar genom *satisficing*. Detta följs av en beskrivning av begreppet avvikare, hur de definieras, identifieras och hanteras under olika förutsättningar. Dessa koncept ligger till grund för applikationen av de algoritmer som presenteras i nästa del av undersökningen.

### 2.1. Satisficing

En risk med frågeundersökningar, vare sig de distribueras i pappersform, via internet eller genomförs per telefon, är så kallad *satisficing*<sup>1</sup>. Begreppet myntades av Herbert Simon som en generell beteckning på suboptimalt beslutsfattande men det var Krosnick (1991) som förädlade konceptet med avseende på frågeundersökningar. Han menade att detta beteende kan uppstå då optimala svar på en frågeundersökning kräver en betydande kognitiv ansträngning. Istället finns risken att vissa respondenter endast ger tillfredställande<sup>2</sup> men suboptimala svar (Krosnick 1991). Ett i sammanhanget optimalt svar kräver en betydande ansträngning i flera faser. Respondenten ska noggrant tolka meningen med frågan, göra en betydande sökning i det egna minnet efter relevant information, varsamt integrera denna information till sammanfattande bedömningar och slutligen svara på det sätt som tydligast uttrycker dessa bedömningar (Krosnick et al. 1996). Om detta inte genomförs är risken att suboptimala svar ges, t.ex. genom att respondenten väljer det första svarsalternativ som tycks överensstämma, underlåter att skilja på olika svarsalternativ, svarar ”vet ej” istället för att uttrycka sin egentliga åsikt, eller slumpmässigt väljer svarsalternativ (Krosnick 1991). Man menade att tre faktorer påverkar sannolikheten att en respondent ger efter för *satisficing*: frågans svårighetsgrad, respondentens förmåga att genomföra uppgiften samt respondentens motivation att genomföra uppgiften, dvs. svara på enkäten (Krosnick et al. 1996).

Sedan Krosnick applicerade begreppet *satisficing* för att beskriva icke önskvärda beteenden vid frågeundersökningar har flera artiklar, framförallt inom folkopinions- och utbildnings-vetenskap, skrivits med syftet att utvärdera kvaliteten av data samt identifiera metoder för undersökningens design som lindrar *satisficing*. Exempelvis har redogjorts för *straight-lining* – respondenters benägenhet att svara på frågor enligt ett givet mönster (Cole et al. 2012; Kaminska et al. 2010), hur svarstiden påverkar kvaliteten (Christensen et al. 2012; Malhotra 2008), samt hur incitament påverkar benägenheten för *satisficing* (Barge et al. 2012). Dessutom har metoder presenterats för att säkerställa att respondenter följer instruktioner, eller gallrar bort dem som inte gör det (Kapelner et al. 2010; Oppenheimer et al. 2009). Ingen överenskommelse rörande vilken metod som är mest framgångsrik för att säkerställa högre datakvalitet tycks dock ha nåtts.

---

<sup>1</sup> Begreppet tolkas analogt med vad bl.a. Meade et al. (2012), Johnson (2005) och Kurtz et al. (2001) kallar slarvigt svarsbeteende (*carelessness*).

<sup>2</sup> Eng: satisfactory



## 2.2. Avvikare

Problemet att identifiera *satisficing* i frågeundersökningar skulle även kunna omformuleras till ett statistik- och informationsutvinningsproblem. Detta enligt tolkningen att det suboptimala svarsbeteendet kan leda till paradoxala svar, eller svarmönster som kraftigt skiljer sig från det normala. Under detta antagande definieras problemet att identifiera så kallade avvikare<sup>3</sup> i ett datamaterial.

### 2.2.1. Definition

Hawkins (1980) föreslår en generell definition av en avvikare som:

*“(...) en observation som avviker så mycket från andra observationer att det finns anledning att misstänka att den genererats av en annan mekanism.”*  
(egen översättning)

För surveydata definierar Chambers (1986) två möjliga typer av avvikare. Representativa avvikare är korrekta mätvärden som avviker från resten av urvalet men för vilka det bör finnas motsvarande värden i populationen. Icke-representativa avvikare är de vars värden till exempel genom felaktig kodning är inkorrekta eller inom populationen unika så att de inte bidrar till en representativ beskrivning av populationen. Vid analys av avvikare i surveydata bortses ofta från de icke-representativa (t.ex. Chambers 1986; Chambers et al. 2004).

Är avvikare genererade som resultat av *satisficing* torde de enligt Chambers klassificering vara icke-representativa och även överensstämma med Hawkins generella definition.

### 2.2.2. Applikationer för detektion av avvikare

Applikationer för identifikation av avvikare förekommer i vitt skilda fält, möjliggjort av en ökad mängd insamlad data och ökad datorkraft som tillåter mer komplexa och kraftfulla algoritmer. Hodge et al. (2004) ger exempel på applikationer för att hitta tillverkningsfel inom produktion, genom bildanalys hitta anomalier och identifiera landminor, detektera obehöriga inkräktare i nätverk, identifiera potentiellt problematiska låntagare eller avslöja avvikande beteende som indikerar kreditkorts- eller försäkringsbedrägeri.

Detektion av avvikare, och förslag på hantering av desamma har även gjorts avseende surveydata. Till exempel används av den kanadensiska statistiska centralbyrån Hidioglou et al.:s (1986) metod för redigering av datapunkter för att minska inverkan från avvikare i periodisk data insamlad från företag. Även Ishikawa et al. (2010) undersöker anomalier i ekonomisk data från företag, genererad av den japanska centralbankens årliga ”Tankan”-undersökning. Olika metoder för att hantera endimensionella avvikare, och dess inverkan på resultatet diskuteras.

---

<sup>3</sup> Eng: outliers

### 2.2.3. Typ av data

En distinktion av datamaterialet görs med avseende på om tidigare kunskap om vad som anses som normala respektive onormala observationer. I idealfallet finns ett träningsdata tillgängligt där observationerna är flaggade med statusen normal eller onormal. Genom efterföljande genomkörningar av ny data kan avvikare identifieras med hjälp av denna tidigare information. Detta angreppssätt kallas övervakad inlärning<sup>4</sup>. Det motsatta angreppssättet används då tidigare kunskap om observationernas status saknas, så kallad oövervakad inlärning<sup>5</sup>. Datan antas då ofta följa någon uppskattad fördelning, och de observationer som avviker från vad som kan anses vara "normalt" för denna fördelning flaggas som möjliga avvikare. Vidare kan semiövervakad inlärning användas om en del av datamaterialet är tidigare klassificerad (Hodge et al. 2004).

Andra karakteristika hos datamaterialet som avgör vilken metod som lämpar sig för detektion av avvikare är antalet observationer  $N$ , antalet variabler (dimensioner)  $M$  samt vilken typ av mätskala variablerna tillhör, dvs. om det rör sig om kvalitativa (kategoriska) variabler, kvantitativa variabler eller en kombination av de två. Variabler som representerar olika kategorier som saknar inbördes rangordning, t.ex. partisympati eller kön, är kvalitativa och representeras av en nominalskala. De observationer som för den kvalitativa variabeln tillhör samma kategori ges samma värde, t.ex. en siffra. Det skulle dock sakna betydelse att fastställa statistiska mått som medelvärde eller median för dessa variabler. Det enda relevanta statistiska måttet är frekvenser av antal förekomster av samma värde. Kvantitativa variabler avser de som representeras numeriskt, och vars siffervärde kan ha statistisk betydelse. Den enklaste formen av kvantitativ variabel kan representeras av en ordinalskala. Värden kan rangordnas, men då intervallen mellan värdena inte kan anses vara konstanta bör inte statistiska mått som medelvärde och standardavvikelse användas. Ett exempel på ordinalskala som används i enkäter är så kallade Likertskalor, där respondenters attityder till ett påstående graderas enligt en numerisk skala. Vidare krävs för en intervallskala att den har bestämda intervall mellan möjliga värden, och de flesta statistiska mått kan därför användas på dessa. Nollpunkten på en intervallskala är inte absolut utan kan förskjutas. Det är därför meningslöst att tala om t.ex. en fördubbling av ett värde på en intervallskala, och även negativa värden är accepterade. Till sist definierar man en kvotskala om tidigare förutsättningar gäller, men även en absolut nollpunkt kan fastställas. Värden på kvotskalor kan jämföras relativt. Exempel på variabler som representeras av kvotskalor är ålder och svarstid (Stevens 1946; Likert 1932).

I det multivariata fallet är algoritmer framförallt utvecklade för att hantera kvantitativa variabler (t.ex. Filzmoser et al. 2008; Franklin et al. 1997; Loureiro et al. 2004). Som Otey et al. (2006) konstaterade innehåller de flesta datamaterial i realiteten en blandning av kvantitativa och kvalitativa variabler. Modellering som bortser från dessa kvalitativa variabler resulterar i en betydande informationsförlust.

---

<sup>4</sup> Eng: supervised learning

<sup>5</sup> Eng: unsupervised learning

#### 2.2.4. Metoder för detektion av avvikare

Hodge et al. (2004) redogör för metoder som använts för detektion av avvikare. Dessa har sitt ursprung i tre olika forskningsfält: statistik, neuronät och maskinlärning. Några hybridmetoder använder algoritmer från flera forskningsfält. Metoder baserade på statistiska modeller antar att data kan beskrivas med en fördelning och dess parametrar, t.ex. normalfördelning. För datamaterial av få dimensioner kan med fördel användas statistiska grafiska metoder för att identifiera avvikande värden, t.ex. är lådagram en vanligt förekommande teknik för identifikation av avvikare i det endimensionella fallet (Hodge et al. 2004). Statistiska metoder används oftast i fådimensionella dataset, då fler dimensioner gör dem känsliga för "the Curse of Dimensionality". För att lindra detta problem kan t.ex. principalkomponentsanalys göras för att projicera mot de variabler som mest bidrar till variansen i data och på så sätt minska antalet dimensioner (Filzmoser et al 2008). Distansbaserade metoder beräknar olika mått på proximitet mellan observationer, t.ex. euklidisk eller mahalnobisk distans, och flaggar som avvikare de observationer som har stort avstånd från någon referenspunkt. Exempelvis använder *nearest neighbour*-tekniker medelavståndet till de närmaste observationerna som ett mått på avvikelse. Identifikation av avvikare genom distansbaserade metoder har visats utförbar och meningsfullt även för multivariata, storskaliga datamaterial innehållande kvantitativa variabler (Knorr et al. 2000). Densitetsbaserade metoder uppskattar täthetsfördelningen i datamaterialet och flaggar som avvikare de observationer som förekommer i områden med låg densitet (t.ex. Breunig et al. 2000). Från neuronätsforskningen härstammar bland annat supportvektormetoder för detektion av avvikare, och maskinlärningen bidrar med t.ex. beslutsträd för samma ändamål (Hodge et al. 2004). Många av nämnda metoder är dock anpassade för kvantitativa variabler där någon slags rangordning av värden är meningsfullt. I fallet med kvalitativ data saknas mening i att rangordna värden och beräkna t.ex. distansmått dem emellan. Andra metoder kräver upprepade bearbetningar av datamaterialet för att utvinna avvikare vilket gör dem ineffektiva för stora, multidimensionella datamaterial (Koufakou et al. 2007).

Tidigare forsknings fokus på detektion av avvikare i datamaterial med kvantitativa variabler har senare ifrågasatts, med motiveringen av många applikationer innehåller data med övervägande kvalitativa variabler, eller en kombination av de två. Koufakou et al. (2007) uppmärksammar den bristande forskningen rörande detektion av avvikare i datamaterial med kvalitativa variabler. De presenterar exempel på algoritmer av bland annat He et al. (2006) och Otey et al. (2006) som behandlar kvalitativa variabler eller en kombination av kvalitativa och kvantitativa variabler. De presenterar även en egen algoritm. En jämförande undersökning av dessa algoritmer görs med avseende på precision, tidsåtgång och skalbarhet i att detektera avvikare i kvalitativ data. Jämförelsen genomförs genom att låta algoritmerna behandla ett antal riktiga datamaterial, t.ex. ett antal patientjournaler där algoritmerna särskiljer elakartade tumörer från godartade baserade på olika attribut hos patienterna. Tidsåtgång och skalbarhet testades även genom att låta algoritmerna behandla simulerade datamaterial där antalet observationer, variabler och sökta avvikare varierades. Algoritmerna bygger på olika metoder, baserade på teori om *frequent itemsets*, entropi respektive frekvens av variabelvärde (Koufakou et al. 2007). De ansågs ge tillfredställande resultat och är därför intressanta att undersöka i detalj för användning av identifikation av avvikande svar i frågeundersökningar.

### 2.2.5. Särskiljning av avvikare

Little et al. (1987) studerar olika metoder för att approximera fördelningen av multivariat surveydata. Man diskuterar olika antaganden gällande fördelningen och baserar sina resultat på antagandet att representativa observationer har en viss fördelning, medan avvikare följer en annan fördelning. Vidare ges förslag på hur robusta uppskattningar beräknas för respektive fördelning för att kunna identifiera avvikare som sedan detaljstuderas och redigeras om de anses felaktiga. Då studien endast intresserar sig för kvantitativ data är metoderna för uppskattningar av fördelningarnas parametrar inte användbara för undersökningen som följer. Dock är av intresse att notera att det för surveydata anses relevant att göra uppskattningar om fördelningen av multivariata anomalimått, och flagga som avvikare de observationer som skiljer sig märkvärt från denna fördelning.

För att fastställa det avgränsningsvärde som särskiljer avvikare från ”normal” data föreslår Amidan et al. (2005) användning av Chebyshevs olikhet. Ofta är fördelningen av data okänd, eller det saknas annan information som möjliggör fastställande av pålitliga undre och övre gränser. De nödvändiga förutsättningarna för tillämpning av Chebyshevs olikhet är att datamaterialets observationer är oberoende och att en relativt liten andel av dem är avvikare (Amidan et al. 2005). Metoden anges inte vara skapad för kvalitativa datamaterial, men då undersökningens uppmärksammade algoritmer alla beräknar ett kvantitativt mått på samlad avvikelse, en anomalipoäng, antas Chebyshevs olikhet vara tillämpbar även för denna typ av data.

Chebyshevs olikhet är designad för att fastställa ett undre gränsvärde för andelen data som ligger inom  $k$  standardavvikelser från medelvärdet. Olikheten uttrycks enligt formeln:

$$P(|X - \mu| \leq k\sigma) \geq \left(1 - \frac{1}{k^2}\right) \quad (1)$$

där  $X$  är en slumpvariabel vars fördelning är en god beskrivning av data samt  $\mu$  och  $\sigma$  dess medelvärde respektive standardavvikelse. Ett värde  $p$  bestäms för att avgöra vilka datapunkter som är potentiella avvikare. Detta värde borde vara större än sannolikheten att stöta på en förväntad avvikare. Rimliga värden för  $p$  är 0,10, 0,05 eller 0,01. Olikheten anpassas ytterligare om datans spridning visas vara, eller antas vara unimodal, dvs. endast ha en topp. Följande ekvation används då för att beräkna  $k$  (Amidan et al. 2005):

$$k = \frac{2}{3\sqrt{p}} \quad (2)$$

Antaget att man förväntar sig att färre än fem procent av observationerna är avvikare så vore med tidigare resonemang rimligt att ansätta  $p = 0.05$ . Detta ger för unimodal data, vilket antas för genererade anomalipoäng i denna undersökning, att  $k \approx 3$ . Betydelsen av detta är att observationer som faller utanför intervallet  $\mu \pm 3\sigma$ , där medelvärde respektive standardavvikelse approximeras till urvalets medelvärde respektive standardavvikelse, kan flaggas som avvikare.

### 2.2.6. Hantering av avvikare

Avvikare kan genereras genom felaktiga mätinstrument, naturliga avvikelser i populationer, bedrägligt beteende, felaktigheter i system eller mänskliga misstag.

Hantering av avvikare beror på hur de antas ha genererats samt på tillämpningsområdet. Beror det på ett inmatningsfel kan misstaget åtgärdas genom att mata in korrekt värde, observationer genererade med mätfel kan avfärdas, medan naturliga avvikelser i populationen som flaggar som avvikare idealiskt sett sedan verifieras som korrekta och behålls i bearbetningen (Hodge et al. 2004).

Barnett (1978) beskriver fyra möjliga ansatser för att hantera avvikare: anpassning, införlivande, identifiering eller avfärdande. (i) Man kan *anpassa* analysen och skydda sig mot avvikare genom användning av robusta metoder. Exempelvis är medianen ett mer robust mått än medelvärdet vid tillämpning på en kvantitativ variabel. En annan metod för anpassning av kvantitativa variabler är Winsorisering, där en avvikares värde ersätts med värdet av en bestämd percentil av datan. (ii) Att *införliva* en avvikare innebär att behandla den som den gör i nuläget, dvs. anta att den är giltig. Istället revideras modellen för populationens fördelning för att innefatta även avvikare. (iii) Vidare tillämpas *identifiering* om den avvikande observationen detaljstuderas för att förklara någon viktig egenskap i populationen. Det kan leda till att alternativa modeller testas för att förklara dess förekomst, eller till att nya undersökningar baseras på någon passande modell för en population som genererar avvikare. (iv) Slutligen kan de *avfärdas*, något Barnett (1978) menar att många ser som den enda möjligheten för behandling av motstridiga observationer.

Eltinge et al. (2006) diskuterar olika metoder för behandling av inflytelserika avvikare i enkäter som samlar in data om organisationer. Förutom tidigare nämnd Winsorisering diskuteras även att lägga mindre vikt vid de observationer som identifierats som avvikare. Det anses vara till viss del problematiskt att applicera för multivariata avvikare. Tolkningen är att undersökningen antar att avvikare genereras av misstag, t.ex. via inmatningsfel, och att man därför felaktigt skulle vikta ner även korrekta svar från samma observation. Om antagandet att en misstänkt observation innehåller övervägande felaktigheter torde nedviktning inte vara lika problematisk.

Ytterligare en metod för hantering av avvikare som bygger på antagandet att felaktig data producerats oavsiktligt är uppföljning av respondenter, förslagsvis under så kallad selektiv redigering. Battipaglia (2002) applicerar denna metod på årliga surveyundersökningar av företag. Misstänkta svar identifieras, men då uppföljning av respondenter kan vara kostsam fokuserar man på att följa upp de uppgifter som orsakar störst förändringar på totalresultatet om ersätta med modellbestämda förväntade värden.

Barge et al. (2012) diskuterar problematiken med *satisficing* i frågeundersökningar (analogt med denna undersöknings tolkning av avvikare) och medger att det är oklart vad man ska göra med de observationer man identifierat. Åtminstone föreslår man att resultatet analyseras både med och utan de observationer som tros genererats via *satisficing* för att undersöka dess eventuella påverkan. De föreslår även en eventuell viktning av observationerna, omvänt proportionerligt med graden av *satisficing*, för att på så sätt minska deras inverkan på totalresultatet.

## 3. Algoritmer

Här följer en beskrivning av de algoritmer som ligger till grund för bearbetning av datamaterialet samt jämförelse av egenskaper och resultat. Urvalet av algoritmer baseras på den jämförelse som presenteras av Koufakou et al. (2007).

### 3.1. Otey

Algoritmen bygger på konceptet utvinning av associationsregler<sup>6</sup> som vidareutvecklades av Agrawal et al. (1994). Utgångspunkten för deras forskning var modeller för beslutsstöd för detaljhandelsorganisationer, då ny teknologi för scanning av streckkoder möjliggjort insamling och lagring av stora mängder försäljningsdata. Kundens varukorgar analyseras, och ett exempel på en associationsregel som genereras är att 98 % av kunder som köpt bildäck och reservdelar i samma transaktion även köpt bilservice. Att identifiera den här typen av regler anses användbart för ökad förståelse av kundens konsumtionsmönster och möjligheten att skraddarsy kampanjer och erbjudanden för att öka försäljningen. Man presenterade själva den egenutvecklade algoritmen Apriori som visade sig överträffa tidigare algoritmer för identifikation av samtliga associationsregler i ett datamaterial. Algoritmens inledande steg, och av vikt för denna undersökning, är att den finner alla *frequent itemsets* - de *itemsets* som är gemensamma för minst en förutbestämd andel (*minsup*) av alla transaktioner (Agrawal et al. 1994).

Otey et al. (2006) använder sig av Apriori för att generera ett datamaterials alla *frequent itemsets*, *FIS*, upp till en viss storlek *MAXLEN*. För varje enskild observation applicerar de sedan återigen Apriori för att identifiera de *itemsets* *I* av storlek mindre än eller lika med *MAXLEN* som representeras. En anomalipoäng beräknas för varje observation *x*, omvänt proportionerlig till de *itemsets* som är infrekventa, dvs. vars support är mindre än *minsup* och därmed inte ingår i *FIS*:

$$Oteypoäng(x) = \sum_{I \in x, \text{sup}(I) \leq \text{minsup}} \frac{1}{|I|} \quad (3)$$

Implikationen av att ta storleken av infrekventa *itemsets* i beräkning är att de med mindre storlek ges högre inflytande över anomalipoängen, då *itemsets* av liten storlek antas vara mer karakteriserande för avvikare (Otey et al. 2006).

Då fokus, i likhet med Koufakou et al. (2007) är på kvalitativa variabler, bortses från den del av Oteys algoritm som genererar en kovariansmatris för kvantitativ data. Med denna tolkning blir algoritmens komplexitet linjär med avseende på antal observationer och exponentiell med avseende på antalet variabler, dvs.  $O(N \times M^M)$ .

---

<sup>6</sup> Eng: association rule mining

Input:	Datamaterial – $D$ Minimum support – $minsup$ Maxstorlek – $MAXLEN$
Output:	Vektor med anomalipoäng för varje observation – $Otey\_score$

FIS = med Apriori, generera alla *itemsets* (vars storlek  $\leq MAXLEN$ ) i  $D$  som förekommer med frekvens  $\geq minsup$

För varje observation  $x$  i  $D$ :

$I =$  med Apriori, generera alla *itemsets* (vars storlek  $\leq MAXLEN$ ) i  $x$

$Y =$  alla *itemsets* i  $I \notin FIS$ , dvs. förekomst  $< minsup$

För varje *itemset*  $y$  i  $Y$ :

$Otey\_score(x) += \frac{1}{|y|}$

Slut på loop

Slut på loop

Returnera  $Otey\_score$

Figur 1. Pseudokod för Oteys algoritmen.

### 3.2. Greedy

He et al. (2005; 2006) använder sig av konceptet entropi som verktyg för att utvinna avvikare ur datamaterial bestående av kvalitativa variabler. Man formulerar problemet som ett optimeringsproblem med målet att separera en delmängd med  $k$  observationer så att den förväntade entropin av återstående data minimeras. Entropi presenterades av Shannon (1948) som ett mått på informationsinnehåll och osäkerhet i en stokastisk variabel. Den totala entropin av ett datamaterial motsvarar den minsta möjliga komprimering som kan göras utan att information går förlorad. Lee et al. (2001) utvärderar olika entropimått för detektion av digitala bedrägerier. Man menar att entropin mäter regelbundenheten i datamaterialet, där datamaterial med låg entropi har färre olika observationer och större redundans. He et al. (2006) utvecklade sin ursprungliga Local Search Algorithm till den mer effektiva Greedyalgoritmen. Den baseras på antagandet att avvikare är observationer som om de exkluderas leder till lägre total entropi, dvs. mindre osäkerhet och oordning i återstående data.

Om  $X$  är en stokastisk variabel och  $S(X)$  den uppsättning värden som  $X$  kan anta och  $p(x)$  sannolikhetsfunktionen av  $X$ , så definieras entropin  $E(X)$  som:

$$E(X) = - \sum_{x \in S(X)} p(x) \log_2 p(x) \quad (4)$$

Givet ett multivariat datamaterial  $D$  med  $M$  variabler  $X_1, \dots, X_M$  och under antagandet att variablerna är oberoende, är entropin av  $D$  lika med summan av entropin för varje variabel:

$$E(D) = E(X_1) + \dots + E(X_M) \quad (5)$$

Optimeringsproblemet är att reducera datasetet med den delmängd  $O$ , bestående av  $k$  st observationer som minimerar entropin av  $D - O$ :

$$\min E(D - O) \quad \text{där } O \in D \text{ och } |O| = k \quad (6)$$

Då denna metod kräver ett förutbestämt antal sökta avvikare,  $k$ , och då algoritmen blir mycket beräkningstung på grund av en komplexitet på  $O(N \times M \times k)$  görs i denna undersökning en förenkling av algoritmen genom att iterera över alla observationer och beräkna den totala entropin som avsaknaden av varje observation innebär.  $O$  motsvarar alltså i varje iteration en observation  $x$ , dvs.  $k = 1$ :

$$\text{Greedy}_\text{poäng}(x) = E(D - x) \quad (7)$$

Ett lågt poäng för en observation innebär alltså att ordningen i datasetet minskas då observationen exkluderas, vilket kvalificerar den som en potentiell avvikare (He et al. 2006).

Input:	Datamaterial – $D$
Output:	Vektor med Greedy-poäng för varje observation – $\text{Greedy\_score}$
För varje observation $x$ i $D$ :	
	Sätt $O = x$
	Exkludera observationen, dvs. delmängden $O$ , från $D$
	$\text{Greedy\_score}(x) =$ Beräkna den totala entropin för återstående datamaterial, dvs. $E(D-O)$
	Slut på loop.
	Returnera $\text{Greedy\_score}$

Figur 2. Pseudokod för Greedyalgoritmen.

### 3.3. AVF

Koufakou et al. (2007) bidrog med den egna algoritmen AVF<sup>7</sup>, vilken man menade var mer effektiv för ändamålet. Då avvikare är observationer som är infrekventa i ett dataset menade man att en ideal avvikare i kvalitativ data är då varje variabelvärde är extremt infrekvent. Ett variabelvärdes infrekvens kan mätas genom att beräkna andelen förekomster av värdet i motsvarande variabel i datamaterialet. Ett mått för att avgöra om en observation är en avvikare definieras enligt:

$$\text{AVF-poäng}(x) = \sum_{l=1}^M f(x_{il}) \quad (8)$$

där  $f(x_{il})$  är andelen förekomster av värdet på variabel  $l$  i det undersökta datasetet. Ju lägre AVF-poäng, desto oftare antar variablerna ovanligt förekommande värden, och sannolikheten är större för att observationen är en avvikare. Komplexiteten är linjär med avseende på både antalet observationer och antalet variabler:  $O(N \times M)$ .

<sup>7</sup> Attribute Value Frequency



---

Input:	Datamaterial – $D$
Output:	Vektor med AVF-poäng för varje observation – $AVF\_score$

---

$Freq\_table$  = generera  $M$  st. tabeller med relativa frekvenser av värden för varje variabel

För varje observation  $x$  i  $D$ :

    För varje variabel  $l$  i  $x$ :

$AVF\_score(x) +=$  motsvarande relativa frekvens för värdet  $x_l$  i  $Freq\_table$

    Slut på loop

Slut på loop

Returnera  $AVF\_score$

---

Figur 3. Pseudokod AVF-algoritmen.

## 4. Experiment

Experimenten genomfördes på en dator med en 2,27 GHz Intelprocessor och 4GB RAM-minne. Alla algoritmer implementerades i den statistiska mjukvaran och programspråket R.<sup>8</sup> Förutom programvarans generiska funktioner användes metoden *apriori* från paketet *arules* och metoden *entropy* från paketet med samma namn. Dessa användes för att utvinna *frequent itemsets* från, respektive beräkna entropin av ett datamaterial.

### 4.1. Datamaterial

Det undersökta datamaterialet är resultatet från Väljarbarometern i november 2013. Väljarbarometern är en månatlig opinionsundersökning som United Minds genomför i samarbete med Aftonbladet. Huvudsyftet är att kartlägga partisympatier och förändringar i opinionen. I tillägg till dessa frågor undersöks även t.ex. vilka politiska frågor som anges som viktiga, mediekonsumtion samt bakgrundsvariabler som kön, ålder, sysselsättning och inkomst. Totalt antal registrerade svar var  $N = 1125$  respondenter. Frågor som endast vissa respondenter svarat på, dvs. de variabler som saknade minst ett värde, bortsågs från. Dessutom eliminerades variabler vars värden bestod av textsträngar inmatade av respondenterna vid frågor med öppna svar. Återstående antal frågor var  $M = 130$ , varav tre representerade kvantitativ data (svarstid, inkomst och ålder) och övriga kvalitativ, huvudsakligen binär data. Kvantitativa variabler diskretiserades uniformt i tio kategorier för att anpassa för algoritmer som behandlar uteslutande kvalitativ data. Se Tabell 1 för ett exempel med observationer och variabler i datasetet.

Tabell 1. Exempel från datasetet Väljarbarometern.

Observationer/Variabler	Q <sub>1</sub> : Kön	Q <sub>2</sub> : Ålder	...	Q <sub>M</sub> : Röstade i valet 2010
X <sub>1</sub>	Kvinna	20-29	...	Ja
X <sub>2</sub>	Man	10-19	...	Nej
...	...	...	...	...
X <sub>n</sub>	Kvinna	60-69	...	Ja

Till data adderades för testsyfte tio slumpmässigt genererade observationer. Varje värde slumpades inom spannet av de existerande värden som registrerats för respektive variabel i det ursprungliga datasetet. Dessa observationer motsvarar en simulering av helt slumpmässiga svar, och bör alltså flaggas som avvikare av fungerande algoritmer.

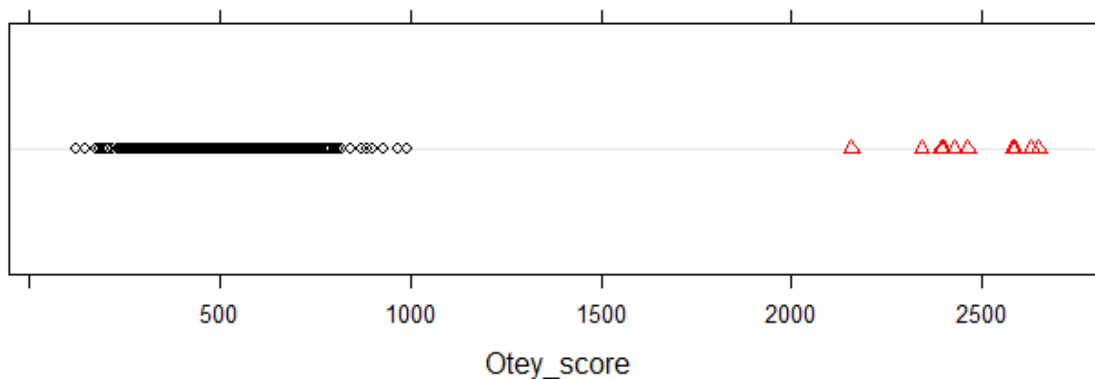
### 4.2. Resultat och diskussion

Inledningsvis redogörs för resultaten då slumpgenererade observationer adderats till data. Sedan avhandlas resultaten för det ursprungliga datasetet.

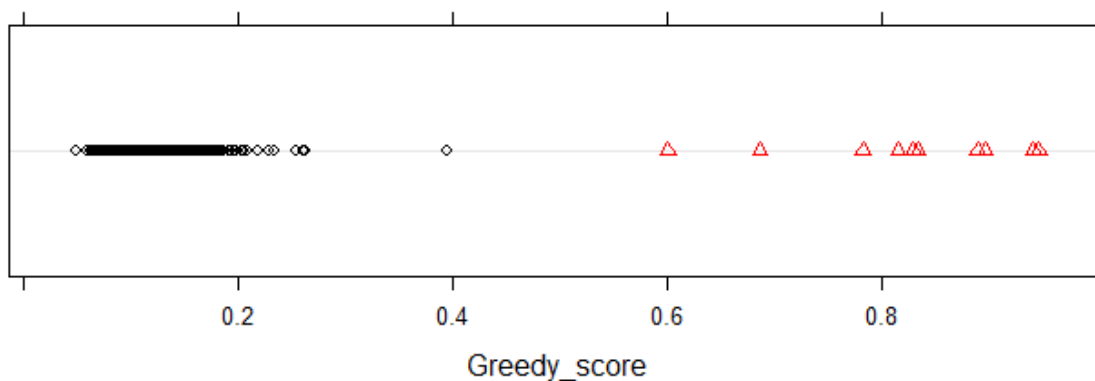
---

<sup>8</sup> <http://www.r-project.org/>

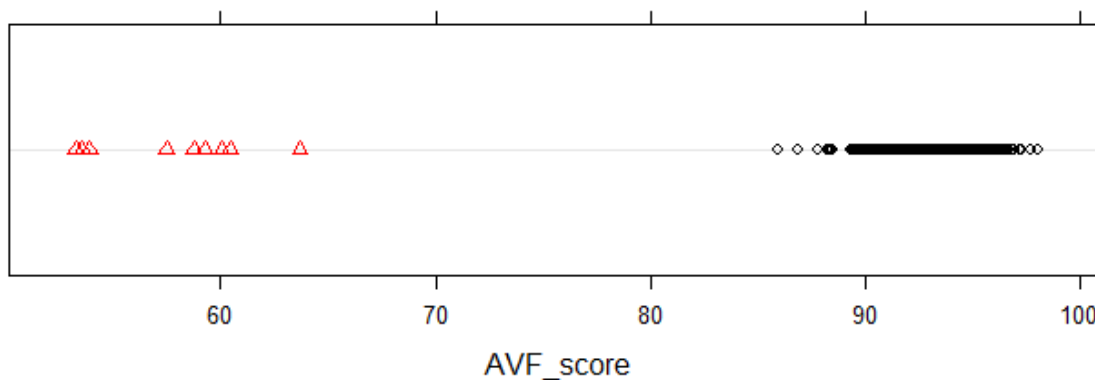
#### 4.2.1. Med slumpade observationer



Figur 4. Spridning av Oteypoäng ( $MAXLEN = 2$ ,  $minsup = 0,1$ ). Svarta ringar motsvarar originalobservationer, röda trianglar är slumpgenererade observationer adderade till originaldata.



Figur 5. Spridning av Greedy-poäng. Svarta ringar motsvarar originalobservationer, röda trianglar är slumpgenererade observationer adderade till originaldata.



Figur 6. Spridning av AVF-poäng. Svarta ringar motsvarar originalobservationer, röda trianglar är slumpgenererade observationer adderade till originaldata.

Samtliga algoritmer tilldelar de slumpgenererade observationerna utmärkande anomalipoäng. Otey och AVF har något tydligare avgränsning mellan originaldata och adderade observationer.

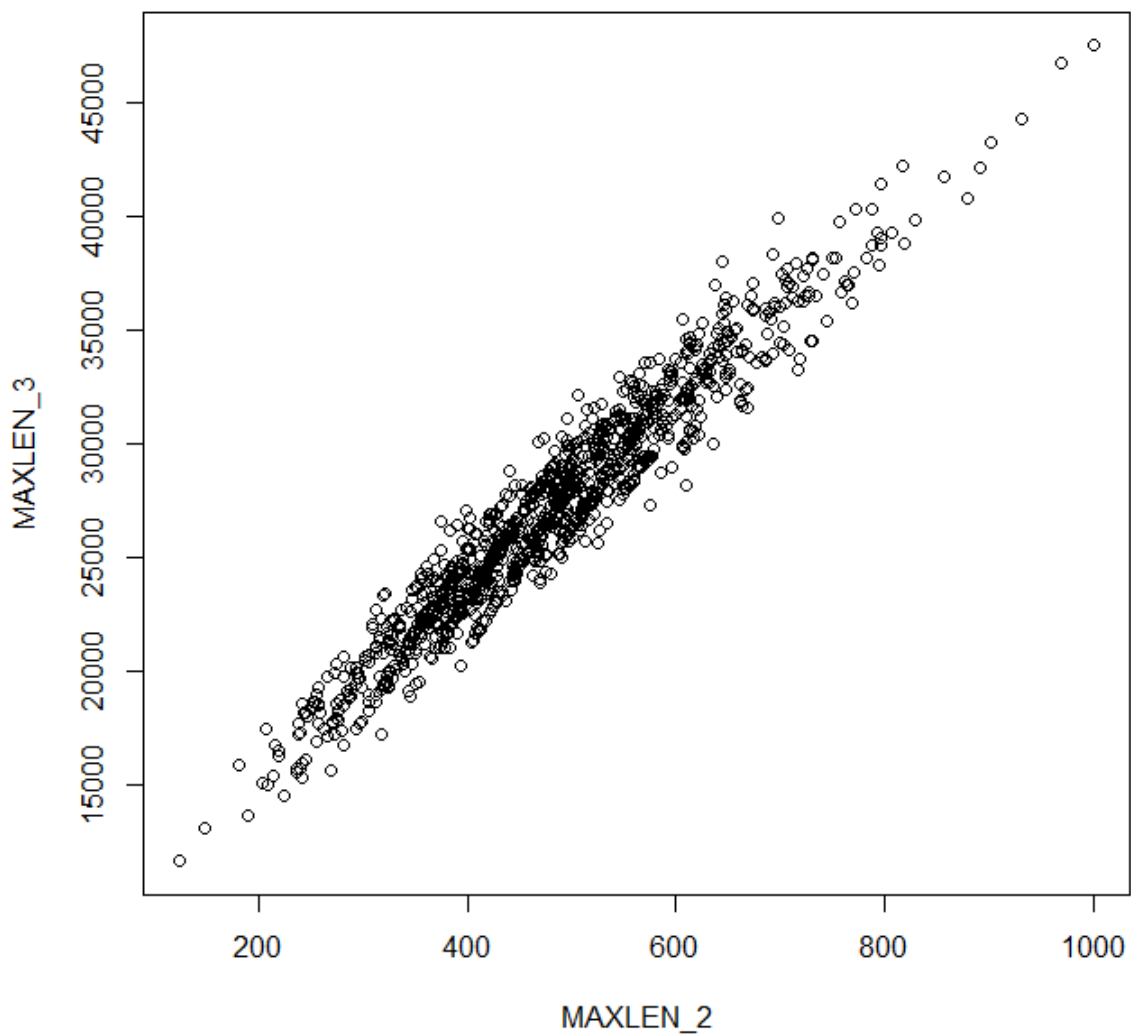
Algoritmerna tycks alltså effektivt kunna särskilja helt slumpmässigt framställda svar, vilket kan ses som ett grundläggande krav för en effektiv detektion av avvikare i surveydata. Originalobservationerna befinner sig inte nära artificiellt framställda observationer vilket föreslår att ingen av dem registrerats slumpartat. Finns misstankar om att ett datamaterial innehåller sådana datapunkter skulle testen ovan kunna vara effektivt i att identifiera dessa.

#### 4.2.2. Otey

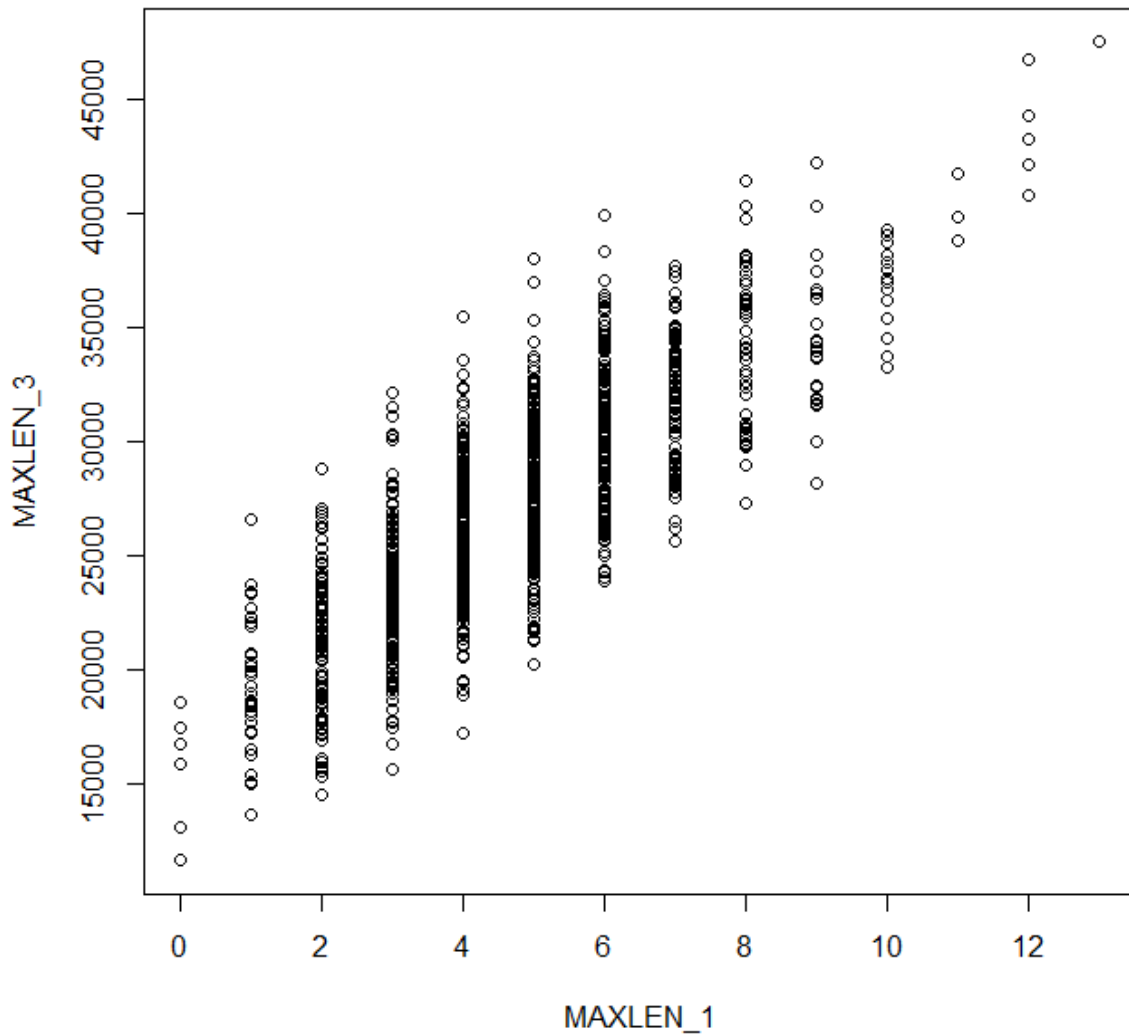
Parametrarna *minsup* respektive *MAXLEN* sattes inledningsvis till 0,1 respektive 2. Betydelsen av detta är att när Apriorialgoritmen gör den inledande utvinningen av *frequent itemsets* i datasetet så inkluderas de *itemsets* som förekommer i minst en tiondel av observationerna. De *itemsets* som utvinns har maxstorlek två. Antalet *frequent itemsets* som utvinns är då  $|FIS| = 19\,429$ . Tidsåtgången för algoritmen, dvs. beräkning av Oteypoäng för samtliga observationer är 128 sekunder.

Teoretiskt sett så förbättras algoritmens precision av att utvinna *itemsets* av större maxstorlek. Används istället parametern  $MAXLEN = 3$ , allt annat lika, tar det dryga timmen för algoritmen att processa datamaterialet. Detta då den inledande utvinningen resulterar i  $|FIS| = 909\,674$  stycken *frequent itemsets*.

En jämförelse mellan resultaten av de olika maxlängderna visar hur resultaten med de olika maxlängderna skiljer sig. I Figur 7 plottas anomalipoäng när parametern  $MAXLEN = 2$  mot motsvarande poäng då  $MAXLEN = 3$ . I Figur 8 illustreras anomalipoäng när  $MAXLEN = 1$  jämfört med när  $MAXLEN = 3$ .



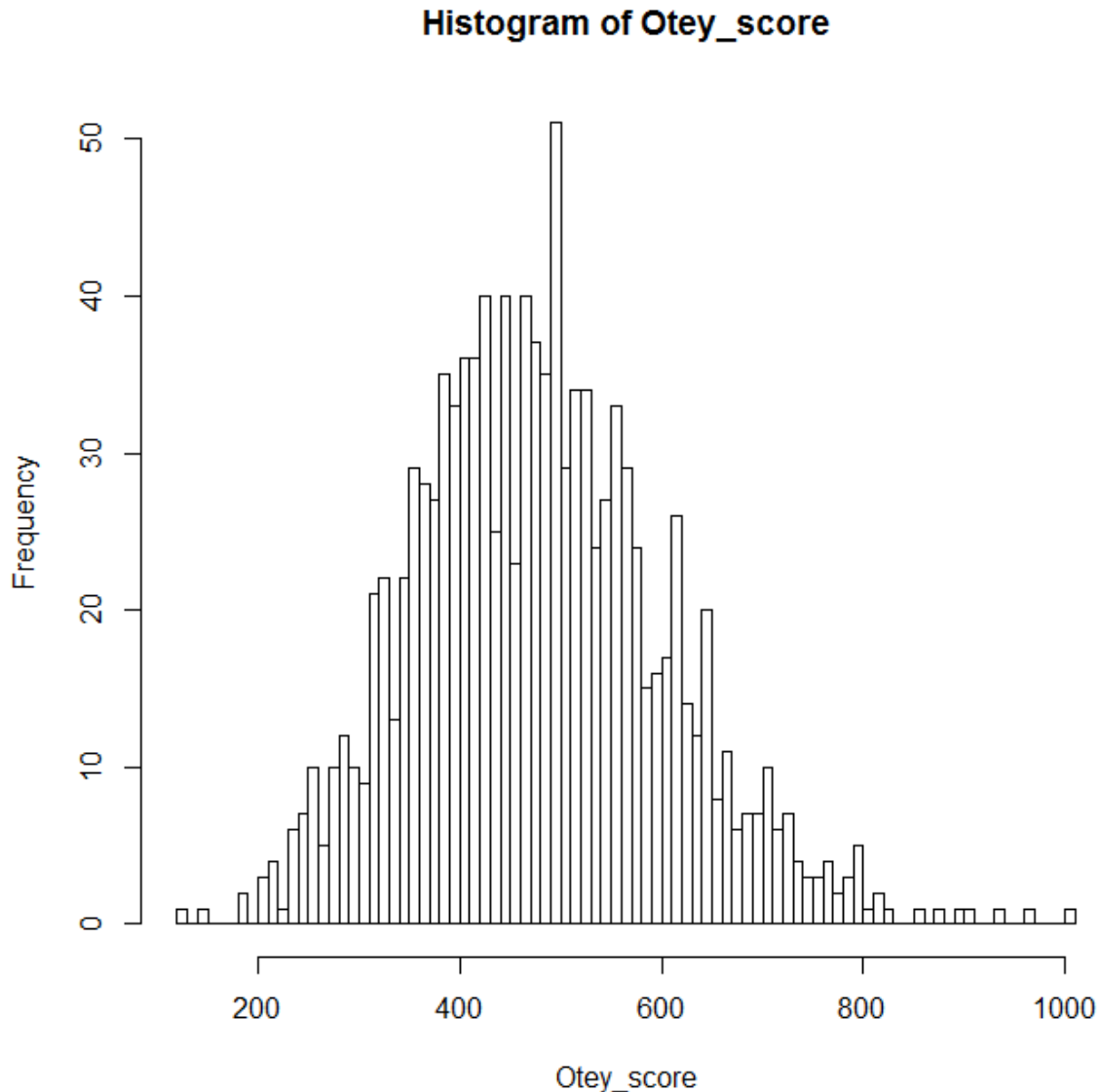
*Figur 7. I y-led Oteypoäng för MAXLEN=3. I x-led Oteypoäng vid MAXLEN=2. Pearsons korrelationskoefficient 0.97.*



Figur 8. I y-led Oteypoäng för  $MAXLEN = 3$ . I x-led Oteypoäng vid  $MAXLEN = 1$ . Pearsons korrelationskoefficient 0.82.

Vi ser i Figur 7 att punkternas spridning följer ett relativt tydligt linjärt mönster, dvs. att en hög anomalipoäng då algoritmen är inställd på  $MAXLEN = 2$  för samma observation även innebär en hög poäng då  $MAXLEN = 3$ . Den mer tidseffektiva inställningen som utvinner färre *itemssets* tycks alltså vara tillräcklig för ändamålet. Är maxlängden inställd till ett är spridningen dock betydande (Figur 8), vilket antyder att utvinning av avvikare med denna inställning är känslig för felbedömningar. Med inställningen  $MAXLEN = 3$  som referensvärde manifesteras även kvalitetsskillnaden av Pearsons korrelationskoefficient då korrelationen med resultatet för  $MAXLEN = 2$  är nästan perfekt positiv (0.97), medan motsvarande för  $MAXLEN = 1$  endast är 0.82.

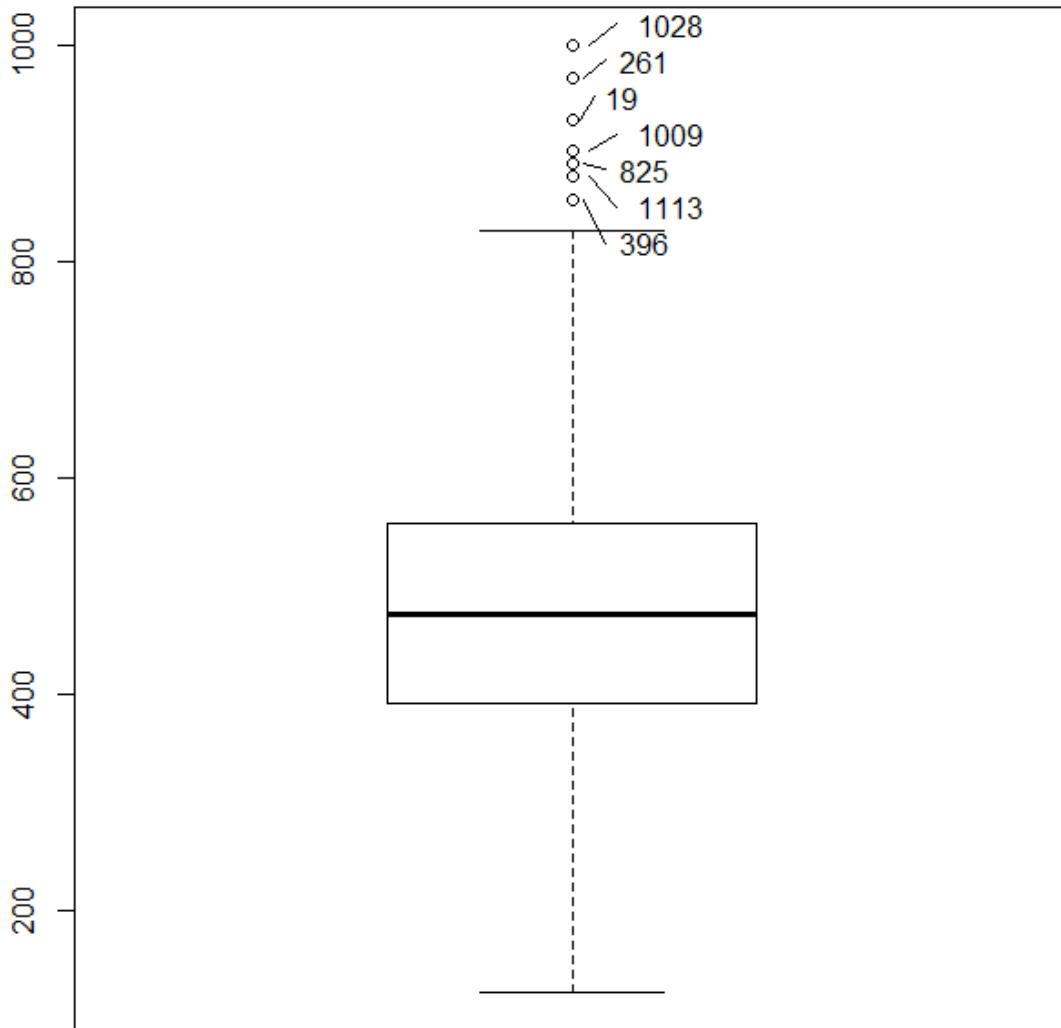
Inställningen  $MAXLEN = 2$  tycks alltså vara optimal med avseende på kvalitet och effektivitet och denna används vid presentation och tolkning av resultat. Figur 9 illustrerar resulterande Oteypoäng med dessa parametrar, uttryckt i ett histogram.



Figur 9. Histogram över Oteypoäng ( $MAXLEN = 2$ ,  $minsup = 0,1$ ).

Histogrammet i Figur 9 visar en spridning som påminner om formen för en normalfördelning. Nollhypotesen att observationernas spridning är normalfördelade förkastas dock enligt Shapiro Wilks test, med  $p$ -värde  $= 4,1 \times 10^{-7}$  (Shapiro et al. 1965). Vi ser dock att spridningen är unimodal, dvs. har en tydlig topp. Vid en analys av spridningen, med fokus på dess extremvärden, ser vi att den vänstra svansen är kortare än den högra. Det är till höger, för höga anomalipoäng, eventuella avvikare borde befinna sig.

En detaljerad analys av observationerna med ovanligt höga anomalipoäng visas i Figur 10, där ett lådagram plottats av anomalipoängen. De observationer som befinner sig bortanför avskärningsvärdena från Chebyshevs olikhet ( $\mu \pm 3\sigma$ ) märks ut med motsvarande index i datamaterialet.



Figur 10. Lådagram över Oteypoäng ( $MAXLEN = 2$ ,  $minsup = 0,1$ ) med markerade index för avvikare.

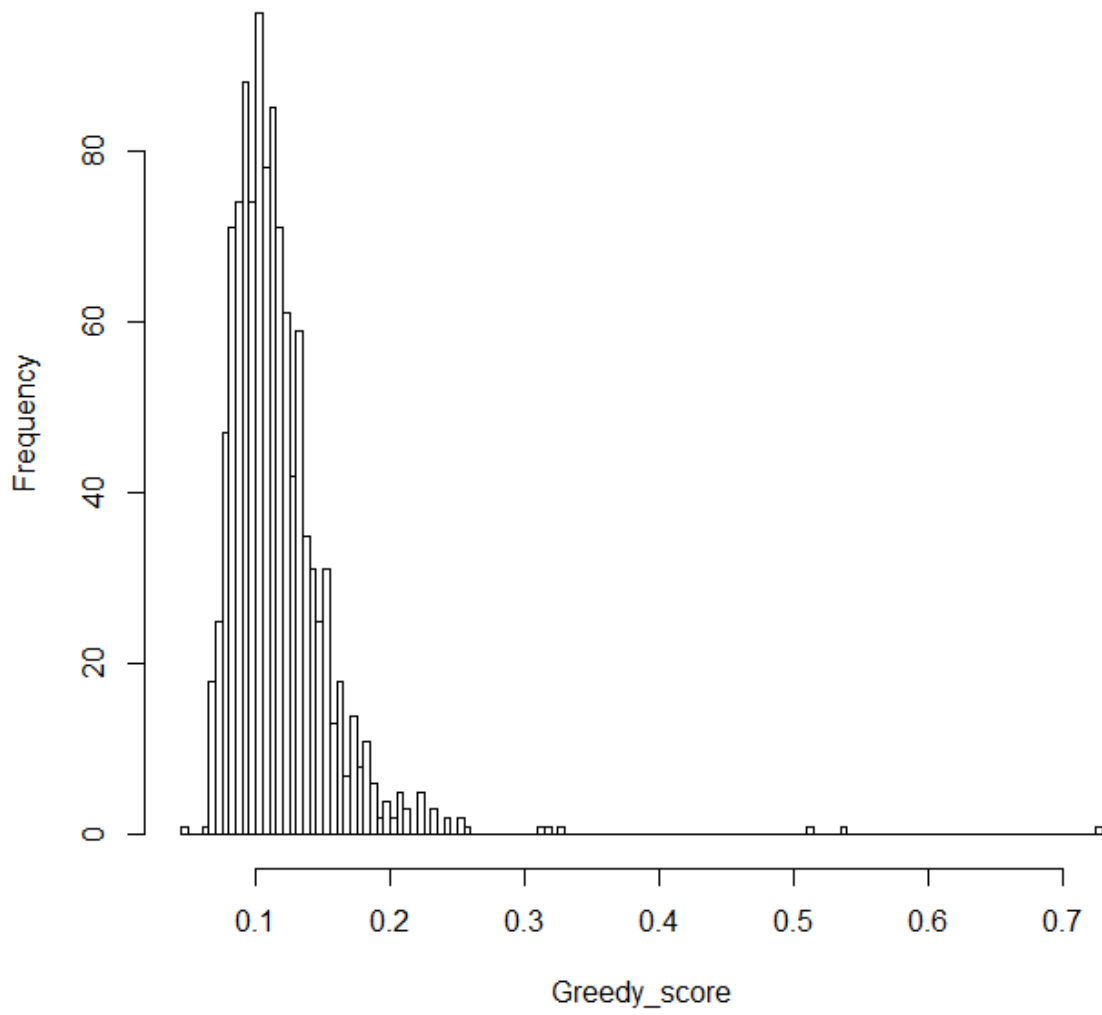
Om avskärningen mellan vad som anses som normala värden och avvikare sker som i Figur 10 utmärker sig alltså sju observationer med ovanligt höga anomalipoäng som extra intressanta att studera i detalj.

#### 4.2.2. Greedy

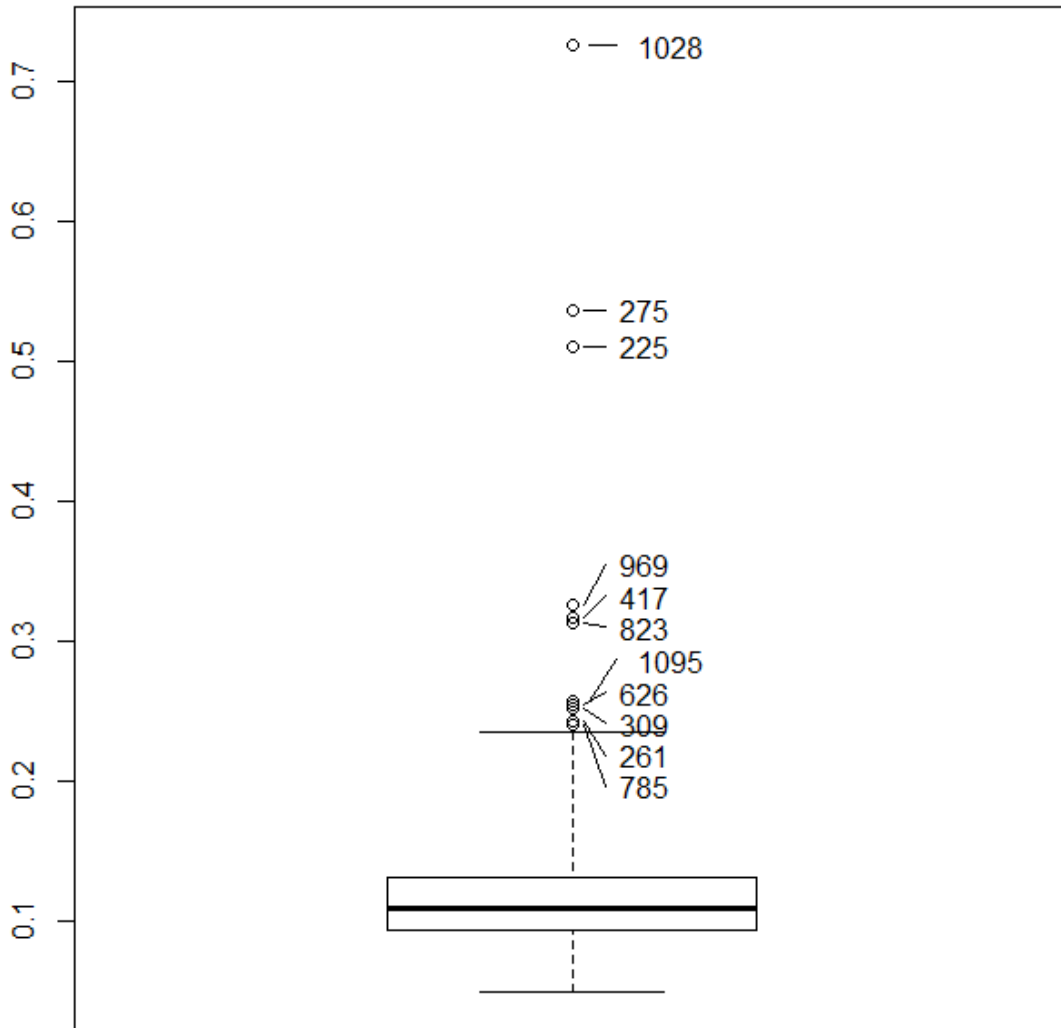
Vid analys av de anomalipoäng som genereras av Greedyalgoritmen framstår en skev fördelning över olika frekvenser (Figur 11). Även denna spridning är tydligt monomodal. Den högra svansen är betydande, och några observationer är tydligt skilda från övriga. Betydelsen av Greedyoängen är ju den resulterande minskning i entropi av datasetet då observationen bortses från. Normalvärden tycks orsaka endast en minskning i storleksordningen  $0,05 - 0,25$ , medan algoritmen urskiljer ett antal observationer med Greedyoäng högre än  $0,3$ .



**Histogram of Greedy\_score**



Figur 11. Histogram över Greedyöäng.



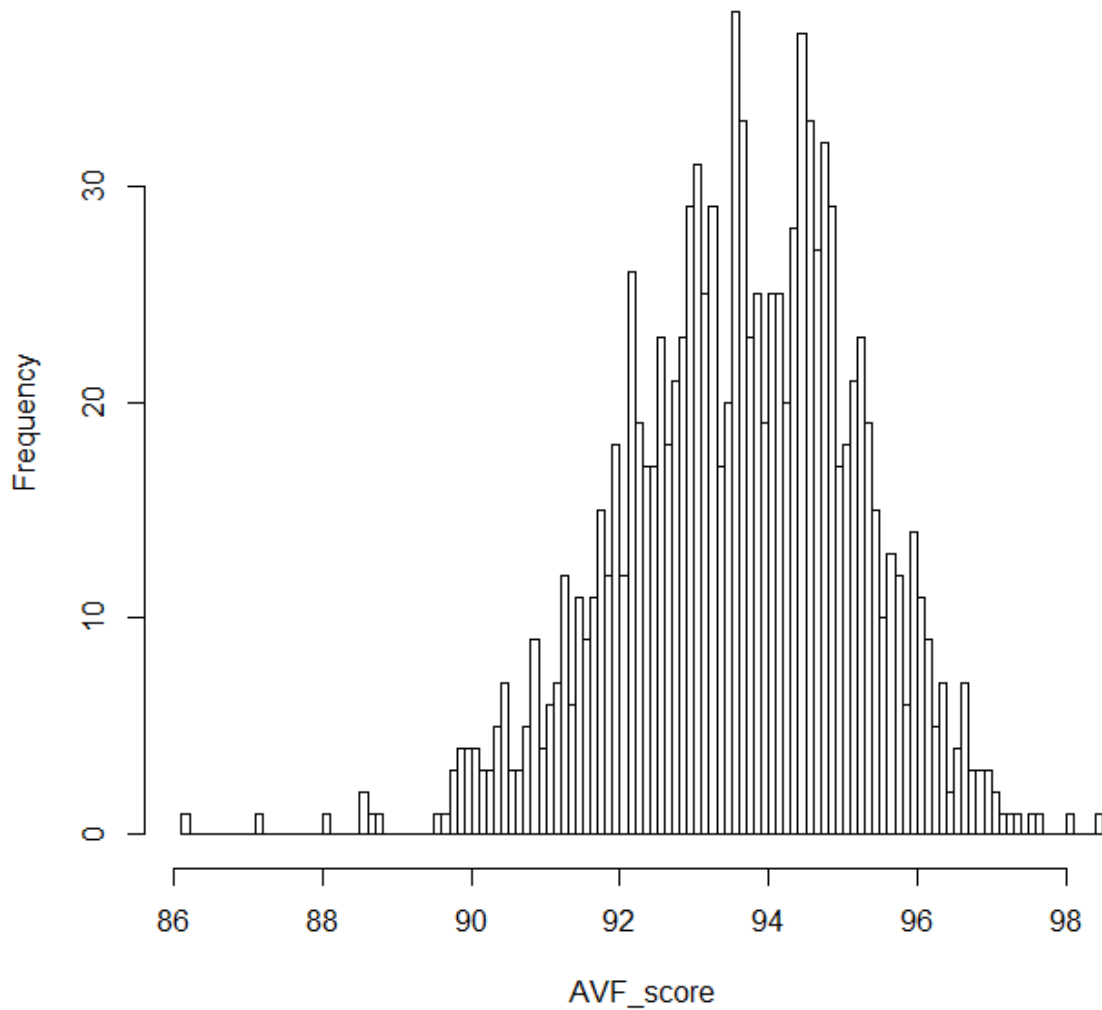
Figur 12. Lådagram över Greedy poäng med markerade index för avvikare.

Enligt Chebyshevs olikhet identifieras elva stycken observationer som avvikare. Vi ser dock att fyra av dessa befinner sig mycket nära normala observationer. Möjligen skulle Figur 12 föreslå att avskärningsvärdet istället bestämdes till 0,3 och därmed räkna även dessa som normala observationer. I denna undersökning tillämpas dock konsekvent Chebyshev för tolkning av resultat.

#### 4.2.3. AVF

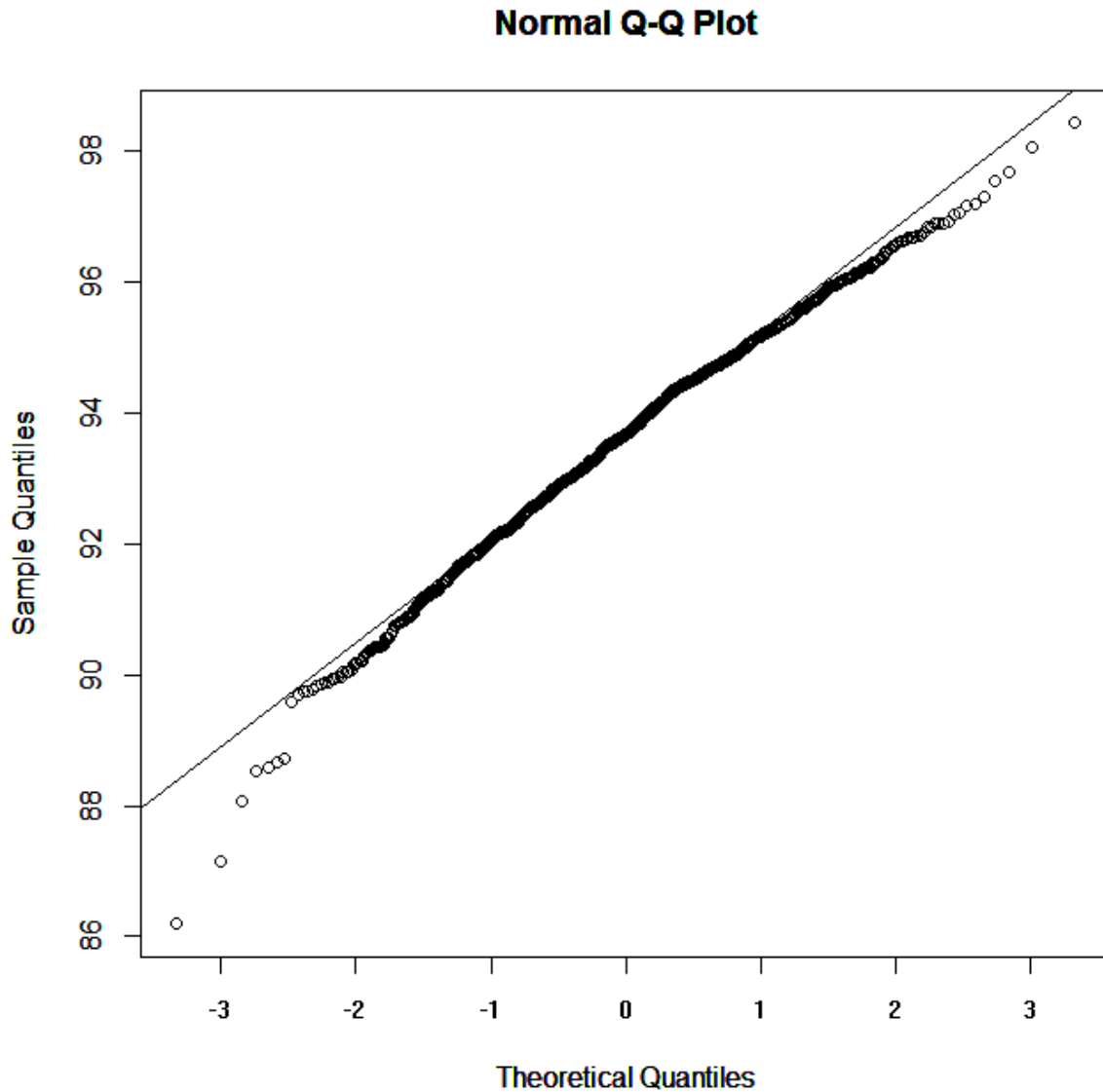
Även resultaten från AVF-algoritmen har en fördelning som liknar en normalfördelning, men med en betydande svans för låga värden, dvs. värden som för algoritmen tyder på att observationerna är avvikare (se Figur 13). Nollhypotesen om en normalfördelning förkastas dock av Shapiro Wilks test ( $p - värde = 3,8 \times 10^{-6}$ ).

Histogram of AVF\_score



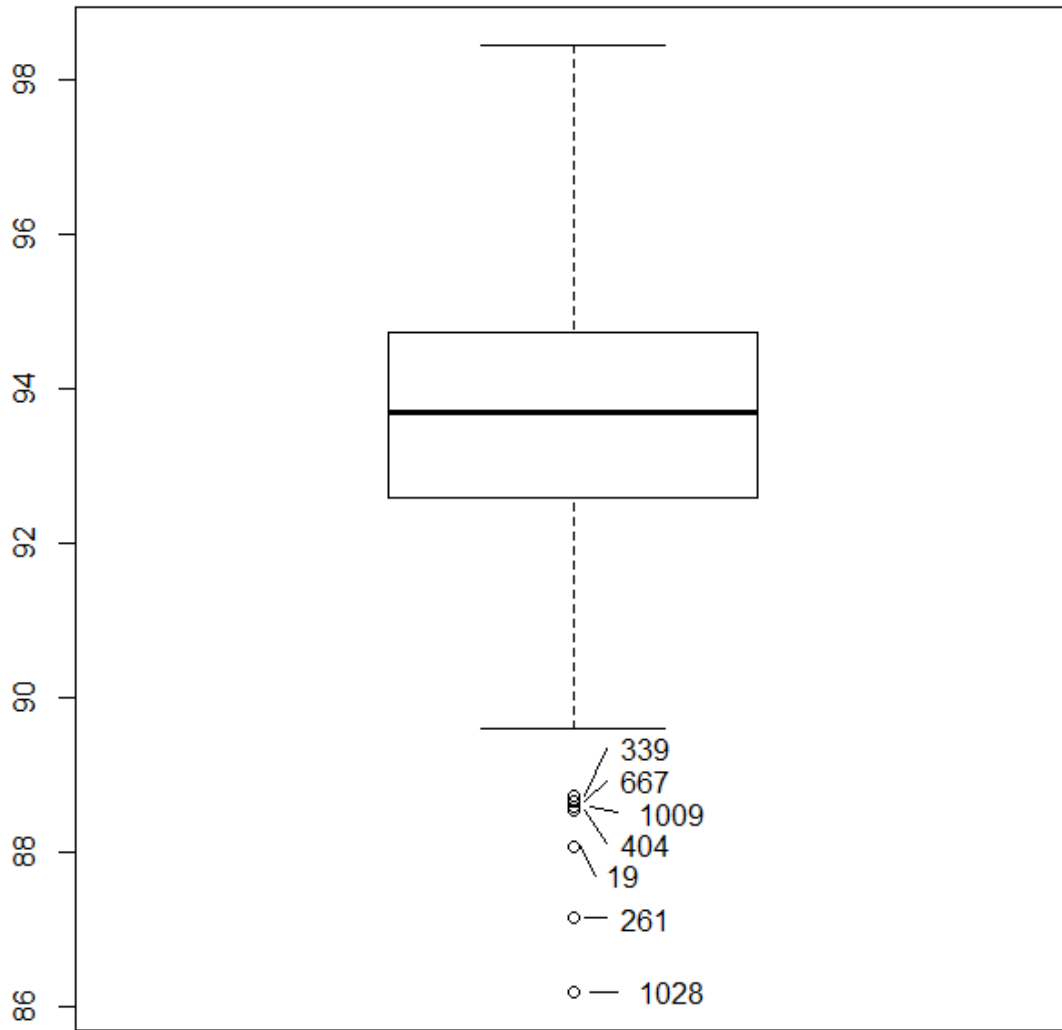
Figur 13. Histogram över AVF-poäng.

I Figur 14 plottas AVF-poängen i en Q-Q-graf med normalfördelningens förväntade spridning vilket ändå visar att spridningen någorlunda väl följer normalfördelningen, undantaget extremvärdena som avviker. För låga värden, där vi förväntar våra avvikare, utmärker sig sju observationer genom betydande avvikelser från en förväntad normalfördelning.



*Figur 14. Q-Q-graf över AVF-poäng samt normalfördelningens referenslinje.*

Samma observationer identifieras som tydliga avvikare av ett lådagram över AVF-poängen då Chebyshevs olikhet tillämpas ( $\mu \pm 3\sigma$ ). Trots att spridningen i Figur 13 antyder en fördelning med två toppar antas för enkelhets skull en unimodal fördelning och att Chebyshevs olikhet beräknas med samma parametrar som tidigare.



**Figur 15.** Lådagram över AVF-poäng med markerade index för avvikare.

#### 4.2.4. Jämförelse

Givet de använda parametrarna och med Chebyshevs olikhet ( $\mu \pm 3\sigma$ ) tillämpad för avskärningsvärden genererar Oteyalgoritmen och AVF vardera sju avvikare, medan Greedy genererar elva stycken.

Tabell 2. Index för avvikare rangordnade efter anomalipoäng. Observationer flaggade som möjliga avvikare av samtliga algoritmer är märkta (\*\*). De som flaggas av två av algoritmerna är märkta (\*).

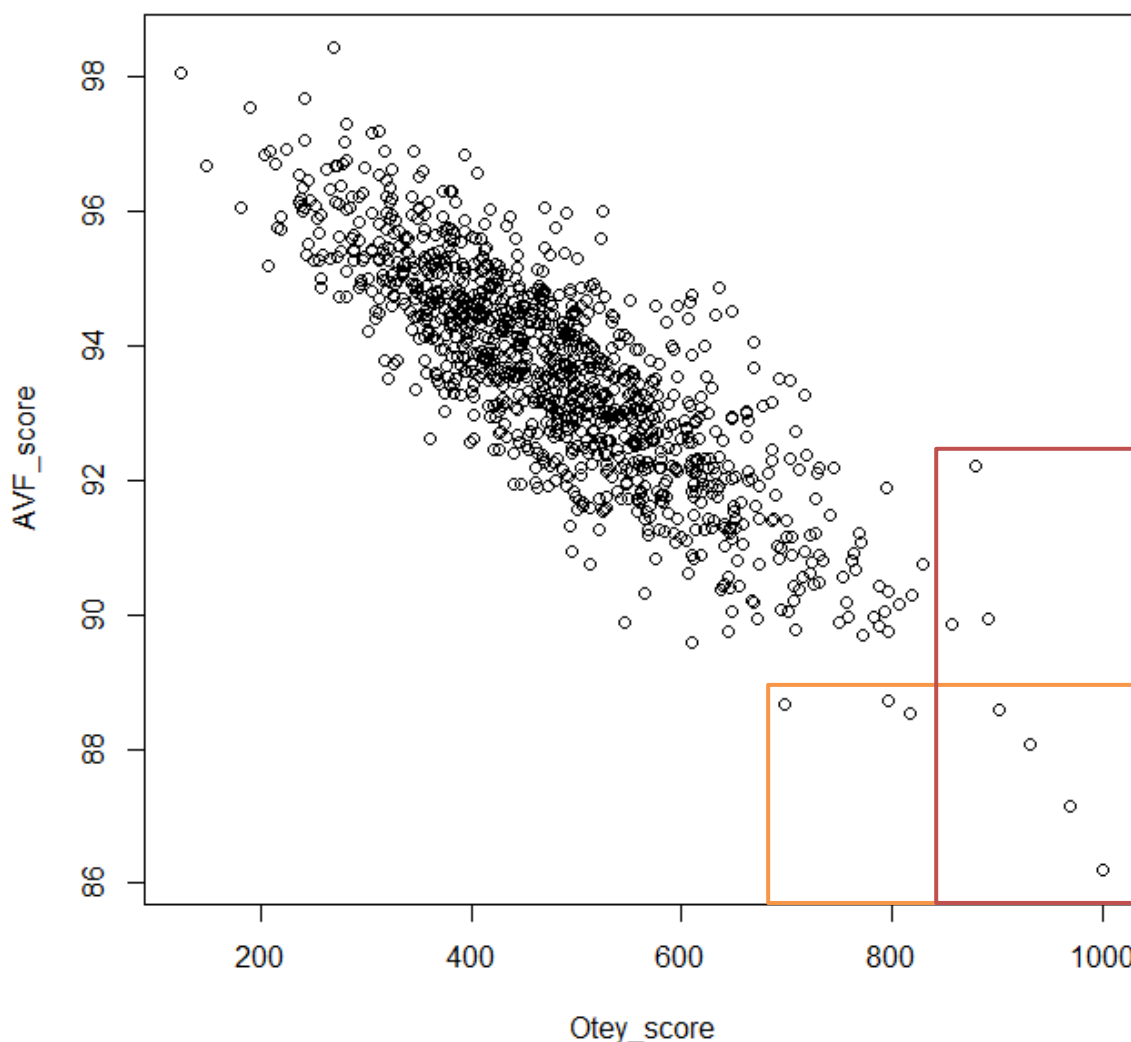
	Otey <sup>9</sup>	Greedy	AVF
1.	1028**	1028**	1028**
2.	261**	275	261**
3.	19*	225	19*
4.	1009*	969	404
5.	825	417	1009*
6.	1113	823	667
7.	396	1095	339
8.	-	626	-
9.	-	309	-
10.	-	261**	-
11.	-	785	-

Med avseende på algoritmernas uppbyggnad finns skillnader som borde påverka deras användbarhet av att hitta avvikare i data från frågeundersökningar. AVF är den minst sofistikerade av algoritmerna då den behandlar varje variabel enskilt och beräknar en anomalipoäng beroende på hur vanligt förekommande värdena är. Det finns alltså ingen egenskap hos AVF-algoritmen som möjliggör korrelation mellan variabler. Algoritmen borde effektivt kunna identifiera observationer som avviker från det normala genom att anta upprepade extrema värden för flertalet variabler. Däremot kan den inte förväntas prestera väl då det egentligen avvikande med en observation endast torde kunna identifieras genom att jämföra flera av dess svar. Exempel på kombinationer av troligtvis ovanliga svarsmönster som AVF inte skulle lyckas identifiera är om en och samma respondent svarat {ålder = 18; använder aldrig internet} eller {röstade på Vänsterpartiet i valet 2010; skulle rösta på Moderaterna om det var val idag}. Denna funktionalitet stöds av Oteyalgoritmen (för  $MAXLEN > 1$ ). Egentligen bygger Otey på samma grundantagande som AVF, att avvikare är observationer med upprepade ovanliga svar registrerade. Egenskapen att även ta korrelation med i beräkningen gör den däremot mer sofistikerad, en egenskap som tros vara betydelsefull för denna tillämpning av detektion av avvikare.

Korrelationen mellan resultaten av de båda algoritmerna är förhållandevis hög, vilket illustreras i Figur 16. Pearsons korrelationskoefficient är  $-0,83$  (negativt värde då AVF

<sup>9</sup> Parametrarna  $MAXLEN = 2$ ,  $minsup = 0,1$ .

genererar låga värden på avvikare och Otey höga), dvs. förhållandevis nära en perfekt negativ korrelation. Framförallt ser vi att korrelationen är relativt hög för avvikande observationer, vilka är intressanta för undersökningen. I figuren innesluts de avvikare som identifierats av respektive algoritm med varsin rektangel. De observationer som innesluts av båda rektanglarna är alltså identiska med av de båda algoritmerna delade avvikare i Tabell 2.



Figur 16. Oteypoäng (x-axeln) plottat mot AVF-poäng (y-axeln) för observationerna i datamaterialet. Den orangea och röda rektangeln innesluter Oteys respektive AVF-algorithmens avvikare.

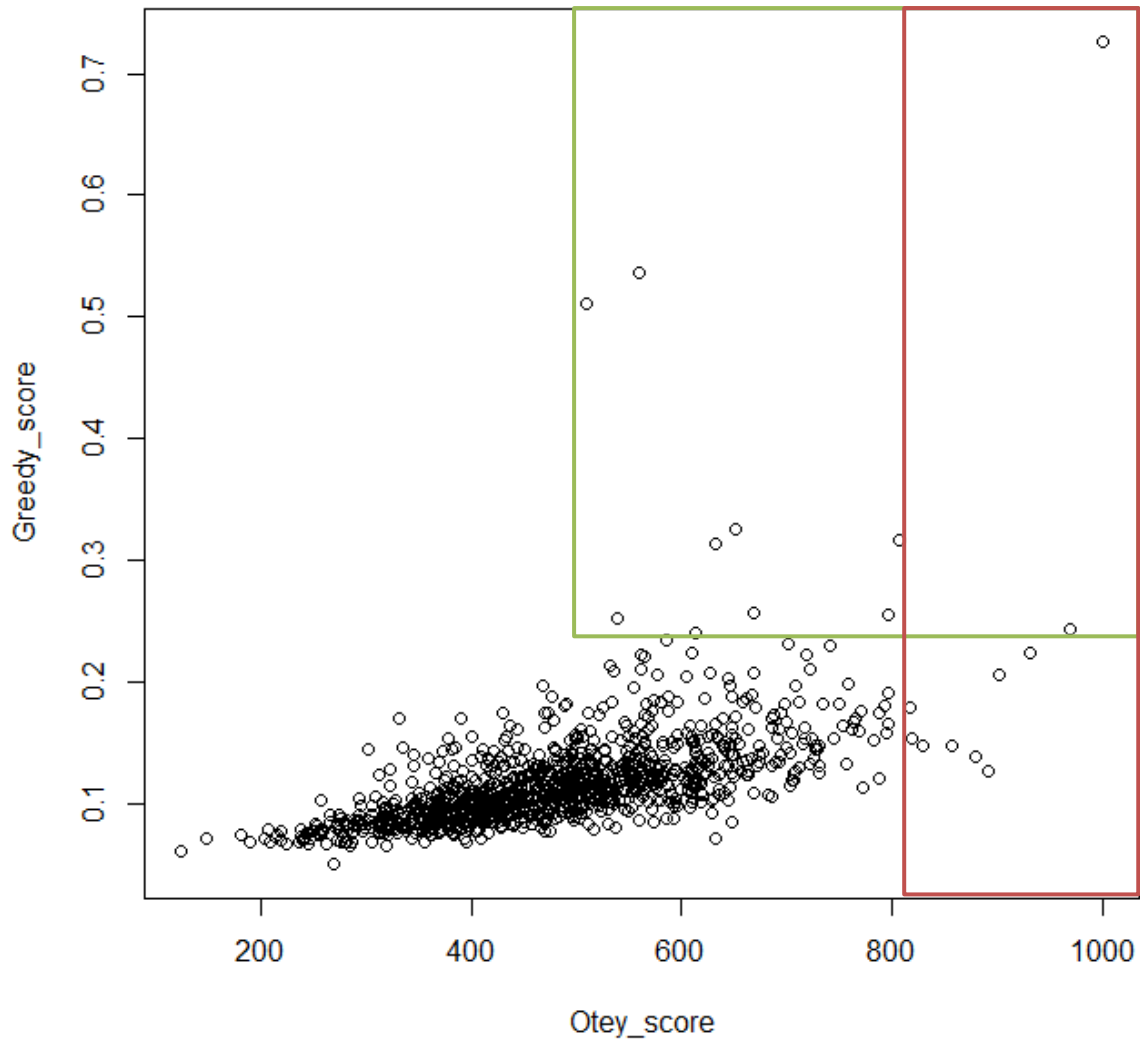
AVF och Otey genererar alltså resultat som är förhållandevis lika, men Otey torde vara den mer sofistikerade av de två enligt tidigare resonemang. AVF:s jämförelsevis korta tidsåtgång (se Tabell 3) och låga komplexitet skulle vara fördelaktig om en större skalbarhet efterfrågades, men anses i den nuvarande tillämpningen inte väga upp bristen på korrelation mellan variablerna, varför Otey antas överlägsen den förra.

Greedyalgoritmen bygger på begreppet entropi och dess informationsteoretiska tillämpning. Den mäter graden av oordning eller heterogenitet i datamaterialet, eller mer specifikt hur mycket mer homogent datamaterialet blir genom att bortse från potentiella avvikare. Huruvida entropikonceptet är tillämpligt på resultat från frågeundersökningar

går att ifrågasättas. Det är en svår balansgång mellan att detektera avvikare som genererats via något *satisficing*-fenomen, men undvika att misstänkliggöra respondenter med riktiga men udda svarsmonster, dvs. distinktionen mellan representativa och icke-representativa avvikare. Entropimåttet mäter graden av regelbundenhet i ett datamaterial, och risken är att det är ett för trubbigt instrument för att göra denna distinktion. Funktionen för beräkning av entropi liknar till viss del AVF i att den summerar över sannolikheten för olika utfall, men där Greedy även multiplicerar med logaritmen. Precis som AVF saknas egenskapen att mäta korrelation mellan olika variabler. Även jämfört med Greedyalgoritmen tycks alltså Otey vara mer sofistikerad. Vi såg i Tabell 3 att få observationer flaggades av båda dessa algoritmer och det är intressant att jämföra resultaten mer detaljerat.

Pearsons korrelation mellan anomalipoängen för Greedy och Otey är 0,59 vilket är förhållandevis långt från en perfekt positiv korrelation. Den långt från linjära spridningen illustreras i Figur 17 tillsammans med rektanglar med föreslagna avvikarkandidater. Två överlappande observationer identifieras (se även Tabell 2). Att flera av de enligt Greedyalgoritmen flaggade observationerna har normala Oteypoäng och vice versa är ytterligare tecken på att de mäter olika egenskaper bland observationerna i datamaterialet. Det är intressant att notera att korrelationen är högre för låga anomalipoäng för de respektive algoritmerna (nedre vänstra hörnet i Figur 17). Det tycks alltså vara så att måtten överensstämmer med avseende på de observationer som är ”ovanligt normala”. Med en minskad normalitet börjar sedan resultaten divergera och korrelationen minska.





Figur 17. Oteypoäng (x-axeln) plottat mot Greedypoäng (y-axeln) för observationerna i datamaterialet. Den röda och gröna rektangeln innesluter Oteys respektive Greedys avvikare.

Tabell 3. Tidsåtgång för respektive algoritm.

Algoritm	Tid (min)
Otey 2 <sup>10</sup>	2,1
Greedy	3,5
AVF	0,9
Otey 3 <sup>11</sup>	63
Otey 1 <sup>12</sup>	1,6

<sup>10</sup> Parametrarna  $MAXLEN = 2, minsup = 0,1$ .

<sup>11</sup> Parametrarna  $MAXLEN = 3, minsup = 0,1$ .

<sup>12</sup> Parametrarna  $MAXLEN = 1, minsup = 0,1$ .

#### 4.2.5. Representativitet

Återstår gör frågan huruvida de avvikare som produceras av den förmodat mest sofistikerade Oteyalgoritmen verkligen är icke-representativa för populationen, och alltså genererats via någon typ av *satisficing*. Tyvärr saknas den information som krävs för att göra denna distinktion. Datamaterialet är till sin karaktär oövervakat, dvs. det finns inget facit som avslöjar vilka observationer som är icke-representativa på grund av någon felaktighet. Denna uppenbara brist anses nödvändig göra ett fastställt tröskelvärde, vilket här etablerats med Chebyshevs olikhet, där de observationer som befinner sig bortom detta värde tolkas som icke-representativa. En fördjupad undersökning av algoritmens kvalitet genom detaljstudier av avvikande observationer krävs för att validera denna tolkning.

#### 4.2.6. Hantering

För att följa på Barnetts karakterisering av olika metoder för hantering av avvikare bortses från *införlivande* av observationerna då detta skulle innebära att behandla datamaterialet som det görs i nuläget, dvs. ingen hantering. Många robusta metoder för *anpassning* är inte användbara på grund av att datamaterialet är multivariat samt ej av kvantitativ typ. Exempel på detta är medianmättet och Winsorisering. Som anpassning skulle dock även nedviktnig kunna kategoriseras, vilket är en metod som skulle kunna användas i detta fall. En *identifiering* av avvikare är ju en del av syftet med undersökningen, men under antagandet att avvikare är icke-representativa bör det vara fruktlöst att med hjälp av dessa förklara någon egenskap av populationen varför denna metod avfärdas. Till sist återstår *eliminering*, vilket vore optimalt om fastställt med absolut säkerhet att alla detekterade avvikare är icke-representativa observationer. Som tidigare nämnt är detta dock inte fallet, och försiktighet bör iaktas i att tillämpa denna metod.

Är antagandet att svar angetts oavsiktligt, genom t.ex. diskriminantanalys avslöjade av att endast ett eller ett fåtal värden bidrar till observationens avvikelse från det normala, skulle en uppföljning av respondenten vara rimlig. Detta analogt med de metoder för undersökningar av företagsdata föreslagna av t.ex. Battipaglia (2002). En avvägning måste dock göras om inflytandet på totalresultatet väger upp för kostnaden i tid och pengar för denna uppföljning, samt om en uppföljning är tekniskt möjlig.

Avsiktligt angivna felaktigheter torde vara mindre meningsfulla att följas upp då respondenter fortsatt kan vara benägna att ange felaktiga svar. Dessa tros kunna identifieras genom att flera eller alla värden bidrar till observationens avvikelse. Med avseende på tidigare resonemang om representativitet, och med vetskapen om svårigheten med att konstruera en algoritm som perfekt hanterar denna nyansskillnad, föreslås en nedviktnig av dessa observationer, omvänt proportionerlig mot dess anomalipoäng. Detta i enlighet med bl.a. Eltinge et al. (2006) och Barge et al. (2012). Risken finns att även representativa avvikare då nedviktas, men denna risk anses underordnad syftet att minska inflytandet från icke-representativa observationer.

## 5. Slutsatser

Definitionenligt är avvikare de observationer som avviker mer än ett visst tröskelvärde från det övriga datamaterialet. För multivariat data är möjligt att beräkna ett sammanvägt mått på anomali vilket ligger till grund för denna jämförelse. I frågeundersökningar är icke-representativa avvikare de observationer som helt eller delvis genererats via någon typ av *satisficing*. Respondenten har inte gjort en noggrann kognitiv ansträngning utan nöjt sig med suboptimala svar. Dessa bör särskiljas från representativa avvikare som är korrekta återgivning av respondentens egentliga, men från det normala avvikande, beteende.

Olika metoder för detektion av avvikare existerar, mer eller mindre användbara beroende på datamaterialets karaktär. Applicerat på datamaterial från frågeundersökningar liknande den studerade används företrädesvis metoder som hanterar kvalitativa variabler, där eventuella kvantitativa variabler diskretiseras. Även metoder som hanterar en kombination av variabeltyper är självklart användbara. Att genomföra multivariata sammanvägningar av datamaterialet har visats användbart i flertalet applikationer och vissa angreppssätt anses vara överförbara även på frågeundersökningar.

Tre algoritmer anses ha de rätta förutsättningarna för applikation på datamaterialet. Samtliga urskiljer korrekt helt slumpmässigt genererade observationer som avvikare, och torde alltså vara effektiva verktyg för att identifiera dessa typer av anomalier. Användningen av någon typ av tumregel, i detta fall Chebyshevs olikhet, anses nödvändigt för att kunna producera användbara resultat. För originaldata identifieras då en handfull observationer som avvikare av vardera algoritmen. Ett fåtal av dessa är gemensamma dem emellan, medan andra flaggas som avvikande av en algoritmen men som normal av en annan. Med avseende på algoritmernas uppbyggnad och effektivitet anses Oteyalgoritmen, inställd att hantera *itemsets* med två element vara överlägsen de övriga. Den främsta anledningen är att den tar hänsyn till korrelationer variabler emellan, vilket anses vara en viktig egenskap för att fastställa ett mått på anomali för denna typ av data. Teoretiskt sett förbättras kvaliteten om även *itemsets* med fler element studeras, men då det i praktiken gör lite skillnad för respektive observations relativa anomalipoäng men leder till en avsevärt högre tidsåtgång, anses detta varken nödvändigt eller eftersträvansvärt.

I den mån uppföljning av svar är tekniskt möjligt och till en låg kostnad bör detta göras för att försöka förklara oavsiktliga icke-representativa svar. Är detta ej genomförbart, och för andra avvikande observationer som antas ha genererats avsiktligt, föreslås en nedviktning av observationerna, omvänt proportionerligt mot dess anomalipoäng. På detta sätt minskas inflytandet från dessa observationer vilket gör att totalresultatet mer rättvist beskriver verkligheten. Verktøget kan även användas för att validera ett datamaterial med fåtalet eller inga avvikare. Ytterligare ett område för tillämpning av detta verktyg är för utvärdering av kvaliteten hos enskilda webbpaneler. Om varje observation tilläts identifieras med avseende på vilken webbpanel respondenten tillhör kan algoritmen identifiera paneler varifrån flera avvikande observationer härstammar.

Undersökningens resultat öppnar för möjligheten att detektera observationer som antas vara icke-representativa och därför orättvist beskriver populationen. Den konkluderat mest effektiva algoritmen, Otey, är ett verktyg för särskiljning av dessa avvikare vilket tillåter detaljstudier, individuell uppföljning eller nedviktning av observationerna. Även

om endast fåtalet observationer detekteras som avvikare kan hantering av dessa, beroende bl.a. på hur totalresultatet återges, leda till betydande skillnader.

# Referenser

## Böcker

Hawkins, D. (1980). *Identification of outliers*. London: Chapman & Hall.

## Vetenskapliga artiklar

Agrawal, R., Srikant, R. (1994). Fast algorithms for mining association rules. I *Proc. of the International Conference on Very Large Data Bases VLDB*, 487–499.

Amidan, B. G., Ferryman, T. A., Cooley, S. K. (2005). Data outlier detection using the Chebyshev theorem. I *Aerospace Conference, 2005 IEEE*, 3814-3819.

Barge, S., Gehlbach, H. (2012). Using the Theory of Satisficing to Evaluate the Quality of Survey Data. *Research in Higher Education*, Vol. 53, Nr. 2, 182–200.

Barnett, V. (1978). The study of outliers: purpose and model. *Applied Statistics*, Vol. 27, Nr. 2, 242-250.

Battipaglia, P. (2002). Selective editing to increase efficiency in survey data processing – An application to the Bank of Italy’s Business Survey on Industrial Firms. *ifc Bulletin*, 149.

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: identifying density-based local outliers. I *ACM Sigmod Record*, Vol. 29, Nr. 2, 93-104.

Chambers, R. L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, Vol. 81, Nr. 396, 1063-1069.

Chambers, R. L., Ren, R. (2004). Outlier Robust Imputation of Survey Data. I *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 3336-3344.

Chang, L., Krosnick, J. A. (2009). National surveys via RDD telephone interviewing versus the internet comparing sample representativeness and response quality. *Public Opinion Quarterly*, Vol. 73, Nr. 4, 641-678.

Christensen, L., Martinsson, J. (2012). Field Work, Survey Completion Times and Data Quality in Citizen Panel 4. *LORE Methodological Notes* 2013:1.

Cole, J. S., McCormick, A. C., Gonyea, R. M. (2012). Respondent use of straight-lining as a response strategy in education survey research: Prevalence and implications. Presenterad vid *Annual meeting of the American Educational Research Association*, i Vancouver, Kanada.

Eltinge, J. L., Cantwell, P. J. (2006). Outliers and influential observations in establishment surveys. *Federal Economic Statistics Advisory Committee*.

Filzmoser, P., Maronna, R., Werner, M. (2008). Outlier identification in high dimensions. *Computational Statistics & Data Analysis*, Vol. 52, Nr. 3, 1694–1711.

Franklin S, Brodeur, M. (1997). A practical application of a robust multivariate outlier detection method. *Proceedings of the Survey Research Methods Section, ASA*, 186-191.

- He, Z., Deng, S., Xu, X. (2005). An optimization model for outlier detection in categorical data. *Advances in Intelligent Computing*. 400-409. Berlin Heidelberg: Springer.
- He, Z., Deng, S., Xu, X. (2006). A Fast Greedy algorithm for outlier mining. *Proc. of PAKDD*.
- Hidiroglou, M. A., Berthelot, J.-M., (1986). Statistical Editing and Imputation for Periodic Business Surveys. *Survey Methodology*, No. 12, 73-83.
- Hodge, V. J., Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, Vol. 22, Nr. 2, 85-126.
- Ishikawa, A., Endo, S., Shiratori, T. (2010). Treatment of Outliers in Business Surveys: The Case of Short-term Economic Survey of Enterprises in Japan (Tankan). *Working Paper 10-E-8, Bank of Japan*.
- Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of research in personality*, Vol. 39, Nr. 1, 103-129.
- Kaminska, O., McCutcheon, A. L., Billiet, J. (2010). Satisficing among reluctant respondents in a cross-national context. *Public Opinion Quarterly*, Vol. 74, Nr. 5, 956–984.
- Kapelner, A. & Chandler, D. (2010). Preventing satisficing in online surveys: A “Kapcha” to ensure higher quality data. *Proceedings of the 1st CrowdConf Conference*.
- Knorr, E. M., Ng, R. T., Tucakov, V. (2000). Distance-based outliers: algorithms and applications. *The VLDB Journal*, Vol. 8, No. 3-4, 237-253.
- Koufakou, A., Ortiz, E. G., Georgiopoulos, M., Anagnostopoulos, G. C., Reynolds, K. M. (2007). A scalable and efficient outlier detection strategy for categorical data. *In Tools with Artificial Intelligence, 2007. ICTAI 2007. 19th IEEE International Conference on*, Vol. 2, 210-217.
- Krosnick, J. A. (1991). Cognitive demands of attitude measures. *Applied Cognitive Psychology*, Nr. 5, 213–236.
- Krosnick, J. A., Narayan, S., Smith, W. R. (1996). Satisficing in surveys: Initial evidence. *New Directions for Evaluation*, Nr. 70, 29-44.
- Kurtz, J. E., Parrish, C. L. (2001). Semantic response consistency and protocol validity in structured personality assessment: The case of the NEO-PI-R. *Journal of Personality Assessment*, Vol. 76, Nr. 2, 315-332.
- Lee, W., Xiang, D. (2001). Information-theoretic measures for anomaly detection. *Proceedings of the 2001 IEEE Symposium on Security and Privacy*, 130-143.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*, Vol. 22, Nr. 140, 1-55.
- Little, R. J., Smith, P. J. (1987). Editing and imputation for quantitative survey data. *Journal of the American Statistical Association*, Vol. 82, Nr. 397, 58-68.

Loureiro, A., Torgo, L., Soares, C. (2004). Outlier Detection Using Clustering Methods: a Data Cleaning Application. *Proceedings of KNet Symposium on Knowledge-based Systems for the Public Sector*, i Bonn, Tyskland.

Malhotra (2008). Completion time and response order effects in web surveys. *Public Opinion Quarterly*, Vol. 72, Nr. 5, 914–934.

Meade, A. W., Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological methods*, Vol. 17, Nr. 3, 437-455.

Oppenheimer, D. M., Meyvis, T., Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*. Vol. 45, Nr. 4, 867–872.

Otey, M. E., Ghoting, A., Parthasarathy, A., (2006). Fast Distributed Outlier Detection in Mixed-Attribute Data Sets. *Data Mining and Knowledge Discovery*. Vol. 12, Nr 2-3, 203-228.

Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Technical Journal*. 379-423.

Shapiro, S. S., Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, Vol. 52, Nr. 3/4, 591-611.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*. Vol. 103, Nr. 2684, 677-680.

## Tidningsartiklar

DN Debatt (2012). "Dåliga undersökningar ger uppseendeväckande resultat". Hämtad 7 december 2013, från <http://www.dn.se/debatt/daliga-undersokningar-ger-uppseendevackande-resultat/>

Lund Business Review (2013). "Spelar det någon roll vem som mäter opinionen?". Hämtad 7 december 2013, från <http://review.ehl.lu.se/spelar-det-nagon-roll-vem-som-mater-opinionen/>