

Examensarbete
15 högskolepoäng, grundnivå

Klusteranalys av cykelflödesdata för identifiering av viktiga
faktorer och avvikande datapunkter

Cluster analysis of bicycle flow to identify the important factors and abnormal
datasets

Alex Tram
Mohamed Hojeij

Sammanfattning

Studien har för avsikt att förbättra kunskapen om vilka faktorer som påverkar cykelflödet en viss dag i Malmö. Vi har huvudsakligen undersökt frågor om, hur många grupperande kluster är optimalt för att kunna identifiera avvikande dagar och vilka är dess faktorer i en tidsserie cykelvolymdata? Vår arbetsmetod var att använda ett matchande tillvägagångssätt baserat på ett experiment tillsammans med en utvärderingsmetod. Arbetsmetoden skedde i en iterativ process där experimentet var att hitta rätt antal kluster och där utvärderingen var analysen av resultaten som producerades av experimentet. Datan erhållen från en cykelräknare belägen på Kaptensgatan i Malmö fick databearbetas med hjälp av normalisering då volymen av cyklister inte ska ha någon påverkan i studien. Syftet med vårt arbete är att kunna identifiera avvikande datapunkter och dess faktorer med stor inverkan på cykelflöden med hjälp av klusteranalys då detta kan leda till mer välinformerade beslut vid stads- och transportplanering. Om det gick att analysera cyklister där dessa faktorer elimineras så skulle detta leda till vidare utveckling och forskning av stor betydelse för Malmö stad. Genom att använda oss av klusteranalysen K-means och Euklidisk distans som används som beräkning av distanser inom liknande områden kunde vi finna relevanta kluster med avvikande datapunkter och faktorer med stor inverkan på cykelflödet. Vårt resultat visar att 7 kluster varav 2 av de delades upp till 6 mindre kluster, var det mest optimala för studien och faktorerna med en stor inverkan på de antal registrerade cyklister under vissa dagar kunde då identifieras bäst. Faktorerna som identifierades var evenemang, festivaler, fotbollsmatcher, konserter, lovdagar, nederbörd och röda dagar.

Nyckelord: Klusteranalys, K-Means, Cykeldata, Maskininlärning.

Abstract

This study aims to provide a deeper understanding of the different factors and their impact on the bicycle flow in Malmö during a certain day. We mainly examined the questions, what is the most optimal number of clusters needed in order to identify discrepancies, and which key factors have huge impact in a dataset? The choice of the method used in this study is a matching approach based on experiment together with an evaluation method. The work method occurred in an iterative process, where the experiment was conducted to find the right number of clusters and the evaluation was the analysing of the results that were produced by the experiment. The collected data from a bicycle counter, located in Kaptensgatan in Malmö, had to be processed with normalization to ensure that the volume of the bicycles does not affect the study. The purpose of our study is to identify discrepancies and key factors that have huge implications on the bicycle flow with the help of cluster analysis that might lead to more well-informed decision in urban planning and transportation planning. If it were possible to analyze cyclists where these factors are eliminated, this would lead to further development and research of great importance for Malmö City. By using the cluster algorithm K-means, and Euclidean distance, which is used as calculation of distances in similar areas, we could then find relevant clusters with deviating data points and key factors with great impact on the bicycle flow. Our results shows that 7 clusters, 2 of which were divided up to 6 smaller clusters, were the most optimal for the study and the factors with a large impact on the number-registered cyclists during certain days could then be best identified. The factors identified were events, festivals, football matches, concerts, rainfalls and holidays.

Key terms: Cluster analysis, K-means, Cycle data, Machine learning.

Innehållsförteckning

1	Inledning	1
1.1	Bakgrund	1
1.2	Syfte	1
1.3	Forskningsfrågor	2
1.4	Relaterad forskning	2
1.4.1	Jämförelse av klusteralgoritmer	2
1.4.2	Orsaker och påverkan av cykeltrafikanter	3
1.5	Målgrupp	4
1.6	Avgränsningar	4
2	Teori	5
2.1	Maskininlärning	5
2.1.1	Övervakad inlärning	5
2.1.2	Förstärkt inlärning	6
2.1.3	Oövervakad inlärning	6
2.2	Klusteranalys	6
2.2.1	K-Means	7
2.3	Normalisering	7
2.4	Euklidisk distans	7
2.5	Armbågsmetod	8
3	Forskningsmetod	9
3.1	Litteraturstudie	9
3.2	Databearbetning	10
3.3	Iterativt experiment	11
3.3.1	Tillämpning av Euklidisk distans och Armbågsmetod	12
3.4	Analys av kluster	13
3.5	Metodval	13
4	Resultat och analys	14
4.1	Iteration 1	14
4.1.1	Val av antal kluster	14
4.1.2	Iteration 1 - Avståndsdigram	16
4.1.3	Analys av Iteration 1	19
4.2	Iteration 2	19
4.2.1	Val av antal kluster	20
4.2.2	Iteration 2 - Avståndsdigram	22
4.2.3	Analys av Iteration 2	25
4.3	Analys av kluster	25
4.4	Resultat av Kluster 1	25
4.5	Resultat av Kluster 2	27
4.6	Resultat av Kluster 3	29
4.6.1	Resultat av Kluster 3.1	30
4.6.2	Resultat av Kluster 3.2	32
4.6.3	Resultat av Kluster 3.3	33
4.7	Resultat av Kluster 4	34
4.8	Resultat av Kluster 5	36
4.9	Resultat av Kluster 6	38

4.9.1	Resultat av Kluster 6.1	39
4.9.2	Resultat av Kluster 6.2	41
4.9.3	Resultat av Kluster 6.3	42
4.10	Resultat av Kluster 7	43
5	Diskussion	46
5.1	Användning av klusteranalys	46
5.2	Optimalt antal kluster och definition av en avvikande dag	46
5.3	Faktorer till avvikelser	46
6	Slutsats och vidare forskning	48
6.1	Slutsats	48
6.2	Vidare forskning	48
7	Referenser	49

Lista över figurer

1	Inlärningssätt inom maskininlärning.	5
2	Klusteranalys i 3D.	7
3	Exempel av armbågsmetod	8
4	Flödesdiagram av vår forskningsmetod	9
5	Cykel- och befolkningstillväxten i Malmö 2003-2017	10
6	Cykelflöde 2006–2014 vid Kaptensgatan, Malmö	11
7	Tillämpning av armbågsmetoden i Iteration 1	14
8	Tillämpning av armbågsmetoden av värdena i Tabell 1	16
9	Avståndsdiagram för kluster 1	16
10	Avståndsdiagram för kluster 2	17
11	Avståndsdiagram för kluster 3	17
12	Avståndsdiagram för kluster 4	17
13	Avståndsdiagram för kluster 5	18
14	Avståndsdiagram för kluster 6	18
15	Avståndsdiagram för kluster 7	18
16	Sammanställning av avståndsdiagrammen	19
17	Tillämpning av armbågsmetoden i Iteration 2 för kluster 3	20
18	Tillämpning av armbågsmetoden av värdena i Tabell 2	21
19	Tillämpning av armbågsmetoden i Iteration 2 för kluster 6	21
20	Tillämpning av armbågsmetoden av värdena i Tabell 3	22
21	Avståndsdiagram för kluster 3.1	23
22	Avståndsdiagram för kluster 3.2	23
23	Avståndsdiagram för kluster 3.3	23
24	Avståndsdiagram för kluster 6.1	24
25	Avståndsdiagram för kluster 6.2	24
26	Avståndsdiagram för kluster 6.3	24
27	Sammanställning av avståndsdiagrammen	25
28	Kluster 1 med avvikande punkter	26
29	Kluster 1 utan avvikande punkter	26
30	Kluster 1 med endast avvikande punkter	26
31	Kluster 2 med avvikande punkter	27
32	Kluster 2 utan avvikande punkter	28
33	Kluster 2 med endast avvikande punkter	28
34	Kluster 3 med avvikande punkter	29
35	Kluster 3 utan avvikande punkter	29
36	Kluster 3 med endast avvikande punkter	30
37	Kluster 3.1 med avvikande punkter	30
38	Kluster 3.1 utan avvikande punkter	31
39	Kluster 3.1 med endast avvikande punkter	31
40	Kluster 3.2 med avvikande punkter	32
41	Kluster 3.2 utan avvikande punkter	32
42	Kluster 3.2 med endast avvikande punkter	32
43	Kluster 3.3 med avvikande punkter	33
44	Kluster 3.3 utan avvikande punkter	33
45	Kluster 3.3 med endast avvikande punkter	34
46	Kluster 4 med avvikande punkter	35
47	Kluster 4 utan avvikande punkter	35

48	Kluster 4 med endast avvikande punkter	35
49	Kluster 5 med avvikande punkter	36
50	Kluster 5 utan avvikande punkter	37
51	Kluster 5 med endast avvikande punkter	37
52	Kluster 6 med avvikande punkter	38
53	Kluster 6 utan avvikande punkter	38
54	Kluster 6 med endast avvikande punkter	39
55	Kluster 6.1 med avvikande punkter	39
56	Kluster 6.1 utan avvikande punkter	40
57	Kluster 6.1 med endast avvikande punkter	40
58	Kluster 6.2 med avvikande punkter	41
59	Kluster 6.2 utan avvikande punkter	41
60	Kluster 6.2 med endast avvikande punkter	41
61	Kluster 6.3 med avvikande punkter	42
62	Kluster 6.3 utan avvikande punkter	42
63	Kluster 6.3 med endast avvikande punkter	43
64	Kluster 7 med avvikande punkter	44
65	Kluster 7 utan avvikande punkter	44
66	Kluster 7 med endast avvikande punkter	44

Lista över tabeller

1	Medelavståndet till centroiden per kluster	15
2	Medelavståndet till centroiden per kluster	20
3	Medelavståndet till centroiden per kluster	22
4	Sammanställning av kluster 1	27
5	Sammanställning av kluster 2	28
6	Sammanställning av kluster 3	30
7	Sammanställning av kluster 3.1	31
8	Sammanställning av kluster 3.2	33
9	Sammanställning av kluster 3.3	34
10	Sammanställning av kluster 4	36
11	Sammanställning av kluster 5	37
12	Sammanställning av kluster 6	39
13	Sammanställning av kluster 6.1.	40
14	Sammanställning av kluster 6.2	42
15	Sammanställning av kluster 6.3	43
16	Sammanställning av kluster 7	45

1 Inledning

I detta avsnitt presenteras bakgrunden till vår studie, dess syfte, vår problemställning och tidigare forskning.

1.1 Bakgrund

Cykeln har blivit en allt viktigare del av stadens transporter, bland annat på grund av dess förmåga att bidra till snabba, hållbara, miljövänliga och kostnadseffektiva transporter. Att cykla bidrar också till ett mer hälsosamt och aktivt liv [1]. Eftersom cykling ger många positiva effekter har myndigheterna stort fokus på att öka andelen cykelresor i våra städer. För att utvidga cykelns attraktionskraft har man genomfört olika typer av åtgärder, till exempel byggnation och förbättring av cykelbanor, samt att erbjuda förbättrade parkeringsmöjligheter för cyklar. Ett annat exempel är införanden av cykeldelningssystem, där man smidigt och effektivt kan låna och lämna tillbaka cyklar vid olika stationer, som i sin tur underlättar resandet med kollektivtrafiken. För att stads- och transportsplanerare ska kunna öka cykelns attraktionskraft behöver man kunskap om cykeltrafiken i den stad man vill genomföra åtgärder. Genom installation av cykelräknare på specifika platser, kan man samla in värdefull data för att öka kunskapen om cykelflöden i en stad. Cykelräknare finns till för att räkna antalet passerande cyklister i respektive riktning vid en specifik plats över tid. Med hjälp av informationen kan man sedan uppskatta cyklisternas beteenden och på så sätt hjälpa stads- och transportplanerare i sitt arbete.

I vårt arbete kommer vi att analysera en serie data som kommer från en cykelräknare som är placerad längs med Kaptensgatan i centrala Malmö. Genom klusteranalys kommer vi söka samband mellan olika dagar och på så sätt identifiera vilka faktorer som tycks ha en stor inverkan på antalet registrerade cyklar under en viss dag. Tidigare forskning har visat att exempel på faktorer som spelar in är väder, tid och veckodag [2, 6]. Arbetet kommer att utföras med hjälp av klusteranalys som är en populär metod för dataanalys. Klusteranalys är en form av maskininlärning och används mycket inom bland annat statistik, informationsutvinning, mönsterigenkänning och bioinformatik [3].

1.2 Syfte

Inom ramen för vårt arbete kommer vi använda klusteranalys för att identifiera avvikande datapunkter och faktorer med stor inverkan på cykelflöden vid en cykelräknare i Malmö. Genom att identifiera orsakerna till de avvikande datapunkterna där cykelflödet skiljer sig från en ordinär dag, syftar vårt arbete till att bidra till kunskap som kan användas till att möjliggöra mer hållbara persontransporter i urbana miljöer. Till exempel, är målet att vår studie kan bidra till att förbättra planeringen av kollektivtrafiken i Malmö. Kunskap om hur cykelflödet förändras, till exempel vid festivaler, fotbollsmatcher, etc, skulle kunna användas för att planera kollektivtrafiken effektivare för att dämpa trafiken och därmed undvika trafikstopp och trafikolyckor. Med hjälp av våra resultat är vår förhoppning även att Malmö stad ska kunna ta mer välinformerade beslut vid stads- och transportplanering.

1.3 Forskningsfrågor

För att uppnå studiens syfte kommer följande forskningsfrågor att studeras:

- FF1: Hur kan klusteranalys användas för att gruppera dagar och identifiera avvikande dagar i en tidsserie cykelvolymdata.
 - FF1.1 – Hur många kluster är optimalt?
 - FF1.2 – Hur definierar man en avvikande dag?
- FF2: Vilka faktorer har huvudsakligen en inverkan på cykelflödet vid Kaptensgatan i Malmö?

I våra forskningsfrågor är det underförstått att den tidsserie cykelvolymdata som avses är insamlad vid samma plats, av en cykelräknare vid Kaptensgatan i Malmö. För att kunna besvara FF1 har forskningsfrågan brytits ner till två underfrågor: FF1.1 och FF1.2. Genom att besvara dessa underfrågor är målet att indirekt besvara FF1.

1.4 Relaterad forskning

Malmö stad är platsen för vår undersökning och tidigare analyser på olika former av trafikflöden har undersökts men de summerar endast antalet för att se om förändringar sker över tid och mellan geografiska områden. Av vår litteratursökning fanns det inte mycket relaterat forskning inom klusteranalys gällande cykelflöden även om liknande analyser har genomförts där man eftersträvar en prediktion av antalet cyklar i en cykelstation.

1.4.1 Jämförelse av klusteralgoritmer

Inom maskininlärning finns det olika typer av oövervakad maskininlärning som har olika för- och nackdelar. Det som skiljer metoderna åt, är mängden och sortens data man har till förfogande. Archana och Prateek [11] jämför olika oövervakade maskininlärningsmetoder för att få fram bästa möjliga resultat. De använder samma data under forskningen för att testa de olika algoritmerna, utöver det så analyserade de resultatens effekt vid en normaliserad data. Metoden och algoritmen som visade bäst resultat var en normaliserad data vid användning av K-means. Medans alla andra algoritmer visade snarlika resultat med lite störningar när de grupperades.

Harous m.fl. [15] presenterar en jämförelse av olika algoritmer för oövervakad maskininlärning. I studien jämförs 13 olika klusteralgoritmer, bland annat K-means, K-modes, K-prototype etc. Genom att använda data med olika dimensioner och storlekar kan de testa varje algoritmen. Två olika typer av data används i studien, en datamängd bestående av 200 rader och 20 kolumner och en annan större datamängd bestående av 600 rader och 60 kolumner. Harous m.fl. kommer fram till att klusteralgoritmen K-means ger det bästa resultatet med lägst fel när de använder sig av stora datamängder med samma dimensioner. Nackdelen med K-means är att algoritmen själv inte bestämmer antal kluster som är optimalt, utan användaren måste experimentera sig fram till det.

1.4.2 Orsaker och påverkan av cykeltrafikanter

Tsapakis m.fl. [2] studerar hur vädret kan påverka cykeltrafikanter. Studiens fokus ligger på färdtiden och hur den påverkas av olika väderförhållanden. Deras resultat visar att regn och snöfall påverkar färdtiden negativt. Vid lätt regn kan färdtiden förlängas mellan 0,1% och 2,1%, vid måttligt regn mellan 1,5% och 3,8% och vid starka regnskuror mellan 4,0% och 6,0%. Färdtiden påverkas mest negativt vid snöfall. Lätt snöfall förlänger färdtiden mellan 5,5% och 7,6%, medans tätt snöfall kan påverka upp till 11,4%. Deras slutsats är att dessa faktorerna är en orsak till att cykeltrafikanterna har en benägenhet att ändra sitt transportmedel till något annat som till exempel bil, buss, tåg, spårvagn, metro, etc.

Mahdie m.fl. [4] presenterar en studie där de matchar korttidsräknare med långsiktiga räknare för att förbättra noggrannheten i AADB (Annual Average Daily Bicyclists) beräkningen. Eftersom kontinuerliga data för längre perioder inte finns tillgängligt på många platser, är det vanligt att samla in korttidsräknade data för ett urval av platser och sedan tillämpa en extrapoleringsmetod för att konvertera den korttidsräknade datan till en årlig räkning av data. Denna studie baseras på klusteranalys, PAM (Partitioning Around Medoids) och en övervakad inlärningsmetod, KNN (K-Nearest Neighbour). Deras modell utvärderas i termer av AADB-estimate Absolute Percent Error (felprocent i uppskattningen) och resultaten visar att de uppnår mindre eller lika stora fel än befintliga metoder. Genom att integrera en klusteranalys och en övervakad inlärningsmetod uppnådde de en stadig modell som kan förbättra cykelvolymen.

Mohamed m.fl. [5] studerar problemet med missade datavärden på grund av periodiska funktionsfel, som kan uppstå i automatiska räknare, till exempel loop-detektorer och cykelräknare. Sporadiska borttappade datavärden påverkar den totala cykelvolymen som i sin tur påverkar bland annat den väsentliga säkerhetsanalysen av cyklisterna. Modellen som de använder i denna studien är dynamisk eftersom den inte antar någon förkunskap om vilka platser som kan uppleva funktionsfel. Modellen kallas för, Autoencoder neuralt nätverk och metodens prediktionsmodul är avsedd för kortsiktiga framtida prognoser för cykelvolym. Resultatet visar en stark uppskattningskraft med ett genomsnittligt fel på cirka 10%. En mycket stark korrelation observerades i de flesta fallen och visar potential för att utveckla en modell med stark förutsägbar förmåga. Denna forskning gynnar speciellt i trafikplanering och operationsanalys för icke-motoriserad trafik.

Holmgren m.fl. [8] presenterar en jämförelse av maskininlärningsalgoritmer för uppskattning av cykelflöden baserat på en cykelräknare i Malmö och väderdata från SMHI. Målet med deras studie är att få fram den bästa maskininlärningsalgoritmen som ger det mest tillförlitliga resultat för att uppskatta antalet cyklisterna. Algoritmerna som visade bäst resultat för uppskattning av cykelflöden var Random SubSpace och Bagging. Variabeln som hade störst påverkan i cykelflöden var datum. När Holmgren m.fl. avlägsnade temperaturvariabeln så presterade algoritmerna bättre och gav en högre korrelation.

Holmgren m.fl. [7] presenterar en annan studie där de också använder sig av regression för att prediktera antalet cyklar registrerade av en cykelräknare belägen på Kaptensgatan i Malmö. Syftet med denna studien är att jämföra två regressionsproblem med två olika målvariabler, faktiska antal cyklar och avvikelser från en långsiktig trendberäkning av förväntad antal cyklar. Resultaten från studien visade att stödvektorregression och regressionsträd var mest lämpligast för arbetets mål och syfte.

Holmgren m.fl. [6] studerar cykeltrafiken i Malmö för att sedan prediktera framtida cyklisterna med hjälp av regressionsalgoritmer. Syftet med arbetet är att förbättra noggrannheten

i prediktionen av cyklisterna i Malmö. Med hjälp av insamlad data från en cykelräknare så kunde de undersöka datan och komma fram till en uppskattning av cykeltrafiken i Malmö. Ett annat viktigt syfte är att kvantifiera olika faktorer som kan påverka cykeltrafiken vid en viss punkt i trafiknätet. De faktorer som anses som viktiga är veckodag, årstid och väder (temperatur och nederbörd). Proceduren gick till att forskarna jämförde ett antal regressionsalgoritmer för att sedan ta reda på vilken/vilka algoritmer som är bäst för det övervägda problemet. Till exempel i förhållande till det relativa felet lyckades de förbättra resultaten från cirka 90% till 30% för de bästa algoritmerna. De lyckades få ett positivt resultat och förbättrade den väsentliga noggrannheten jämfört med tidigare resultat.

1.5 Målgrupp

Studien riktar särskilt till målgrupper med intresse och ansvar för trafikfrågor, främst kommuner och regioner, men även till personer som har intresse och engagemang inom detta fält. Forskare och analytiker inom snarlika områden kan också ha nytta av studien.

1.6 Avgränsningar

Vi har valt att avgränsa arbetet till en tidsserie cykelflödesdata som är erhållen från en cykelräknare vid Kaptensgatan i Malmö och väderdatan från SMHI mellan årsperioden 2006-2014. Vi kommer att genomföra klusteranalysen med metoden K-Means [9, 10, 15] för att identifiera avvikande datapunkter och dess faktorer. Faktorer som arbetet avgränsas till är huvudsakligen variabler som veckodag och tid, men även faktorer som till exempel evenemang, väder, årstid, röda dagar, festivaler etc. kan dyka upp som en avvikande datapunkt under vissa typer av dagar.

2 Teori

I detta avsnitt beskrivs teorin som är viktig för att kunna tillgodogöra sig vår rapport. Detta för att ge läsaren en övergripande förståelse inom vårt område i arbetet.

2.1 Maskininlärning

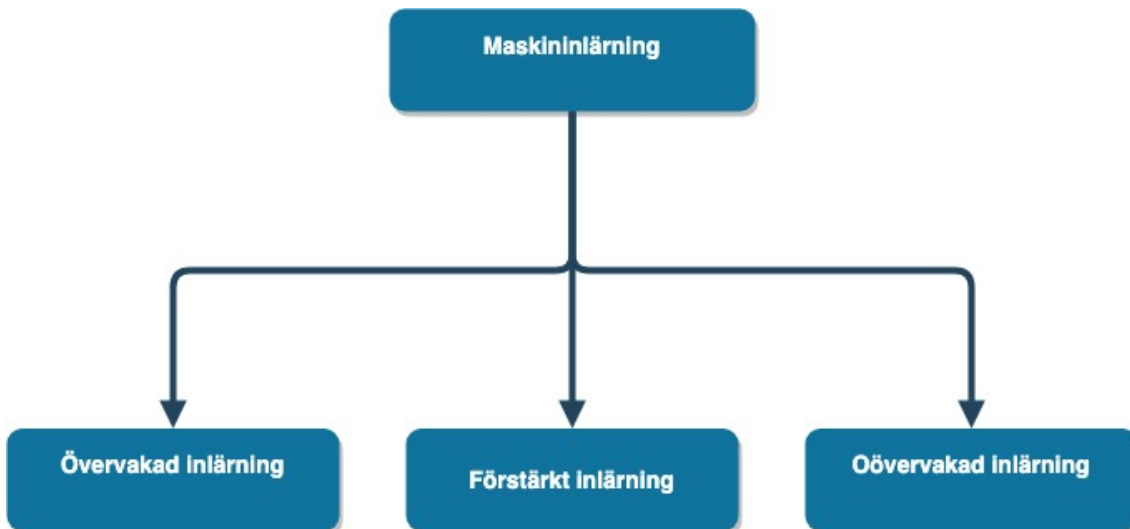
Inom datavetenskap är maskininlärning ett forskningsområde som växt fram ur AI (Artificiell Intelligens) och som går ut på att utveckla maskiners förmåga att självständigt förstå och hantera stora datamängder. Maskininlärning handlar om att bygga system som lär sig utifrån erfarenhet. Med hjälp av algoritmer kan datorn tolka och lära sig från de data den bearbetar för att sedan kunna förutse mönster. Beroende på vilken data man tränar och vilket mönster man vill identifiera så delar man ofta in området i olika inlärningssätt: övervakad inlärning, förstärkt inlärning och oövervakad inlärning [8] (se Figur 1).

Processen maskininlärning består huvudsakligen av tre steg:

Steg 1 – Inlärning. Första steget handlar om att välja inlärningssätt för att sedan träna maskinen och utföra uppgiften.

Steg 2 – Träning. Andra steget efter man har valt inlärningssätt, handlar om att träna modellen för att få så korrekt utfall som möjligt.

Steg 3 – Utvärdering. Med den tränade modellen utvärderar man resultaten med hjälp av den testdatan som återstår. Detta är en ständig process som kommer göras bättre ju mer data man tillför.



Figur 1: Inlärningssätt inom maskininlärning.

2.1.1 Övervakad inlärning

Övervakad inlärning innebär att man förfinar en algoritm genom att träna datorn med märkt data (träningssmängd), det vill säga data som innehåller exempel med givna korrekta svar och därefter utvärderar man hur väl algoritmen fungerar med en annan datamängd

(testmängd). Målet med denna inlärning är att lära ett system att förutsäga värdet av en beroende variabel. Ett exempel på användning av denna typ av inlärning är när man vill identifiera falska kreditkort genom att använda en datauppsättning med både falska och giltiga debiteringar för att träna modellen. Klassificering och regression är två typer av övervakad inlärning som används mycket inom maskininlärning. Skillnaden mellan klassificering och regression är att inom klassificering är syftet att prediktera en klass eller kategori på den nya observationen medans inom regression vill man prediktera ett kontinuerligt värde [6].

2.1.2 Förstärkt inlärning

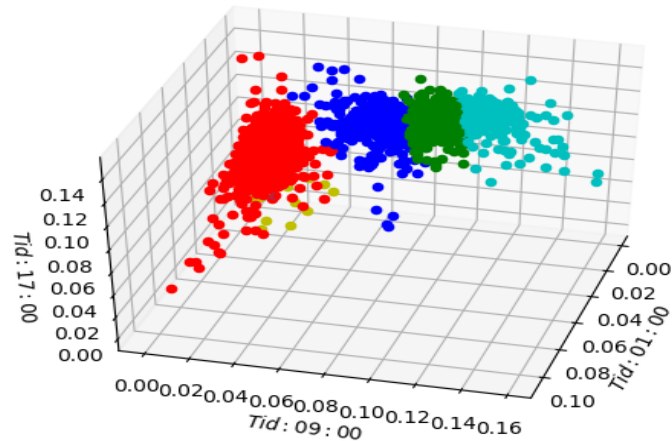
Förstärkt inlärning innebär att en eller flera mjukvaror lär sig den rätta lösningen genom provning utan att ha tillgång till förberedd data. Denna typ av inlärningssätt passar bra i sammanhang där mjukvara behöver komma fram till optimala lösningar i en viss miljö. Målet med denna inlärning är att generalisera och anpassa ett beteende utifrån återkoppling som kommer då och då. En bra jämförelse är till exempel inlärning hos djur [8].

2.1.3 Övervakad inlärning

Övervakad inlärning är en inlärningsalgoritm som grupperar datainstanser så att de instanser blir baserad på deras statistiska egenskaper. Denna typ av inlärning tränar algoritmer med omärkta data, där syftet är att identifiera samband. Ett exempel på användning av denna typ är när man vill identifiera och gruppera kunder med liknande köpvanor. Klusteranalys är en vanlig typ av oövervakad inlärning som används mycket inom maskininlärning [16].

2.2 Klusteranalys

Klusteranalys är ett samlingsnamn för en analysmetod och refererar inte till en enskild algoritm. Klusteranalys är en form av oövervakad maskininlärning som används mycket inom statistik, informationsutvinning, mönsterigenkänning, bildanalys, informationssökning och bioinformatik. Inom datavetenskap och statistik innebär klusteranalys att man grupperar en mängd datainstanser i delmängder som kallas kluster. Målet med denna typ av analys är att hitta kluster, där datainstanserna på något logiskt sätt har ett samband [3]. Processen genomförs genom att man definierar en avståndsfunktion, utifrån de datamängder som beskriver ett element. Därefter beräknas avstånden mellan klustren med hjälp av ett avståndsmått som till exempel euklidiskt distans, som sedan möjliggör visualisering (upp till 3D) av klustren (se Figur 2).



Figur 2: Klusteranalys i 3D.

2.2.1 K-Means

K-means är en oövervakad inlärningsalgoritm som är en populär typ av klusteranalys. Denna algoritm används vanligtvis vid data mining och mönsterigenkänning. Syftet med algoritmen är att identifiera rätt antal kluster och avvikande datapunkter för att tillfredsställa ett viss kriterium. Med hjälp av ett avståndsmått kan man utvärdera hur väl de identifierade klustren är [9].

Tillvägagångsättet utgår från följande tre steg, där steg 2 och steg 3 itereras tills ingen omlokalisering av klustren behövs:

Steg 1: Gruppera in data i k-antal kluster. Där k är ett antagande från användaren. K-means bestämmer sedan en slumpmässig centroid för varje k-kluster. Olika startindelningar kan ge olika slutindelningar.

Steg 2: Associera varje datapunkt till det klustret vars centroid ligger närmast med hjälp av ett avståndsmått, till exempel Euklidiskt distans.

Steg 3: Beräkna avståndet mellan de nya centroiderna och datapunkterna för varje kluster

2.3 Normalisering

Normalisering har olika betydelser inom olika områden, men rent allmänt syftar det på en funktion där man omvandlar ett objekt till en viss önskvärd form. Normalisering inom statistik innebär att man omvandlar värden till jämförbara värden. Till exempel vid beräkning av material, så omvandlas värden till samma måttenheter och tidsperioder sammanfaller.

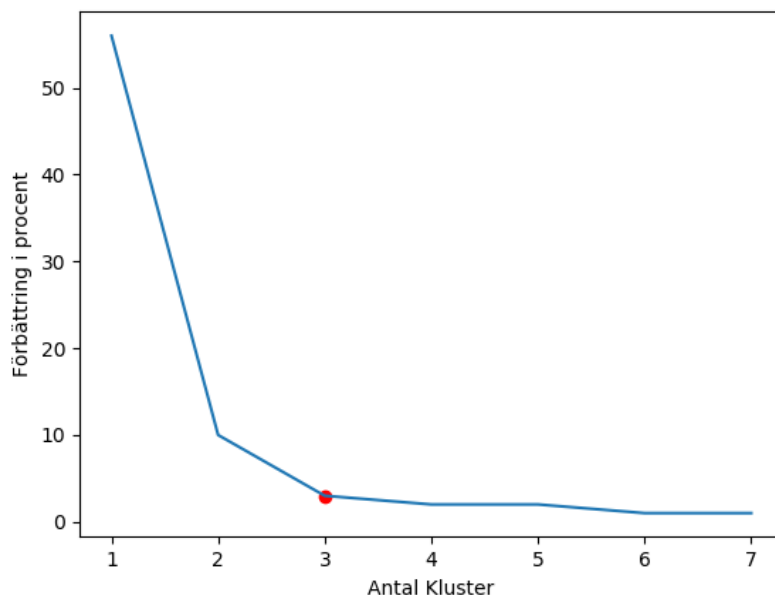
2.4 Euklidisk distans

Euklidisk distans är en metod som används för att beräkna avståndet mellan två olika punkter i en eller flera dimensioner. Euklidisk distans används ofta inom matematik och

fysik, men även inom klusteranalys. Inom klusteranalys används Euklidisk distans för att beräkna avståndet från en datapunkt i klustret till klustrets centroid och med hjälp av avståndet kan man definiera datapunktens innebörd.

2.5 Armbågsmetod

Armbågsmetoden (eng: Elbow-point) är en metod som ser på andelen av variansen i funktionen av antal kluster. Detta är en metod som är essentiell för vår studie eftersom K-means inte har en inbyggd funktionalitet för att bestämma optimalt antal kluster. Tanken med metoden är att man ska välja rätt antal kluster för sin forskning, det ska vara en klar förbättring av modelleringen av data vid ökning av kluster. De första klustren ger en klar förbättring och lägger till mycket information men vid en viss punkt kommer förbättringen att falla dramatiskt och då ges en vinkel i grafen (se Figur 3). Vid vinkeln kan man sedan avläsa hur många kluster som är optimalt för ens forskning [9].

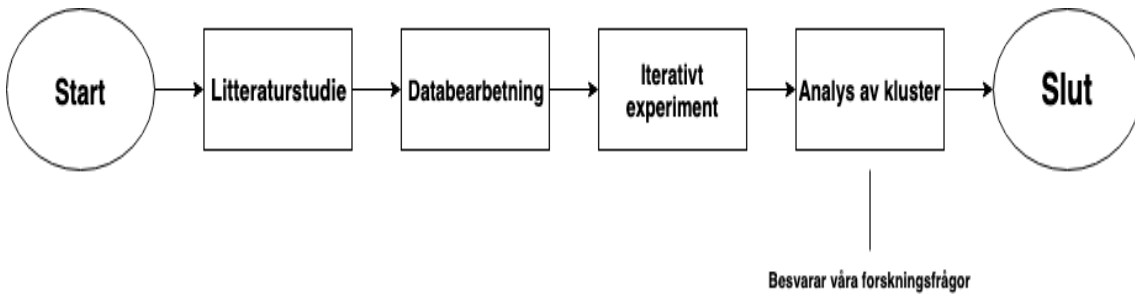


Figur 3: Exempel av armbågsmetod

3 Forskningsmetod

I detta avsnitt redovisas forskningsmetodiken arbetet använt sig av för att besvara forskningsfrågorna. Här presenteras även bakgrunden kring de erhållna datamängderna, arbetets experiment och utvärderingsmetod, samt motiven bakom metodbesluten.

Se Figur 4 för en illustration, i form av ett flödesdiagram, över de moment som ingår i vår forskningsmetod. Vårt iterativa experiment syftar huvudsakligen till att identifiera tillräckligt många homogena kluster för att kunna besvara våra forskningsfrågor i momentet, Analys av kluster.



Figur 4: Flödesdiagram av vår forskningsmetod

Med tydliga steg och ett experiment som är lämpat för studiens syfte blir det lättare att forma arbetet och att analysera vad som lyckats och vad som inte lyckats. Det första momentet, Litteraturstudie innefattar insamling av litteratur och tidigare forskning som relaterar till vårt problem. Syftet med litteraturstudien var att lära oss om de metoder och teorier som är nödvändiga för att konstruera, genomföra och analysera vårt experiment, som bygger på klusteranalys. Moment 2, Databearbetning innefattar bearbetning av de data som vi använt i vår studie, dvs en tidsserie cykelvolymdata som samlats in av cykelräknaren belägen på Kaptensgatan i Malmö och de väderdata som vi laddat ner från SMHI:s API[19]. Med de omvandlade datan kunde vi därefter utföra vårt iterativa experiment i moment 3, med hjälp av klusteranalys. Moment 3 och moment 4 är essentiella för att kunna besvara våra forskningsfrågor.

3.1 Litteraturstudie

För att genomföra vår litteraturstudie använde vi oss av Jacobsen's metodik [12]. Jacobsen menar att den information som en forskare använder sig av, som redan samlats in och blivit presenterad av en annan forskare, är en sekundärkälla som även kallas för sekundärdata. Insamlingen av teoretiska fakta kommer oftast från andra studier med ändamål som inte är lika, men som andra forskare finner intressant att dela med sig av.

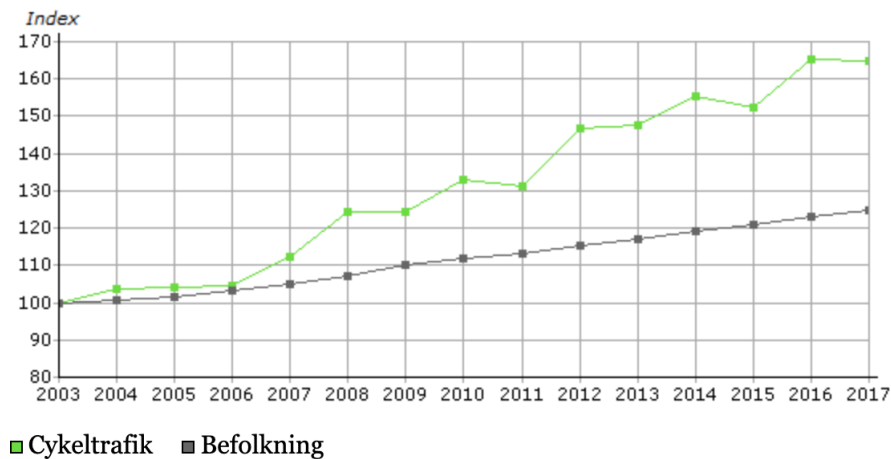
Syftet med vår litteraturstudie var att identifiera forskning inom vårt område samt att bygga upp vår kunskap om klusteranalys för att på ett korrekt sätt kunna genomföra och analysera vårt experiment. Vi har identifierat relaterat arbete i form av tidskriftsartiklar, konferensartiklar och böcker med hjälp av sökmotorerna Libsearch, Goggle Scholar, IEEE, Scencedirect och ACM. För att undvika föråldrad forskning begränsade vi våra sökningar till studier från slutet av 1990-talet och framåt. Genom granskning av studiernas abstrakt och resultat kunde vi därefter skapa ett mindre urval och med det mindre urvalet kunde vi komma fram till ett antal artiklar och böcker som var mest relevanta för vår studie.

Vår litteraturstudie gav oss viktiga förkunskaper om ämnet som studien handlar om. Med dessa förkunskaper underlättades arbetsprocessen och arbetet kunde därmed utföras på ett förberett sätt. Genom artiklarna [9, 10, 15] kunde vi besluta att K-means och Arm-bågsmetoden är två essentiella metoder för vårt arbete.

3.2 Databearbetning

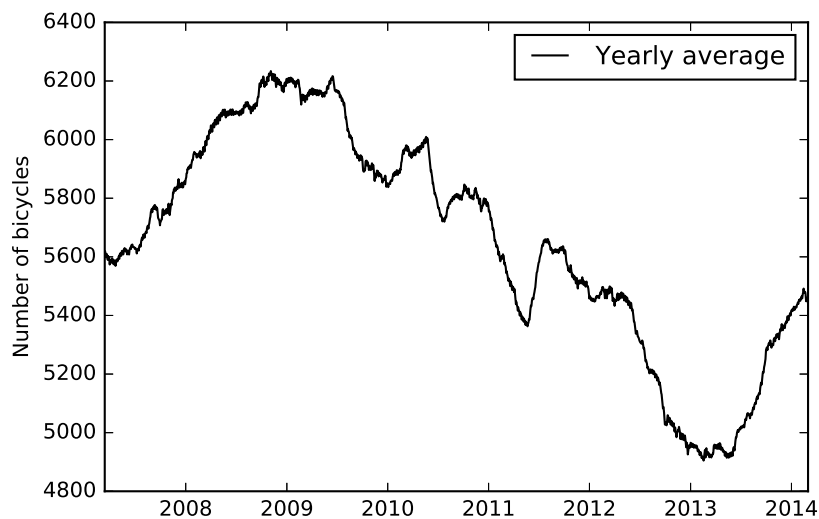
Inom ramen för vår studie har vi använt cykelflödesdata som samlats in av en cykelräknare som är placerad vid Kaptensgatan i Malmö, samt väderdata som vi laddat ner från Sveriges Meteorologiska och Hydrologiska Institut (SMHI). Den tidsserie cykelflödesdata som studien är baserad på är insamlad mellan åren 2006–2014 och den anger information om antal passerande cyklar i cykelvägens båda riktningar för varje timme under den nämnda tidsperioden. De väderdata vi använt, specificerar mängden nederbörd för varje timme.

Ser man till den generella cykeltrafikstillväxten i Malmö de senaste 10–15 åren så ser man en tydlig uppåtgående trend i Figur 5 som är tagen ur Gatukontoret i Malmö [18], dock så följer inte det registrerade cykelflödet i cykelräknaren vid kaptensgatan den generella tillväxten i Malmö, som istället visar en negativ trend efter 2009 enligt tidigare forskning [6] (se Figur 6). En viktig anledning till den nedgång vi observerat vid Kaptensgatan kan bero på att den nya tågstationen vid triangeln öppnades i december 2010, samt att man införde MalmöExpressen (stadsbusslinje som körs parallellt med Kaptensgatan) under den perioden.



Datakälla: Gatukontoret

Figur 5: Cykel- och befolkningstillväxten i Malmö 2003-2017



Figur 6: Cykelflöde 2006–2014 vid Kaptensgatan, Malmö

Vår databearbetning (normalisering) bestod huvudsakligen av att normalisera de datapunkter som ingår i vår klusteranalys och har en viktig betydelse eftersom att antalet cyklar som passerar Kaptensgatan förändras över tid. För att kunna jämföra cykelflödet för varje dag utan att ta hänsyn till antal cyklar under en viss dag utförde vi normalisering av cykelflödesdatan. Normaliseringen utförde vi genom att normalisera de histogram som beskriver hur antalet cyklar under varje dygn fördelas över dygnets 24 timmar. Normaliseringsprocessen genomförde vi genom att dividera varje av de 24 staplarnas värden med totala antalet cyklar för det aktuella dygnet. På så sätt ersätts antalet cyklar för respektive timme med andelen av dygnets cyklar för samma timme. Normaliseringen sker via formeln:

Normaliseringen sker via formeln:

$$x_{td}^n = \frac{x_{td}}{y_d}, \text{ där} \quad (1)$$

x_{td}^n är det normaliserade värdet för timme t och dygn d .

x_{td} är det faktiska antalet cyklar under timme t och dygn d .

y_d är det totala antalet cyklar under dygn d , dvs, $\sum_{t=1}^{24} x_{td}$

3.3 Iterativt experiment

I detta skedet utförde vi ett iterativt experiment med syfte till att identifiera det mest optimala antalet homogena kluster för vår studie, för att sedan kunna analysera och besvara våra forskningsfrågor i nästa avsnitt Analys av kluster 3.4. Detta experiment består av två moment som sker i en iterativ process: klusteranalys och utvärdering av kluster avseende dess homogenitet. För att analysera klustren som producerades av klusteranalysen är det viktigt att i förväg definiera hur detta ska utformas för att kunna uppnå målet med studien.

För att experimentera de normaliserade datapunkterna så använde vi oss av klusteranalysen K-means. K-means används med syftet att identifiera rätt uppsättning av homogena kluster för att underlätta analysen om de avvikande datapunkterna i nästa avsnitt. Denna iterativa processen utgick från följande steg, vi började med ett kluster och utvärderade klustret genom dess avvikelser, datapunkter, felfaktorer, medelavstånd och totala avstånd till centroiden, samt vilka dagar som har hamnat i klustret. För att beräkna medelavståndet och totala avståndet från datapunkterna till centroiden använde vi oss av Euklidisk distans (läs mer i underavsnittet, Euklidisk distans 3.3.1). Därefter la vi till ett kluster och utvärderade på samma sätt tills vi kunde bedöma resultaten på bästa möjliga sätt. Denna iterativa processen varade fram till 30 kluster då vi ansåg att förbättringen efter varje ökning av kluster inte längre var väsentlig. Detta kunde vi sedan bekräfta med hjälp av armbågsmetoden. (läs mer i underavsnittet, Armbågsmetoden 3.3.2)

3.3.1 Tillämpning av Euklidisk distans och Armbågsmetod

Med hjälp av Euklidisk distans kunde vi definiera en datapunkts innebörd. Beräkningarna nedanför var essentiella för att möjliggöra Armbågsmetoden, som i sin tur var essentiell för vår identifiering av rätt antal kluster och avgränsning av avvikande datapunkter. Beräkning av avstånd sker med hjälp av följande formel:

$$D(d) = \sqrt{\sum_{i=1}^{24} (\beta_i - \alpha_i)^2}, \text{ där} \quad (2)$$

$D(d)$ är avståndet från en datapunkt till klustrets centroid och d är datapunkten man vill räkna avståndet för.

β är datapunktens koordinater

α är centroidens koordinater

Med hjälp av föregående ekvation (2) kunde vi beräkna det totala avståndet för ett specifikt kluster genom att vi summerade alla datapunkternas avstånd i klustret. Beräkning av det totala avstånd i ett kluster sker med hjälp av följande formel:

$$t_d(k) = \sum_{i=1}^n D(i), \text{ där} \quad (3)$$

$t_d(k)$ är det totala avståndet för alla datapunkter i ett kluster k

n är datapunkterna i klustret.

$D(i)$ är avståndet för en specifik datapunkt i , dvs $\sqrt{\sum_{i=1}^{24} (\beta_i - \alpha_i)^2}$

Med hjälp av föregående ekvationer (2 och 3) kunde vi sedan beräkna det totala avståndet för alla kluster. Detta gjorde vi för att möjliggöra armbågsmetoden. Beräkning av det totala avståndet för alla kluster sker med hjälp av följande formel:

$$t_{total}(n) = \sum_{i=1}^n t_d(i), \text{ där} \quad (4)$$

$t_{total}(n)$ är det totala avståndet för alla kluster där n är antal kluster.

$t_d(i)$ är det totala avståndet för ett specifikt kluster i , dvs $\sum_{i=1}^n D(i)$

För att kunna bekräfta ett optimalt antal kluster använde vi oss av armbågsmetoden [9]. Detta gjorde vi genom att använda oss av beräkningarna (2), (3) och (4). Med värdena vi fick ut därifrån, kunde vi sedan tillämpa armbågsmetoden med hjälp av följande formel:

$$P = \frac{t_{total}(i-1) - t_{total}(i)}{t_{total}(i)} \cdot 100, \text{ där} \quad (5)$$

P är förbättringen i procent

$t_{total}(i)$ är det totala avståndet för alla kluster i , som bildas av datapunkterna

$t_{total}(i-1)$ är det totala avståndet för alla kluster i , som bildas av datapunkterna men subtraherat ett kluster

3.4 Analys av kluster

Genom att analysera hur klustren är utspridda och beräkna både datapunkternas och klustrens avstånd, kunde vi sedan gruppera dagarna med något logiskt samband och sedan identifiera de avvikande dagarna som forskningen är ute efter. Med de funna avvikande dagarna kunde vi sedan finna de flesta faktorerna genom djupare bakgrundsundersökning för de dagarna.

3.5 Metodval

Den generella metodiken i studien är huvudsakligen baserad på ett kontrollerat experiment [14]. Kontrollerat experiment är ett vetenskapligt test som manipuleras direkt av en forskare för att testa en enda variabel i taget, i studiens fall, k-värdet som motsvarar antal kluster. K-värdet är variabeln som testas och är den oberoende variabeln som justeras för att se effekterna på klusteranalysen som ska utvärderas.

Valet av metod gällande dataanalysen grundas i att det är en kvantitativ metod då vi använde oss av historiska kvantitativa data, vilket lämpar sig väl för en statistik analys med maskininlärning. Enligt Bryman [17] är kvantitativa undersökningar mest lämpade för forskare som vill få fram siffror genom exempelvis statistik där informationen ska bli generaliserade. Bryman [17] menar också att det är forskaren som strukturerar upp hela undersökningen genom sina frågor, vilket skiljer sig mot en kvalitativ undersökning där typiskt är deltagarens perspektiv som är i fokus. En kvalitativ undersökning med till exempel enkätundersökning eller intervjufrågor är nästintill omöjlig för att ge oss den data vi behöver för att kunna både driva tester med klusteralgorithm och besvara våra forskningsfrågor då vår data är objektiv avseende datum, riktning och tid.

4 Resultat och analys

I detta avsnitt presenteras resultaten från vårt experiment. Datan som har bearbetats, genomgått experimentering och analyserats utgör resultaten som sammanställts. Strukturen av avsnittet ska ge läsaren en djupare förståelse gällande våra bästa iterativa körningar med olika antal kluster. Detta avsnitt ligger även till grund för analysdelen och i slutet av varje iteration kommer en större analys att presenteras.

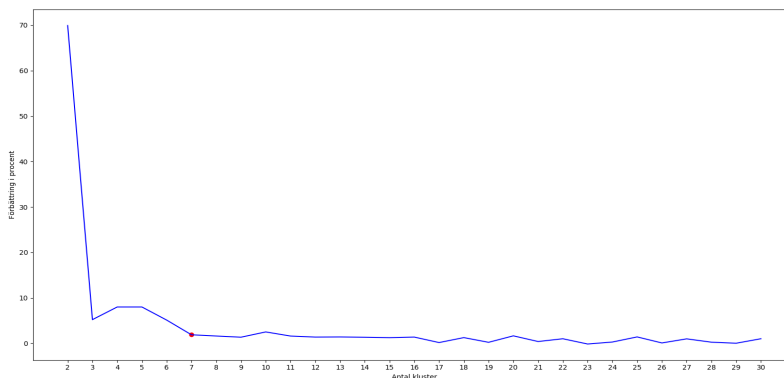
Diagrammet och tabellen nedan (Avsnitt 4.1.1) visualiserar resultaten från armbågsmetoden för val av antal kluster. Figur 7 omfattar en iteration mellan 1-30 kluster, där den röda pricken omger det mest optimala antal homogena kluster för vår studie (7st).

4.1 Iteration 1

Iteration 1 är ett klusteranalys experiment av en iterativ form som utfördes med syftet att hitta det mest optimala antal homogena kluster för vår studie. Iterationen består av klusteranalys (algoritmen K-means) på all vår datamängd, dvs alla dagar i vår tidsserie cykelvolymdata.

4.1.1 Val av antal kluster

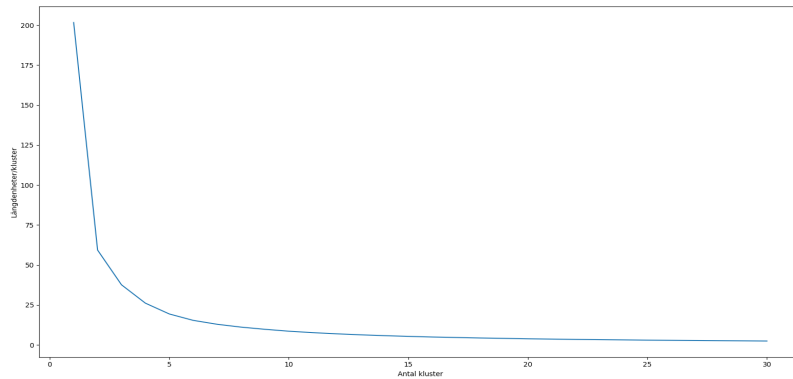
I Figur 7 har vi tillämpat armbågsmetoden (ekvation 5). Ur diagrammet kan man avläsa att grafen stabiliserar sig vid punkterna 7, 17, 19, 21 och 23. Vi ansåg att 7 kluster var det mest optimala för vår studie. För att förstärka vårt antagande så beräknade vi medelavståndet till centroiden för alla kluster och resultaten visar (se Tabell 1) att värdena sjunker dramatiskt fram till sju kluster och stabiliserar sig därefter. Nedan kommer vi presentera en mer detaljerad analys kring de 7 identifierade klustren från iteration 1.



Figur 7: Tillämpning av armbågsmetoden i Iteration 1

Medelavståndet till centroiden / kluster	
Antal kluster (x)	Längdenheter
1	201.65
2	59.33
3	37.60
4	26.11
5	19.35
6	15.34
7	12.90
8	11.12
9	9.75
10	8.56
11	7.66
12	6.93
13	6.30
14	5.78
15	5.33
16	4.93
17	4.63
18	4.32
19	4.08
20	3.81
21	3.62
22	3.42
23	3.28
24	3.13
25	2.97
26	2.85
27	2.72
28	2.62
29	2.53
30	2.42

Tabell 1: Medelavståndet till centroiden per kluster

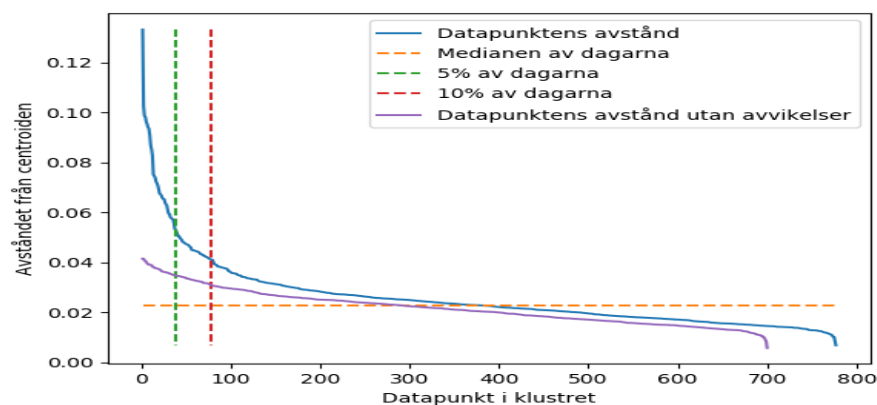


Figur 8: Tillämpning av armbågsmetoden av värdena i Tabell 1

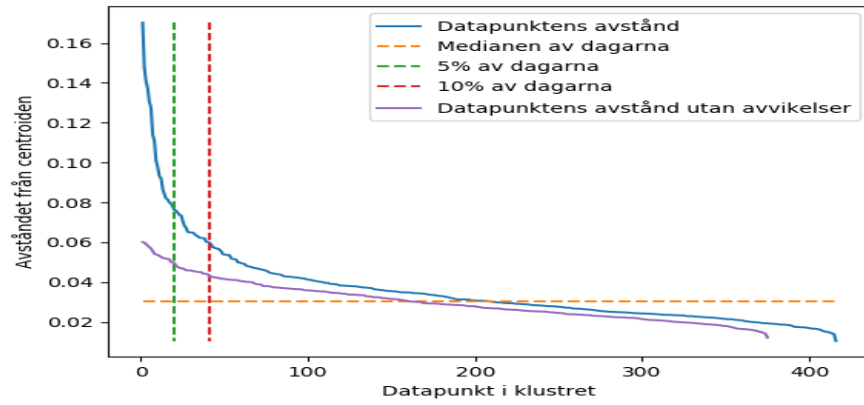
4.1.2 Iteration 1 - Avståndsdigram

I avståndsdigrammen i Figur 9 - 15 illustrerar vi datapunkternas Euklidiska distans från respektive centroid för de sju klustren som vi genererat i iteration I av vår klusteranalys. Varje avståndsdigram innehåller fem olika linjer där linjernas färger har olika tolkningar:

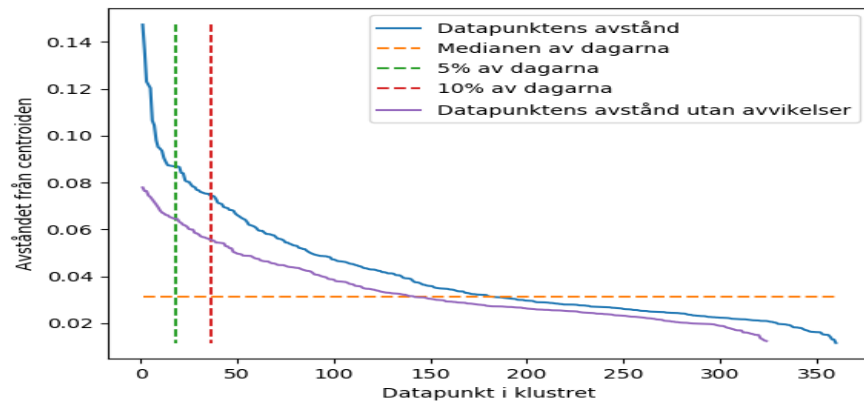
- Blåa linjen motsvarar datapunkternas avstånd från centroiden och är sorterad från störst till minst.
- Gula linjen motsvarar medianen av datapunkterna.
- Gröna linjen motsvarar 5% av datapunkterna med de största avstånden från centroiden.
- Röda linjen motsvarar 10% av datapunkterna med de största avstånden från centroiden.
- Lila linjen motsvarar den blåa linjen utan de eliminerade 10% datapunkterna (avvikelserna) med de största avstånden från centroiden.



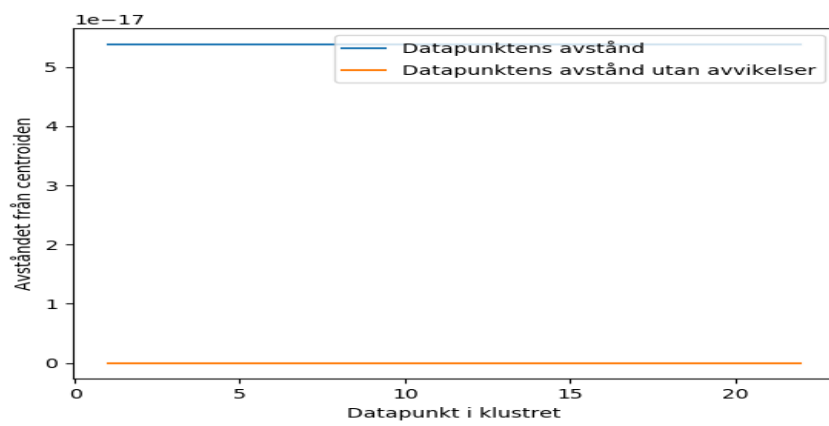
Figur 9: Avståndsdigram för kluster 1



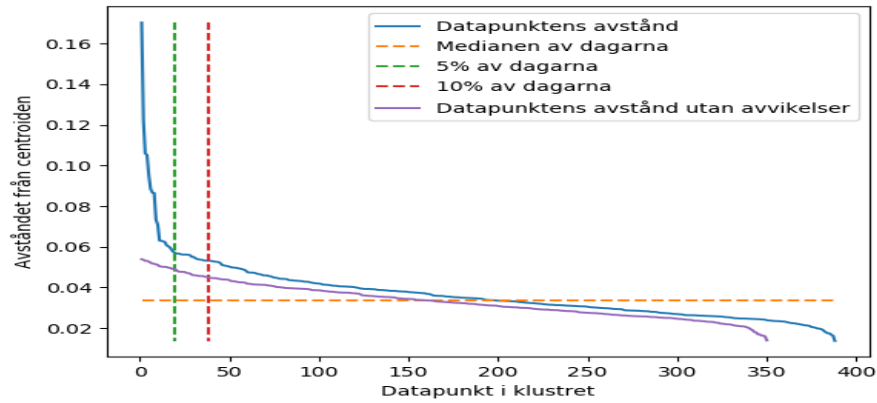
Figur 10: Avståndsdiagram för kluster 2



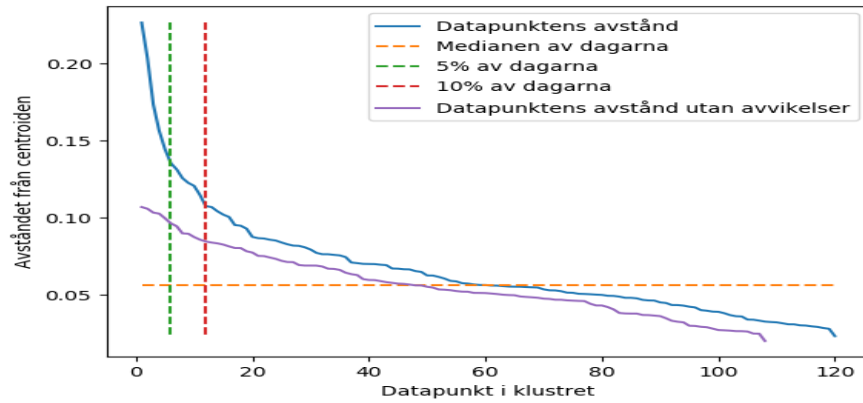
Figur 11: Avståndsdiagram för kluster 3



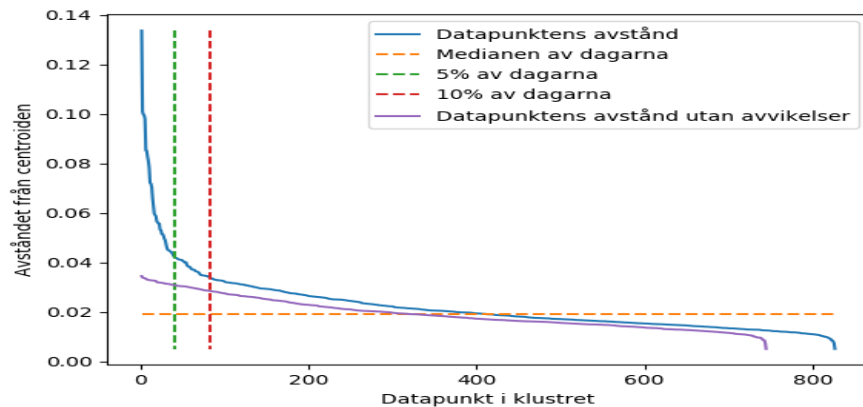
Figur 12: Avståndsdiagram för kluster 4



Figur 13: Avståndsdigram för kluster 5

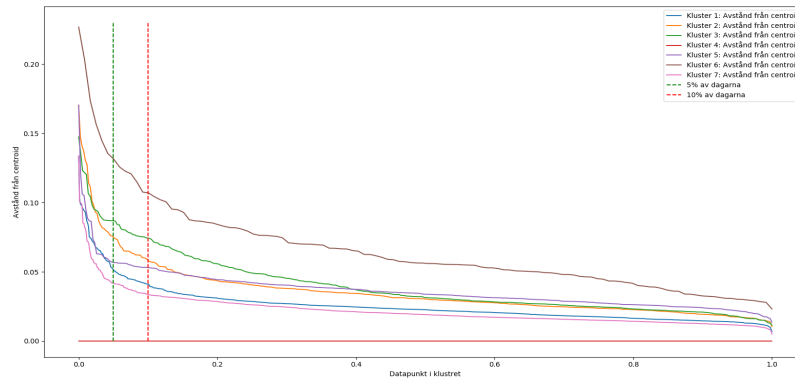


Figur 14: Avståndsdigram för kluster 6



Figur 15: Avståndsdigram för kluster 7

Avståndsdigrammet i Figur 16 är en sammanställning från de sju avståndsdigrammens blåa linjer (datapunkternas avstånd från respektive centroid).



Figur 16: Sammanställning av avståndsdiagrammen

4.1.3 Analys av Iteration 1

Baserat på armbågsmetoden i avsnitt 3.3.1 fick vi fram resultatet på 7 kluster, vilket beror på att andelen av variansen i funktionen av antal kluster avtar efter 7 kluster som man kan avläsa ur grafen i Figur 7. Vi kunde efter vår experimentering och utvärdering konstatera att vid 17, 19, 21 och 23 kluster så existerade det obehövligen kluster. Till exempel så kunde vi få flera kluster som motsvarade samma veckodag med snarlika flöden, skillnaden var att volymen av cyklisterna hade högre höjder vid några tidpunkter under dagen vilket var irrelevant och hamnade utanför vårt forskningsområde.

I utvärderingen angående avståndsdiagrammen kan man avläsa ur graferna i Figur 16 att de flesta klustren (undantag kluster 3 och kluster 6) blir mer homogena någonstans efter 10% linjen.

Genom Iteration 1 förvärvade vi kunskap om klusteralgoritmens beteenden vid olika antal kluster. Även vid sju kluster som ska vara det mest optimala antalet, blev det svårt att avläsa alla graferna oproblematiskt. Genom Figur 16, kan man avläsa att skärningspunkten mellan klustrens avstånd från centroiden och 10% linjen för kluster 3 och kluster 6 skiljer sig mycket från övriga kluster. Baserat på avståndet för kluster 3 och kluster 6 valde vi att exekvera ytterligare en iteration för att bryta ner båda klustren i mindre kluster för att möjligtvis identifiera mer homogena kluster.

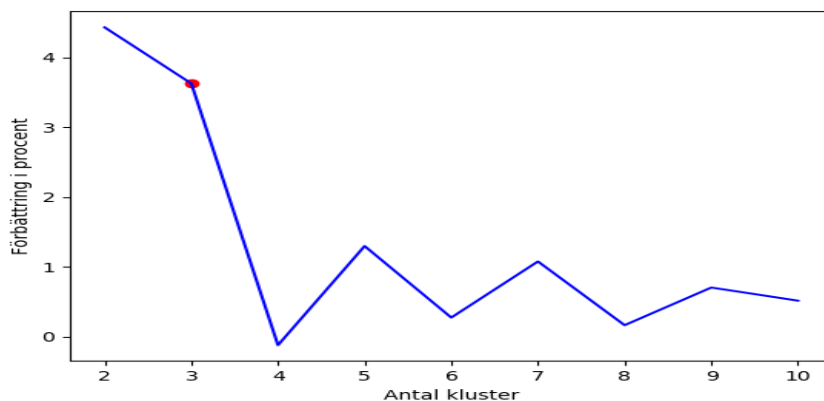
Undantag i vår klusteranalys är kluster 4, då datapunkterna som hamnade i detta kluster registrerades under en period där cykelräknaren var ur funktion.

4.2 Iteration 2

Iteration 2 är ett klusteranalys experiment av en iterativ form likt Iteration 1. Syftet med Iteration 2 var att bryta ner kluster 3 och kluster 6 till mindre kluster med målet att få dessa icke homogena kluster till mer homogena och analyserbara.

4.2.1 Val av antal kluster

Från armbågsmetoden för val av antal kluster i Figur 17 kan man avläsa att grafen förbättrar sig vid punkterna 3, 5, 7 och 9. Vi ansåg att 3 kluster är det mest optimala för denna nedbrytning av kluster 3, eftersom andelen av variansen vid denna punkt utgör störst förbättring. För att förstärka vårt antagande så beräknade vi medelavståndet till centroiden per kluster och resultaten visar (se Tabell 2) att värdena sjunker dramatiskt fram till tre kluster och stabiliserar sig därefter.



Figur 17: Tillämpning av armbågsmetoden i Iteration 2 för kluster 3

Medelavståndet till centroiden / kluster	
Antal kluster (x)	Längdenheter
1	5.87
2	2.94
3	1.96
4	1.47
5	1.17
6	1.00
7	0.84
8	0.74
9	0.65
10	0.59

Tabell 2: Medelavståndet till centroiden per kluster

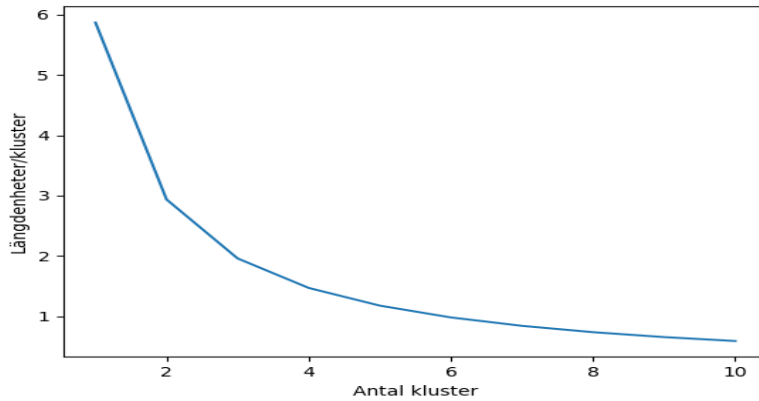
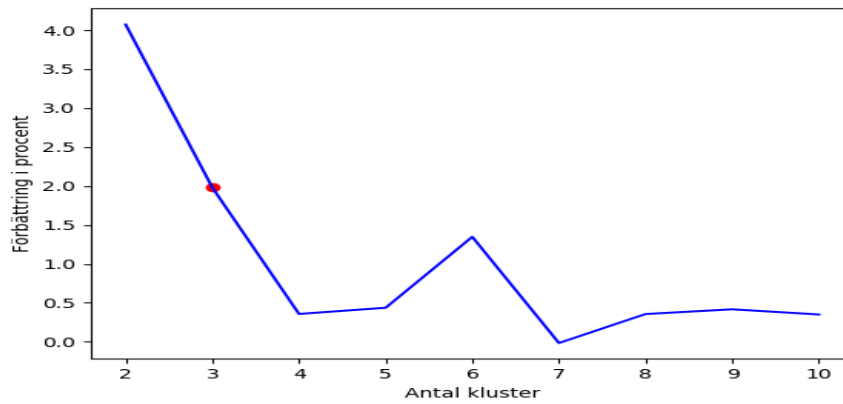


Figure 18: Tillämpning av armbågsmetoden av värdena i Tabell 2

Från armbågsmetoden för val av antal kluster i Figur 19 kan man avläsa att grafen förbättrar sig vid punkterna 3 och 6. Vi ansåg att 3 kluster är det mest optimala för denna nedbrytning av kluster 6, eftersom andelen av variansen vid denna punkt utgör störst förbättring. För att förstärka vårt antagande så beräknade vi medelavståndet till centroiden per kluster och resultaten visar (se Tabell 3) att värdena sjunker dramatiskt fram till tre kluster och stabiliserar sig därefter.



Figur 19: Tillämpning av armbågsmetoden i Iteration 2 för kluster 6

Medelavståndet till centroiden / kluster	
Antal kluster (x)	Längdenheter
1	13.45
2	7.00
3	4.76
4	3.58
5	2.88
6	2.43
7	2.08
8	1.83
9	1.63
10	1.47

Tabell 3: Medelavståndet till centroiden per kluster

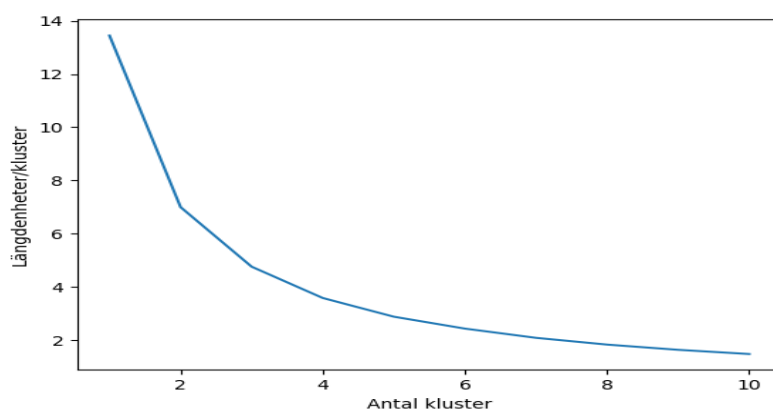
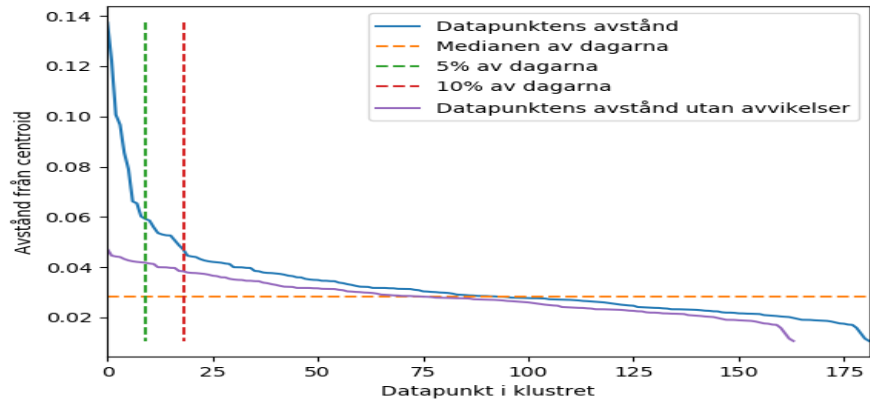


Figure 20: Tillämpning av armbågsmetoden av värdena i Tabell 3

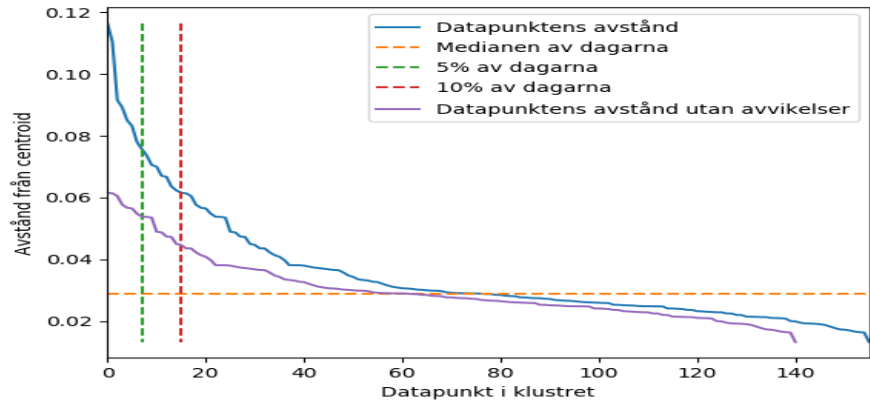
4.2.2 Iteration 2 - Avståndsdiagram

I avståndsdiagrammen i Figur 21 - 26 illustrerar vi datapunkternas Euklidiska distans från respektive centroid för de sex klustren som vi genererat i iteration 2 av vår klusteranalys. Varje avståndsdiagram innehåller fem olika linjer där linjernas färger har olika tolkningar (undantag kluster 4 då klustret innehåller för få datapunkter för att beräkna ut 10% avvikelser):

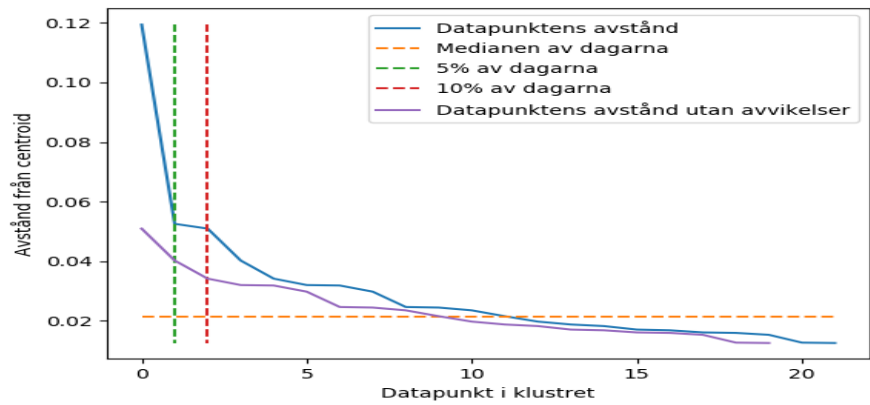
- Blåa linjen motsvarar datapunkternas avstånd från centroiden och är sorterad från störst till minst.
- Gula linjen motsvarar medianen av datapunkterna.
- Gröna linjen motsvarar 5% av datapunkterna med de största avstånden från centroiden.
- Röda linjen motsvarar 10% av datapunkterna med de största avstånden från centroiden.
- Lila linjen motsvarar den blåa linjen utan de eliminerade 10% datapunkterna (avvikelsena) med de största avstånden från centroiden.



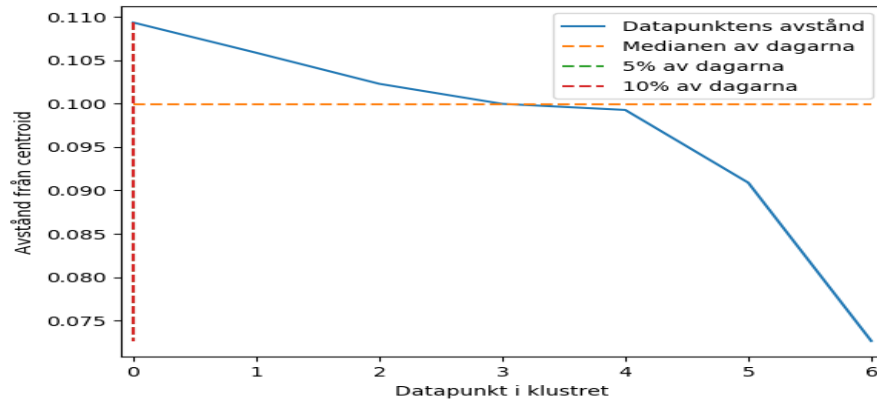
Figur 21: Avståndsdiagram för kluster 3.1



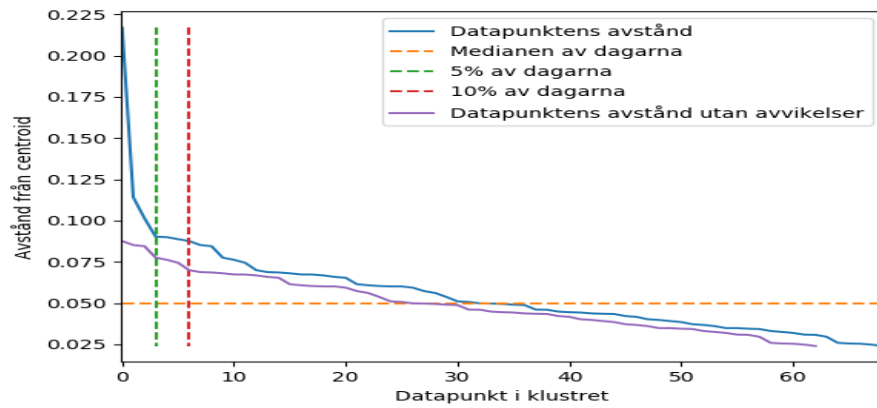
Figur 22: Avståndsdiagram för kluster 3.2



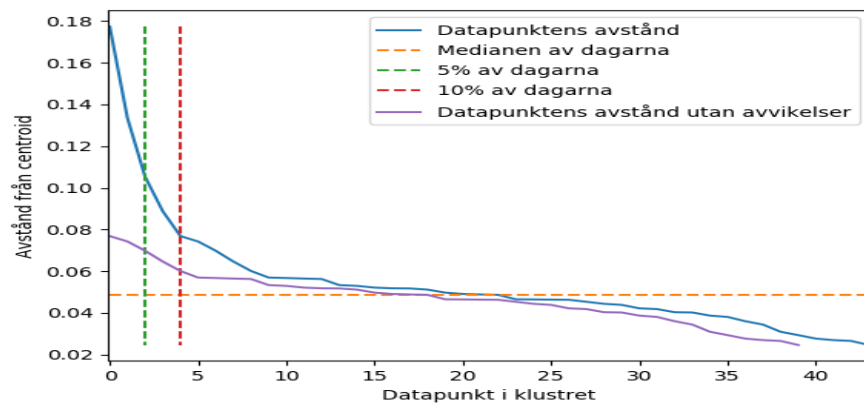
Figur 23: Avståndsdiagram för kluster 3.3



Figur 24: Avståndsdigram för kluster 6.1



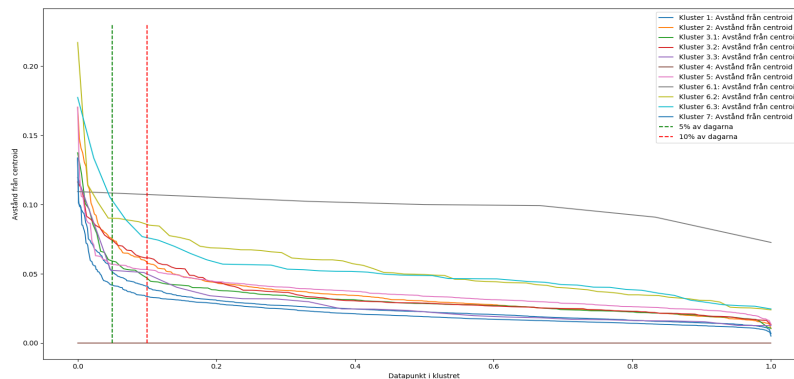
Figur 25: Avståndsdigram för kluster 6.2



Figur 26: Avståndsdigram för kluster 6.3

Avståndsdigrammet i Figur 27 är en sammanställning från avståndsdigrammen för kluster 1, 2, 4, 5, 7 och de nedbrutna klustren 3.1, 3.2, 3.3, 6.1, 6.2, 6.3. Linjerna motsvarar

varje avståndsdiagramms blåa linje (datapunkternas avstånd från respektive centroid).



Figur 27: Sammanställning av avståndsdiagrammen

4.2.3 Analys av Iteration 2

I experimentet angående, Val av antal kluster för kluster 3 och kluster 6 fick vi resultatet på 3 kluster vardera. Detta kan man avläsa ur grafen i Figur 27, där de nya nedbrutna klustren har blivit mer homogent, undantag kluster 6.1.

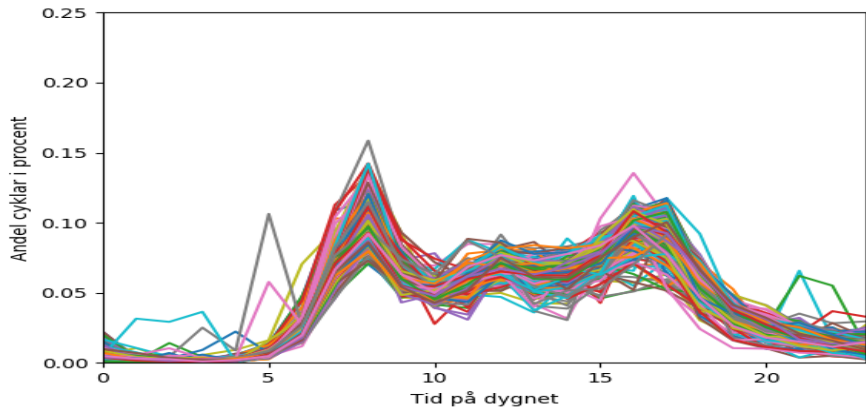
Genom Iteration 2 kan man nu avläsa ur graferna att klustrens avstånd har sjunkit (undantag Figur 24), detta har gett oss värdefull information för att kunna dra en mer noggrann slutsats. Vårt preliminära antagande på 10% kan nu bestyrkas. Vid eliminering av de 10% datapunkterna med längsta avstånden från centroiderna, kan man avläsa att klustren blivit mer homogena jämfört med Iteration 1. Efter Iteration 2 kunde vi även dra slutsatsen att de eliminerade datapunkterna är avvikande datapunkter med hjälp av bakgrundsundersökning.

4.3 Analys av kluster

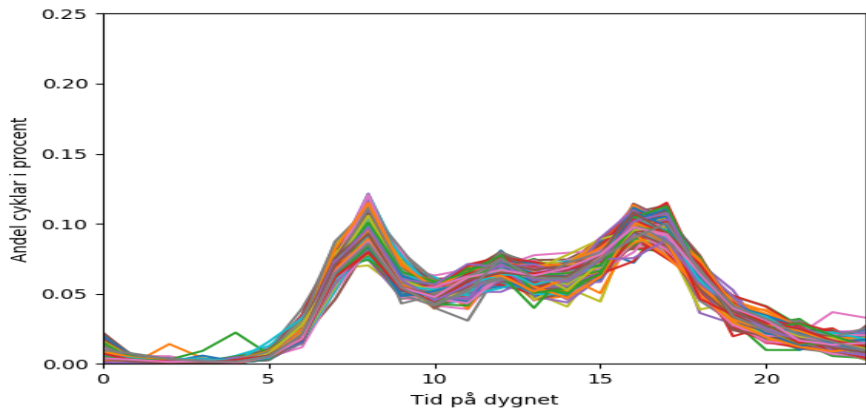
I detta avsnittet presenteras grafer och tabeller som utgör resultatet som har sammanställts från de föregående avsnitten 4.1 och 4.2. Varje delavsnitt nedan omfattar ett kluster som omgrupperas till tre grafer; Med, utan och endast avvikande datapunkter som utgör 10% av datapunkterna med störst avstånd från klustrets centroid. Varje kluster motsvarar ett cykelflöde och varje linje omfattar en dag. För varje kluster sammanställer vi också vilka typer av dagar som har hamnat i klustret samt faktorerna till klustrets avvikande dagar.

4.4 Resultat av Kluster 1

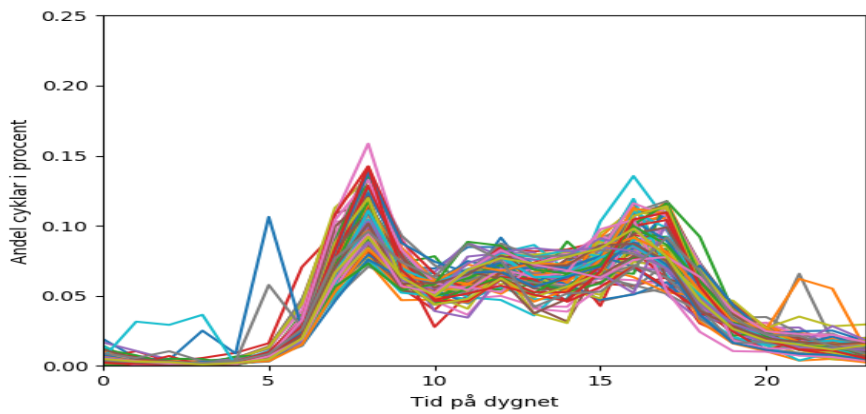
Resultat från kluster 1 visar på hur cykelflödet vid Kaptensgatan i Malmö ser ut på vardagar med vissa avvikande dagar. Faktorerna som vi har funnit med hjälp av bakgrundsundersökning till de avvikande dagarna har varit: större evenemang såsom MECA Raceway, MoneyBrother, Britney Spears och Big Slap, Malmö festivaler, lov dagar, nederbörd samt röda dagar såsom kvalborg & valborg.



Figur 28: Kluster 1 med avvikande punkter



Figur 29: Kluster 1 utan avvikande punkter



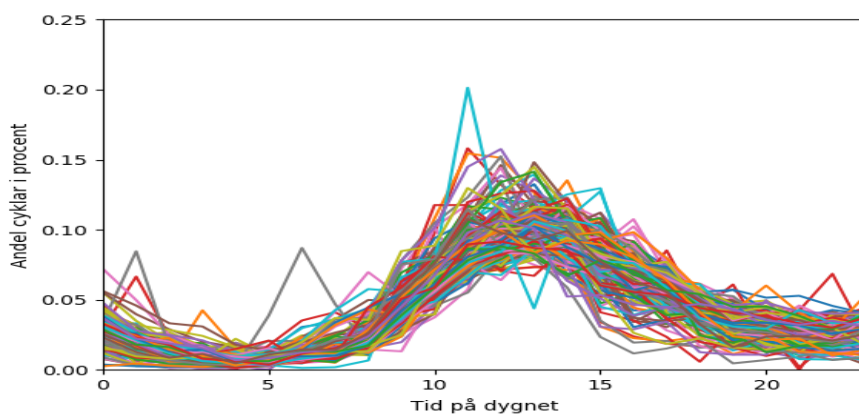
Figur 30: Kluster 1 med endast avvikande punkter

Kluster	1
Antal måndagar:	154
Antal tisdagar:	143
Antal onsdagar:	147
Antal torsdagar:	146
Antal fredagar:	187
Antal lördagar:	0
Antal söndagar:	0
Antal dagar i klustret:	777
Antal avvikande punkter:	77

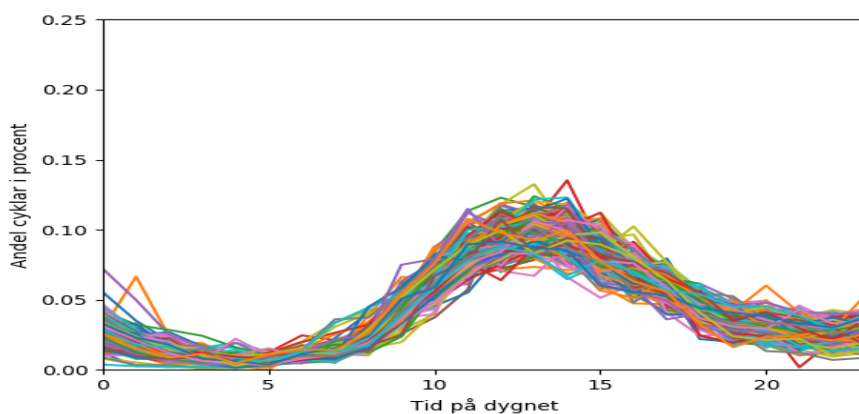
Tabell 4: Sammanställning av kluster 1

4.5 Resultat av Kluster 2

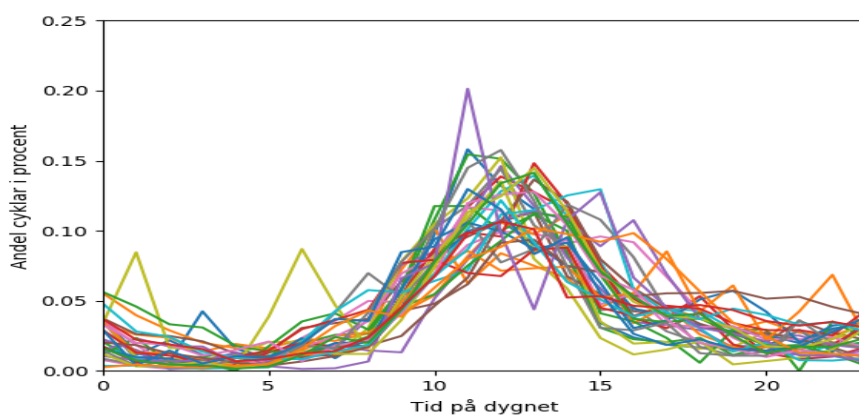
Resultat från kluster 2 visar på hur cykelflödet vid Kaptensgatan i Malmö ser ut på lördagar med vissa avvikande dagar. Faktorerna som vi har funnit med hjälp av bakgrundsundersökning till de avvikande dagarna har varit: röda dagar såsom Julafton, dagar innan röda dagar samt nederbörd.



Figur 31: Kluster 2 med avvikande punkter



Figur 32: Kluster 2 utan avvikande punkter



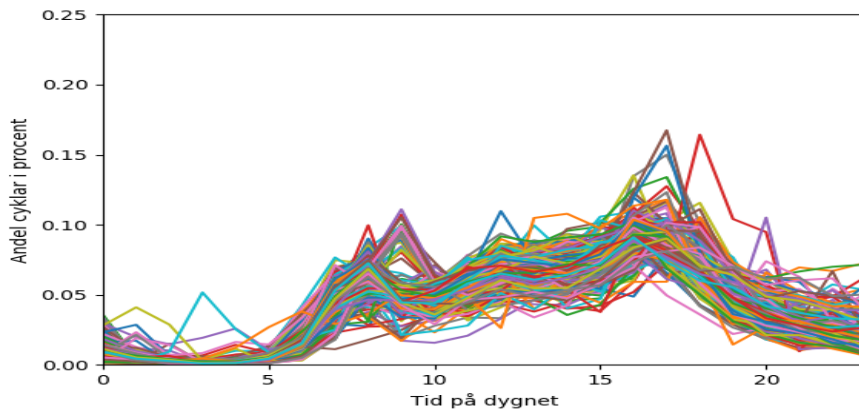
Figur 33: Kluster 2 med endast avvikande punkter

Kluster	2
Antal måndagar:	9
Antal tisdagar:	5
Antal onsdagar:	5
Antal torsdagar:	3
Antal fredagar:	17
Antal lördagar:	362
Antal söndagar:	15
Antal dagar i klustret:	416
Antal avvikande punkter:	41

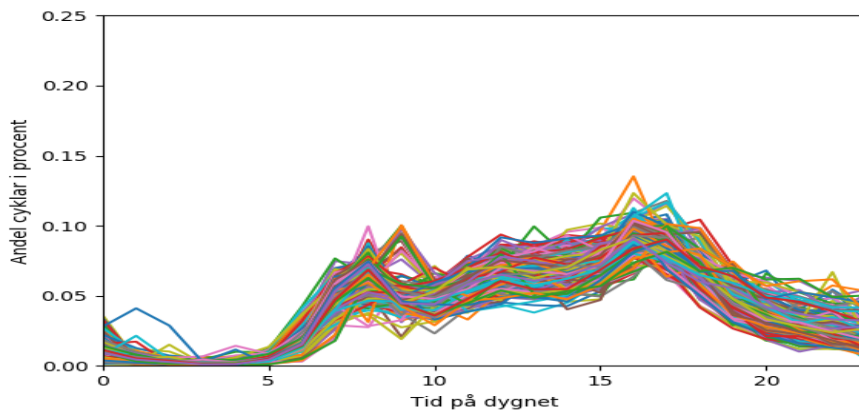
Tabell 5: Sammanställning av kluster 2

4.6 Resultat av Kluster 3

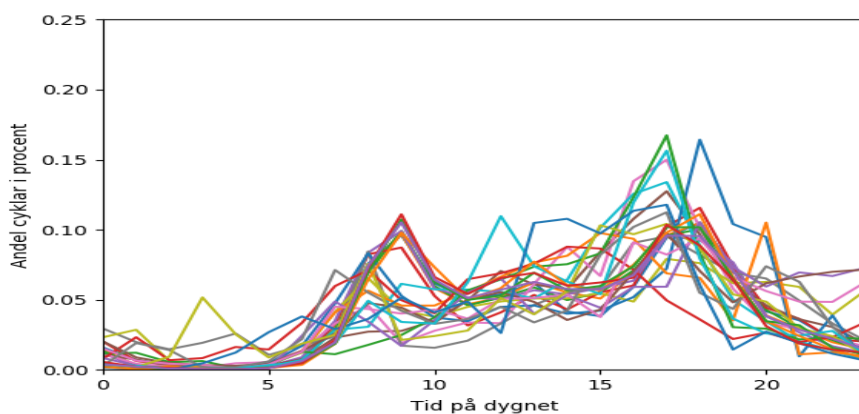
Resultat från kluster 3 visar på hur cykelflödet vid Kaptensgatan i Malmö ser ut på fredagar med vissa avvikande dagar. Faktorerna som vi har funnit med hjälp av bakgrundsundersökning till de avvikande dagarna har varit: lovperioder (sport-, påsk-, sommar-, höst- & jullov).



Figur 34: Kluster 3 med avvikande punkter



Figur 35: Kluster 3 utan avvikande punkter

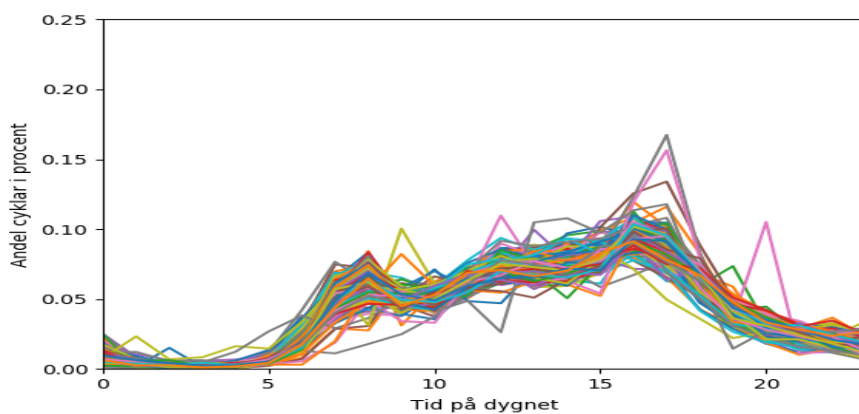


Figur 36: Kluster 3 med endast avvikande punkter

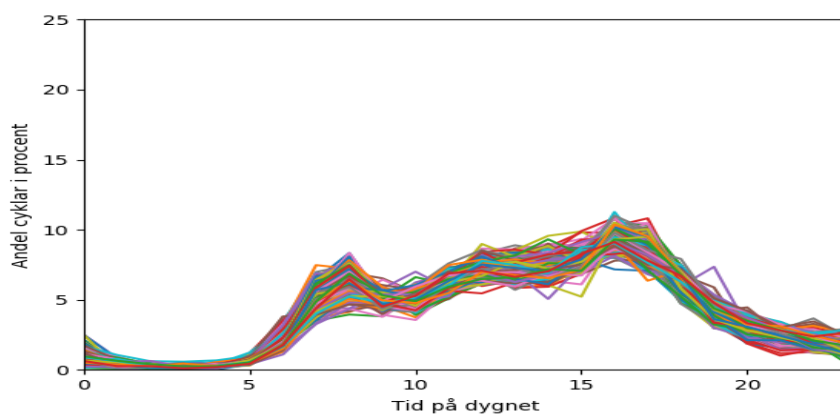
Kluster	3
Antal måndagar:	57
Antal tisdagar:	65
Antal onsdagar:	61
Antal torsdagar:	74
Antal fredagar:	102
Antal lördagar:	0
Antal söndagar:	1
Antal dagar i klustret:	360
Antal avvikande punkter:	36

Tabell 6: Sammanställning av kluster 3

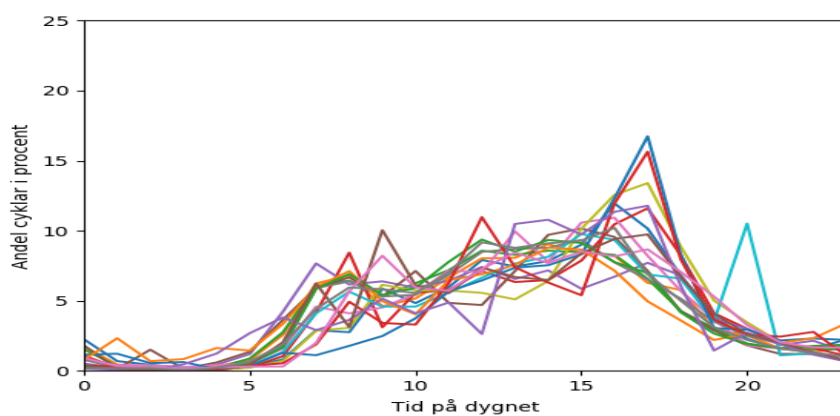
4.6.1 Resultat av Kluster 3.1



Figur 37: Kluster 3.1 med avvikande punkter



Figur 38: Kluster 3.1 utan avvikande punkter

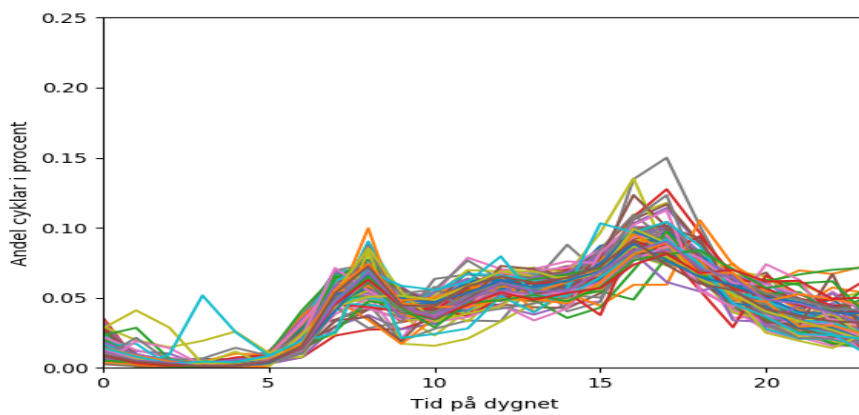


Figur 39: Kluster 3.1 med endast avvikande punkter

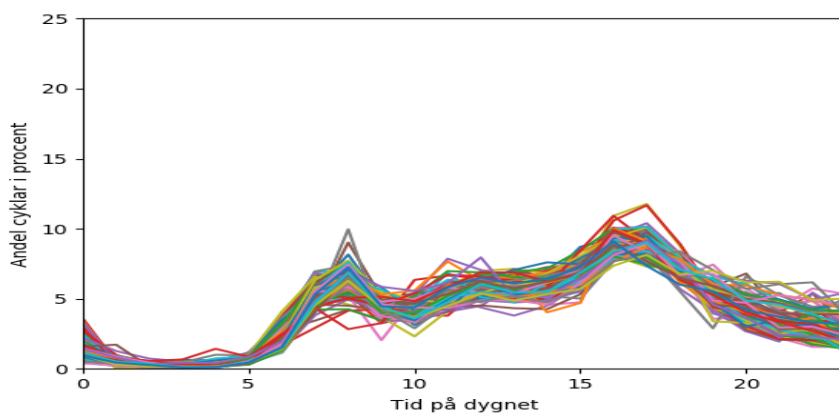
Kluster	3.1
Antal måndagar:	34
Antal tisdagar:	32
Antal onsdagar:	17
Antal torsdagar:	36
Antal fredagar:	63
Antal lördagar:	0
Antal söndagar:	0
Antal dagar i klustret:	182
Antal avvikande punkter:	18

Tabell 7: Sammanställning av kluster 3.1

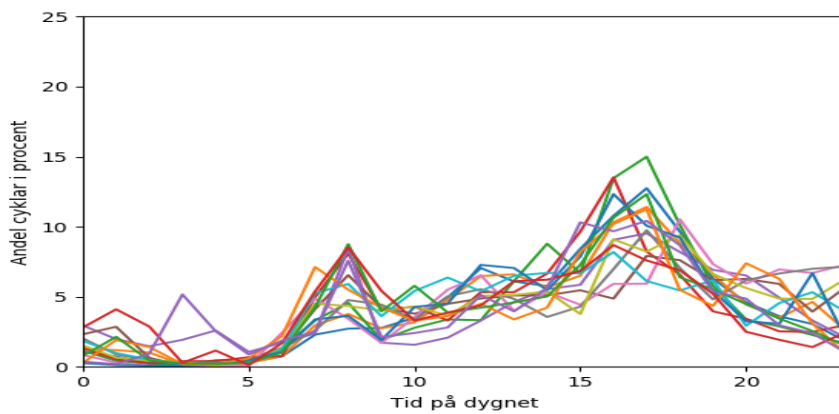
4.6.2 Resultat av Kluster 3.2



Figur 40: Kluster 3.2 med avvikande punkter



Figur 41: Kluster 3.2 utan avvikande punkter

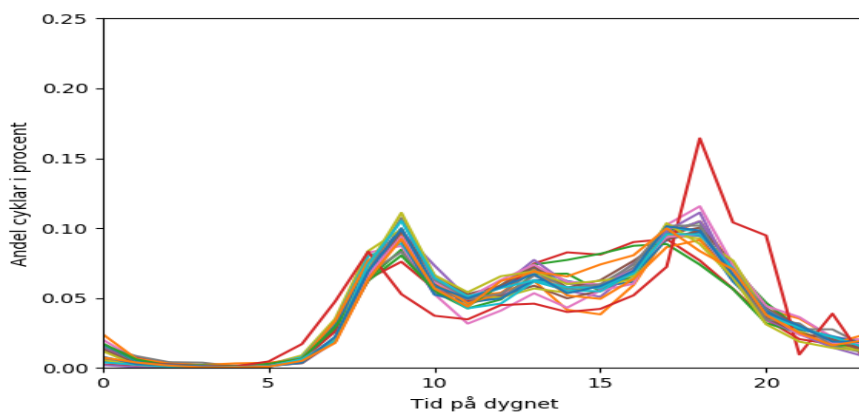


Figur 42: Kluster 3.2 med endast avvikande punkter

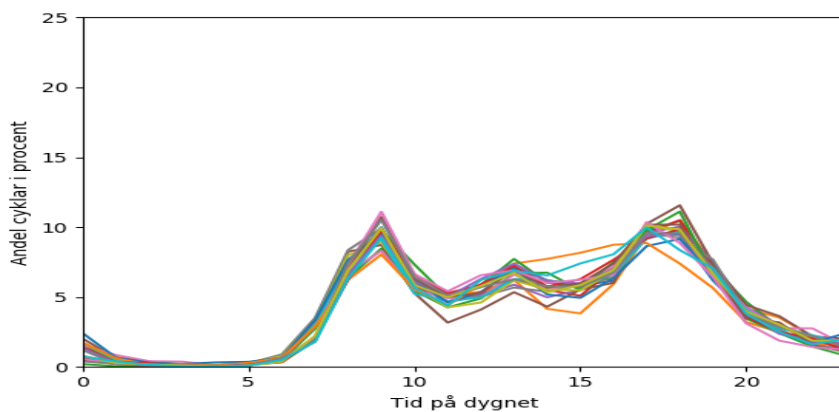
Kluster	3.2
Antal måndagar:	18
Antal tisdagar:	28
Antal onsdagar:	39
Antal torsdagar:	34
Antal fredagar:	36
Antal lördagar:	0
Antal söndagar:	1
Antal dagar i klustret:	156
Antal avvikande punkter:	15

Tabell 8: Sammanställning av kluster 3.2

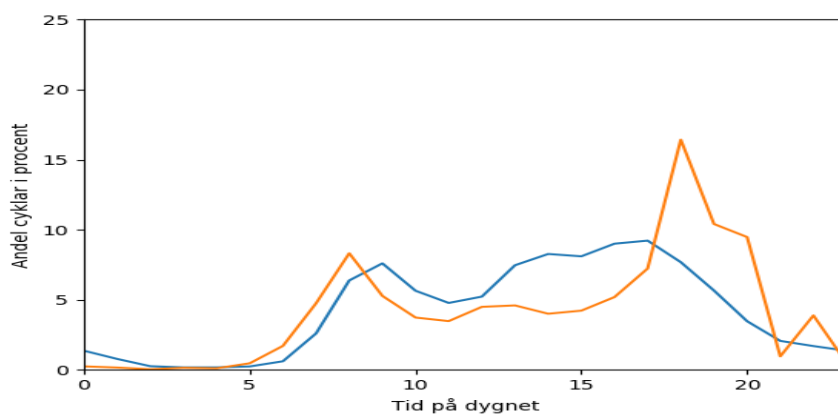
4.6.3 Resultat av Kluster 3.3



Figur 43: Kluster 3.3 med avvikande punkter



Figur 44: Kluster 3.3 utan avvikande punkter



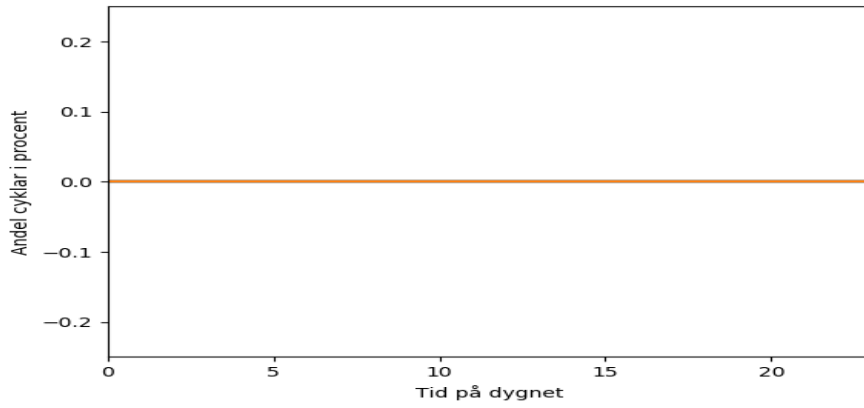
Figur 45: Kluster 3.3 med endast avvikande punkter

Kluster	3.3
Antal måndagar:	5
Antal tisdagar:	5
Antal onsdagar:	5
Antal torsdagar:	4
Antal fredagar:	3
Antal lördagar:	0
Antal söndagar:	0
Antal dagar i klustret:	22
Antal avvikande punkter:	2

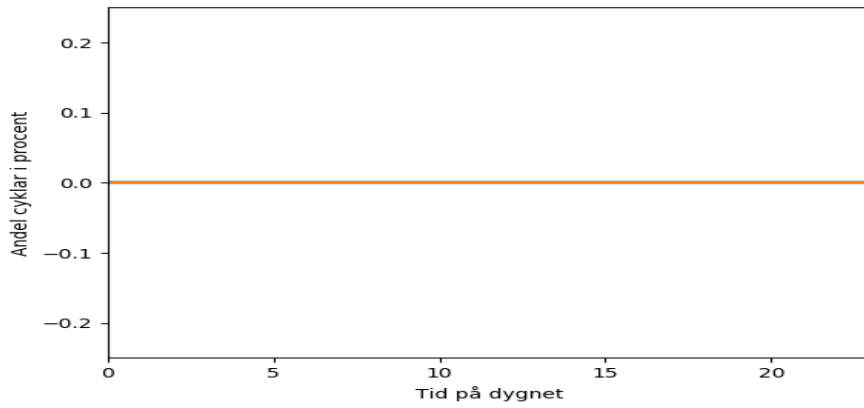
Tabell 9: Sammanställning av kluster 3.3

4.7 Resultat av Kluster 4

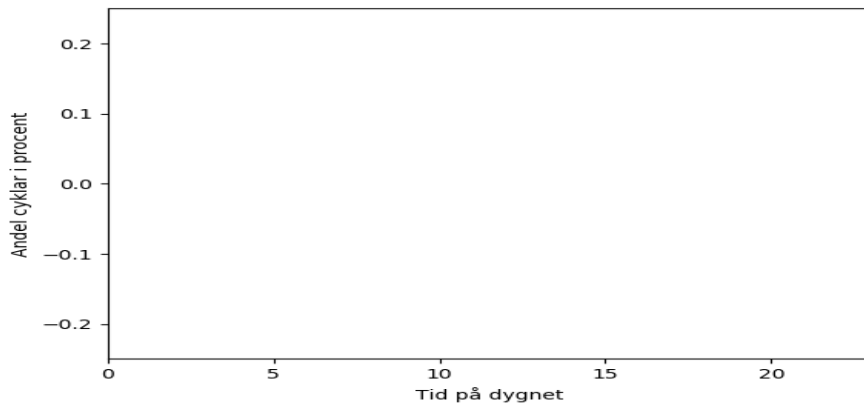
Resultat från kluster 4 visar på hur cykelflödet på Kaptensgatan i Malmö ser ut när cykelräknaren är ur funktion. Faktor: cykelräknaren varit ur funktion.



Figur 46: Kluster 4 med avvikande punkter



Figur 47: Kluster 4 utan avvikande punkter



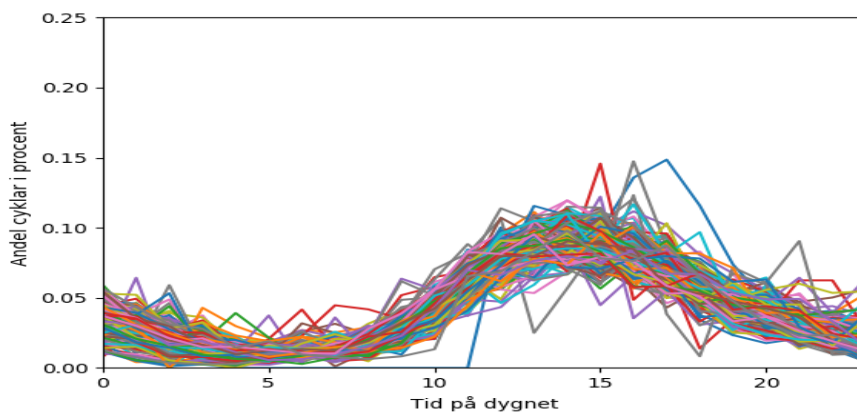
Figur 48: Kluster 4 med endast avvikande punkter

Kluster	4
Antal måndagar:	4
Antal tisdagar:	3
Antal onsdagar:	3
Antal torsdagar:	3
Antal fredagar:	3
Antal lördagar:	3
Antal söndagar:	3
Antal dagar i klustret:	22
Antal avvikande punkter:	0

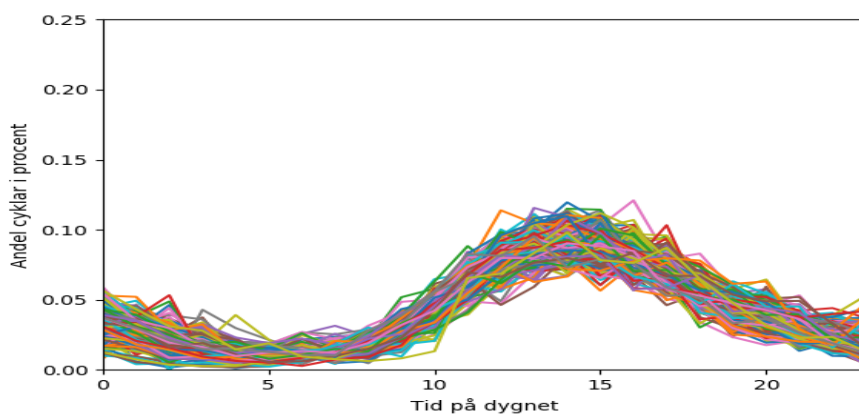
Tabell 10: Sammanställning av kluster 4

4.8 Resultat av Kluster 5

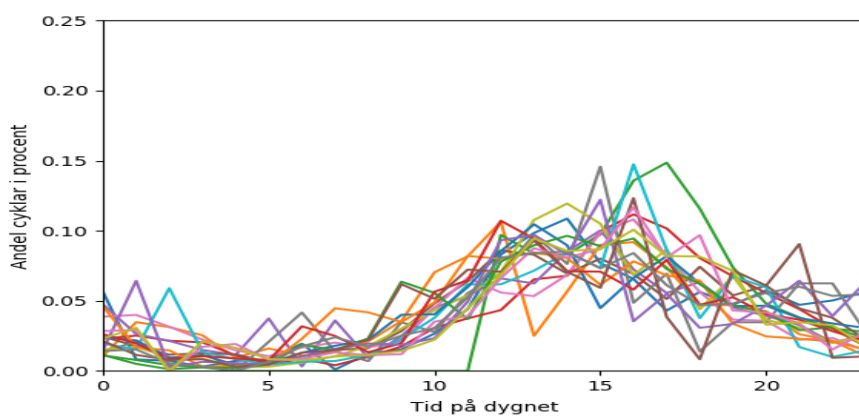
Resultat från kluster 5 visar på hur cykelflödet vid Kaptensgatan i Malmö ser ut på söndagar med vissa avvikande dagar. Faktorerna som vi har funnit med hjälp av bakgrundsundersökning till de avvikande dagarna har varit: röda dagar såsom Juldagen & annandag påsk, fotbollsmatcher samt nederbörd



Figur 49: Kluster 5 med avvikande punkter



Figur 50: Kluster 5 utan avvikande punkter



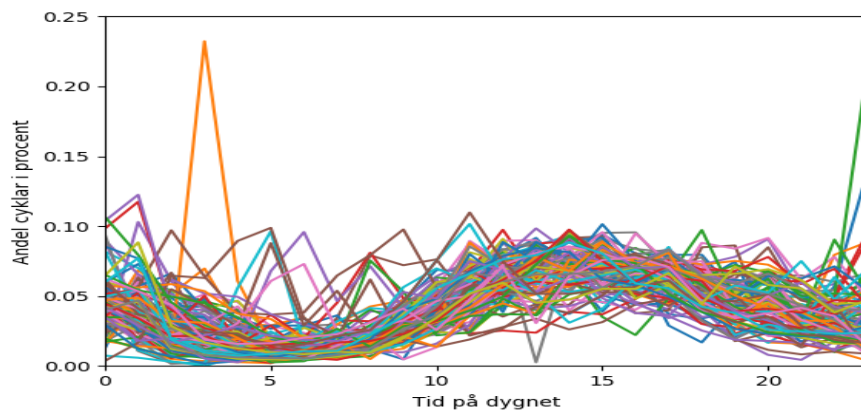
Figur 51: Kluster 5 med endast avvikande punkter

Kluster	5
Antal måndagar:	8
Antal tisdagar:	3
Antal onsdagar:	5
Antal torsdagar:	10
Antal fredagar:	9
Antal lördagar:	25
Antal söndagar:	328
Antal dagar i klustret:	388
Antal avvikande punkter:	38

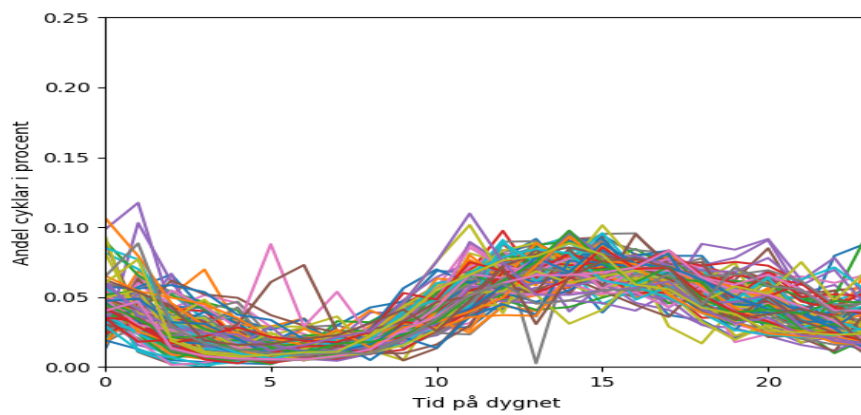
Tabell 11: Sammanställning av kluster 5

4.9 Resultat av Kluster 6

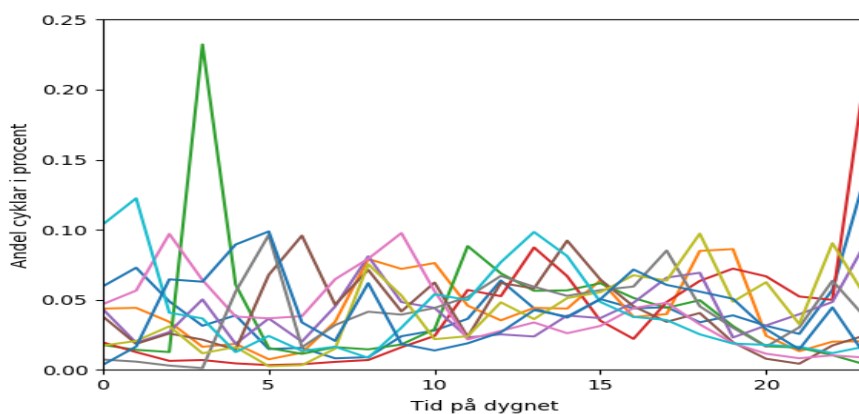
Resultat från kluster 6 visar på hur cykelflödet vid Kaptensgatan i Malmö ser ut på helger med vissa avvikande dagar. Faktorerna som vi har funnit med hjälp av bakgrundsundersökning till de avvikande dagarna har varit: buggar i cykelräknaren, Malmö festivaler, fotbollsmatcher samt konserter med bland annat Carola.



Figur 52: Kluster 6 med avvikande punkter



Figur 53: Kluster 6 utan avvikande punkter

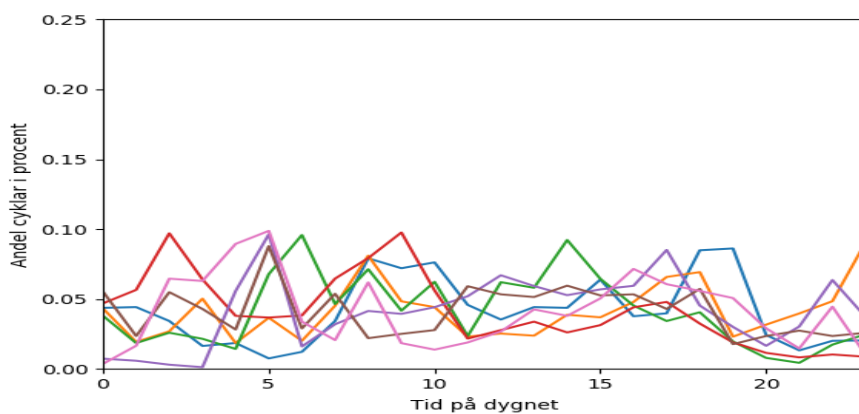


Figur 54: Kluster 6 med endast avvikande punkter

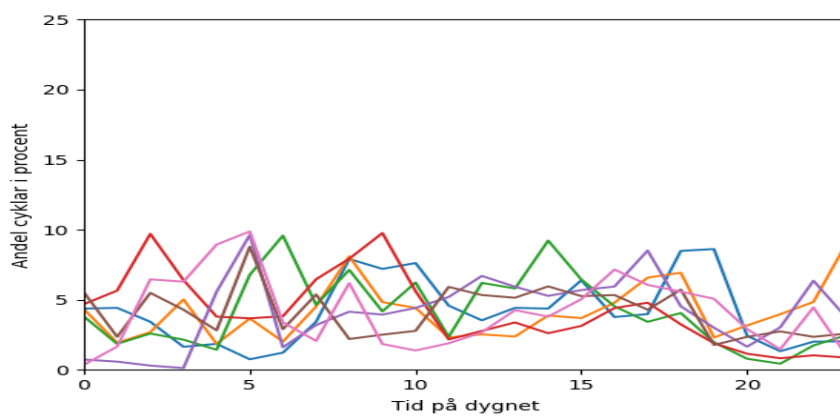
Kluster	6
Antal måndagar:	3
Antal tisdagar:	6
Antal onsdagar:	6
Antal torsdagar:	5
Antal fredagar:	5
Antal lördagar:	26
Antal söndagar:	69
Antal dagar i klustret:	120
Antal avvikande punkter:	12

Tabell 12: Sammanställning av kluster 6

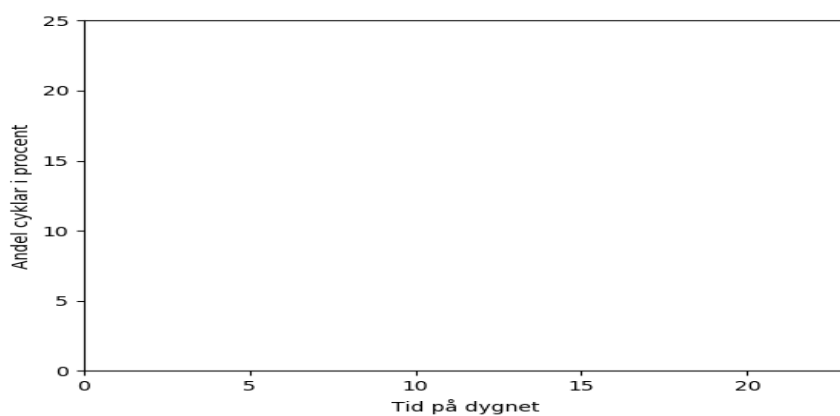
4.9.1 Resultat av Kluster 6.1



Figur 55: Kluster 6.1 med avvikande punkter



Figur 56: Kluster 6.1 utan avvikande punkter

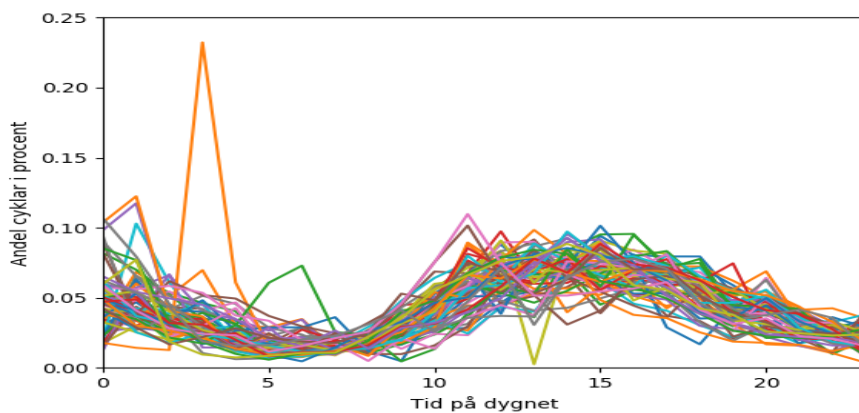


Figur 57: Kluster 6.1 med endast avvikande punkter

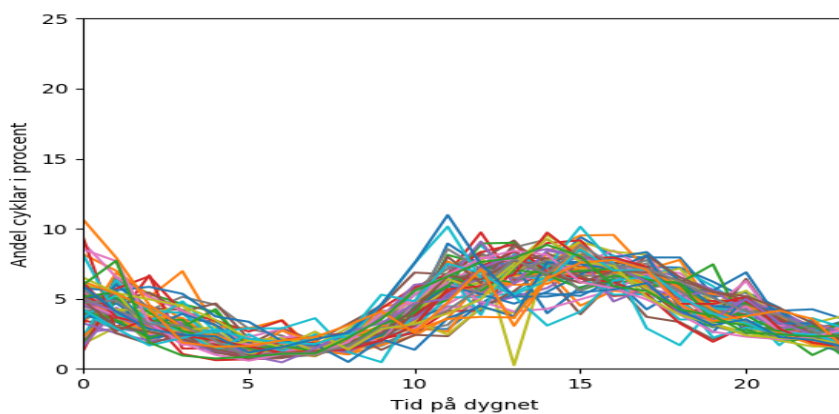
Kluster	6.1
Antal måndagar:	1
Antal tisdagar:	1
Antal onsdagar:	3
Antal torsdagar:	0
Antal fredagar:	0
Antal lördagar:	1
Antal söndagar:	1
Antal dagar i klustret:	7
Antal avvikande punkter:	0

Tabell 13: Sammanställning av kluster 6.1.

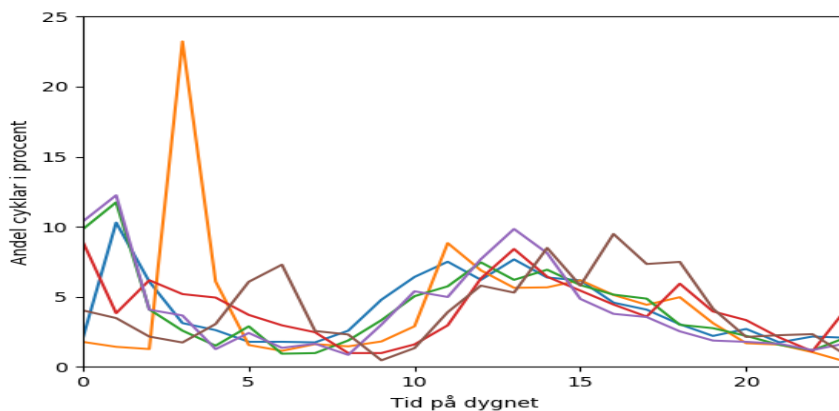
4.9.2 Resultat av Kluster 6.2



Figur 58: Kluster 6.2 med avvikande punkter



Figur 59: Kluster 6.2 utan avvikande punkter

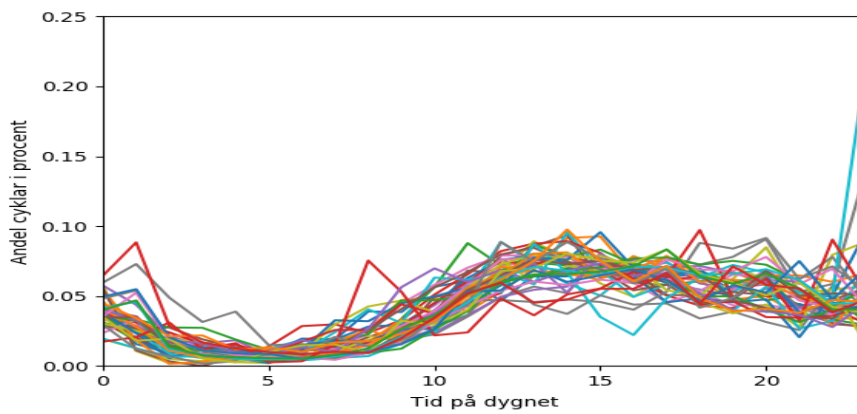


Figur 60: Kluster 6.2 med endast avvikande punkter

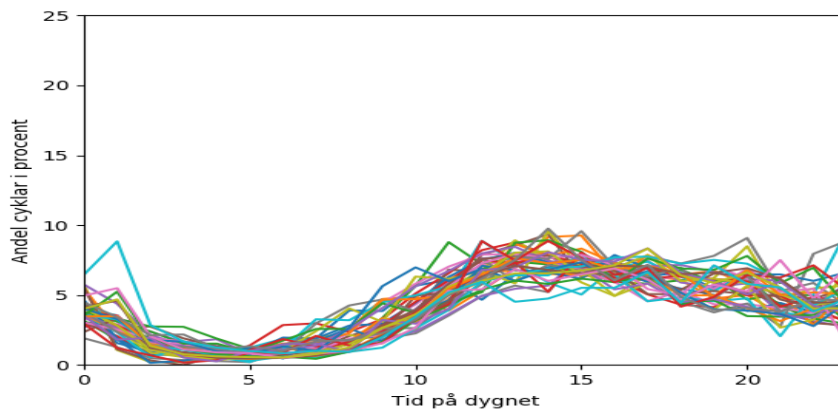
Kluster	6.2
Antal måndagar:	2
Antal tisdagar:	2
Antal onsdagar:	1
Antal torsdagar:	1
Antal fredagar:	2
Antal lördagar:	7
Antal söndagar:	54
Antal dagar i klustret:	69
Antal avvikande punkter:	6

Tabell 14: Sammanställning av kluster 6.2

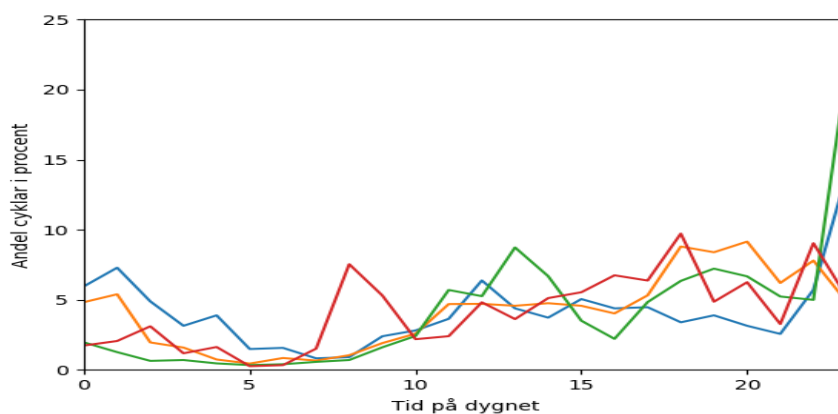
4.9.3 Resultat av Kluster 6.3



Figur 61: Kluster 6.3 med avvikande punkter



Figur 62: Kluster 6.3 utan avvikande punkter



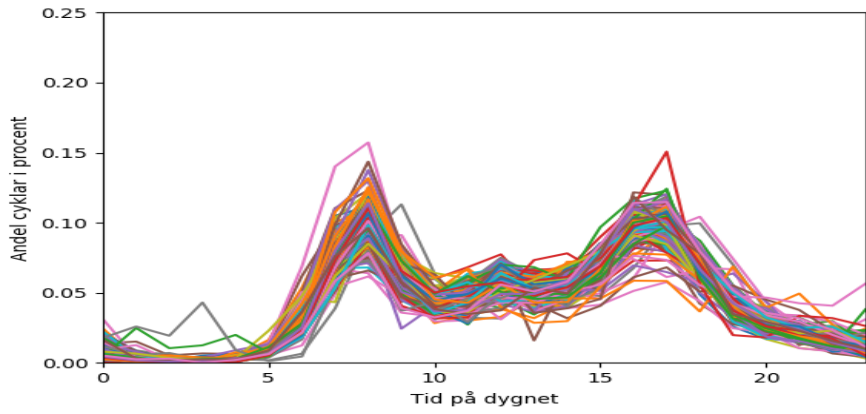
Figur 63: Kluster 6.3 med endast avvikande punkter

Kluster	6.3
Antal måndagar:	0
Antal tisdagar:	3
Antal onsdagar:	2
Antal torsdagar:	4
Antal fredagar:	3
Antal lördagar:	18
Antal söndagar:	14
Antal dagar i klustret:	44
Antal avvikande punkter:	4

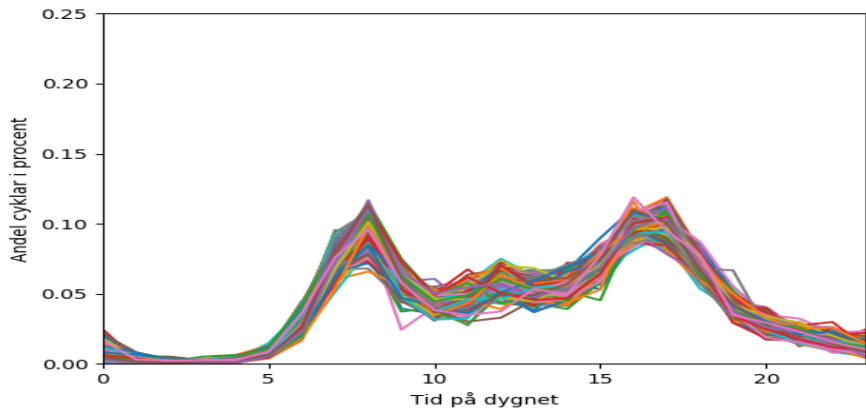
Tabell 15: Sammanställning av kluster 6.3

4.10 Resultat av Kluster 7

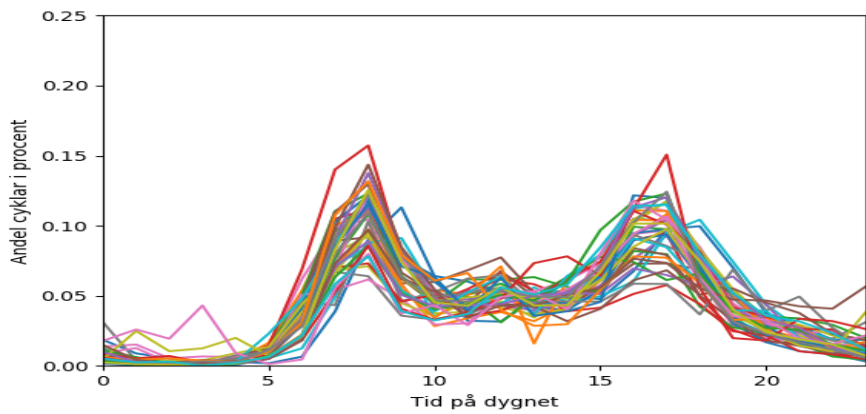
Resultat från kluster 7 visar på hur cykelflödet vid Kaptensgatan i Malmö ser ut på vardagar med vissa avvikande dagar. Faktorerna som vi har funnit med hjälp av bakgrundsundersökning till de avvikande dagarna har varit: demonstrationer, fotbollsmatcher, festivaler, lövdagar samt nederbörd.



Figur 64: Kluster 7 med avvikande punkter



Figur 65: Kluster 7 utan avvikande punkter



Figur 66: Kluster 7 med endast avvikande punkter

Kluster	7
Antal måndagar:	180
Antal tisdagar:	190
Antal onsdagar:	189
Antal torsdagar:	175
Antal fredagar:	93
Antal lördagar:	0
Antal söndagar:	0
Antal dagar i klustret:	827
Antal avvikande punkter:	82

Tabell 16: Sammanställning av kluster 7

5 Diskussion

I detta avsnitt diskuteras detaljer och tolkningar från Avsnitt 4 Resultat och analys. Strukturen är indelad i underrubriker som är relaterade till våra forskningsfrågor. I diskussionen är det underförstått att vi har tagit hänsyn till våra avgränsningar i Avsnitt 1.6.

5.1 Användning av klusteranalys

Tack vare vår litteraturstudie fick vi viktiga förkunskaper om klusteranalys och maskininlärning. För val och uppsättning av klusteranalys tog vi hänsyn till principerna inom K-means. För att gruppera kluster och få ut bästa resultat använde vi oss av samma princip som Archana och Prateek [11] genom att använda samma datamängd (normaliserad) under hela forskningen. Gruppering av dagar underlättades sedan med hjälp av armbågsmetoden där vi använde oss av ett avståndsmått för att utvärdera hur väl de grupperande dagarna är [9].

5.2 Optimalt antal kluster och definition av en avvikande dag

Med hjälp av K-means och ett avståndsmått kunde vi efter Iteration 1 avgöra att klusteralgoritmen producerade bäst resultat när vi använde oss av 7 kluster. Detta var tillräckligt för att besvara underfrågan FF1.1, men för att besvara underfrågan FF1.2 var vi tvugna att exekvera ytterligare en iteration där vi bröt ner de kompakta klustren (3) och (6) till 6st mindre kluster för att möjliggöra en djupare analys. Med de nedbrutna klustren kunde vi sedan se att skärningspunkten mellan klustrens avstånd från centroiden till 10% linjen hamnade i en mer acceptabel avstånd, vilket stärker vår slutsats. Genom att vi eliminerade 10% av datapunkterna med de längsta avstånden från centroiden, kan vi sedan avläsa ur alla avståndsdigram (se Figur 9-15 och 21-26 (undantag 24)) att den lila grafen "datapunktens avstånd utan avvikelser" är mer homogen och därmed kunna dra en slutsats att de datapunkterna hör hemma i klustret.

Diagrammen i Figur 29, 32, 35, 38, 41, 44, 37, 50, 53, 56, 59, 62 och 65 i resultatdelen styrker vår slutsats. Där kan man se att datapunkterna i respektive kluster har på något sätt ett samband. Så hur kan sambandet för cykelflöden i respektive kluster utvärderas? Genom bakgrundsundersökning av datapunkterna i respektive kluster har vi kunnat identifiera sambandet inom vår avgränsning. Sambandet vi har för klustren är att vanliga vardagar hamnade i kluster 1, lördagar med speciella röda dagar som julafton hamnade i kluster 2, fredagar under lovperioder (sports-, påsk-, sommar-, höst- & jullov) hamnade i kluster 3, datapunkterna som registrerades när cykelräknaren var ur funktion hamnade i kluster 4, söndagar med speciella röda dagar som Juldagen och fotbollsmatcher hamnade i kluster 5, helger med buggar i cykelräknaren hamnade i kluster 6 och vardagar med större evenemang som demonstrationer, fotbollsmatcher och festivaler hamnade i kluster 7.

5.3 Faktorer till avvikelser

De 10% avvikande datapunkterna har vi sedan kunnat identifiera faktorerna bakom med hjälp av djupare bakgrundsundersökning. Varför hamnade de avvikande datapunkterna i respektive kluster? En faktor som väder (temperatur, snöfall) kan ha påverkat cyklisternas val för dagen som Tsapakis m.fl [2] har påpekat i sin artikel och detta har lett till att

cykelflödet har blivit annorlunda jämfört med den normala cykelflödet för sitt kluster. En annan faktor är, som Mohamed m.fl [5] betonar i sin artikel att missade datavärden på grund av till exempel funktionsfel i cykelräknaren påverkar cykelflödet. Under vissa dagar, vissa timmar har cykelräknaren varit avslagen och missat att registrera värdefulla cyklister vilket har påverkat cykelflödet negativt och vissa andra dagar har räknaren registrerat helt orimliga värden. Andra faktorer som ledde till avvikelser var att cykelflödet förändrades under vissa timmar under en viss dag på grund av till exempel Malmö festivaler, fotbollsmatcher, större evenemang, nederbörd etc.

6 Slutsats och vidare forskning

6.1 Slutsats

Gällande specifikt antal kluster fungerade olika antal kluster olika bra. Det finns både för- och nackdelar med för få och för många kluster. För vår studie ansåg vi att efter Iteration 1 var 7 kluster det mest optimala inom vårt område och avgränsning. Men för att kunna göra en djupare analys av de avvikande datapunkterna ansåg vi att vi behövde bryta ner 2 av de 7 klustren till 6 mindre kluster, vilket gav oss totalt 11 kluster efter Iteration 2.

När det kommer till definition av avvikande datapunkter ansåg vi att med hjälp av ett avståndsdiagram att 10% av datapunkterna som ligger längst ifrån centroiderna är med största sannolikhet de avvikande datapunkterna med störst inverkan på cykelflöden. Även här kan avgränsningen av studien ha en påverkan på resultaten.

Baserat på vårt resultat levererade vår lösning tillräckligt med resultat för att besvara på våra forskningsfrågor. Enligt resultatet noterade vi att med hjälp av K-means, så uppnådde vi en högre nivå av noggrannhet och gruppering av cykelflödesdata genom att vi eliminerade de avvikande datapunkterna och faktorerna som identifierats.

Lösningen var optimal till en viss gräns, därför att avgränsningen av studien kan ha påverkan på resultatet. Om en annan data hade undersökts, till exempel en annan tidsperiod, kan oförutsägbara faktorer som till exempel evenemang, fotbollsmatcher, nederbörd och röda dagar påverka resultatet annorlunda.

6.2 Vidare forskning

För vidare forskning föreslår vi att studera med annan data (tidsperiod) från samma cykelräknare. Vidare forskning inom detta området skulle kunna bidra till ytterligare utvärdering av vår klusteranalys och resultat till att uppnå förbättrade uppskattning av cykelflöden och hur stor påverkan faktorerna vi identifierade har på cyklisterna på kaptensgatan i Malmö stad.

Vi föreslår även att liknande forskning och experiment utförs vid andra cykelräknare för att kunna utvärdera cykelflöden vid andra områden i stan men även i andra städer. Framtida arbete hade kunnat vara att utöka testningen av metoden men i andra städer över andra tidsperioder med andra attribut som till exempel cyklisternas riktning och förhållanden.

Spännande användningsområden för metoden skulle vara i andra städer med andra förutsättningar som sämre eller bättre väder, andra lov dagar och arbetstider, turiststäder, högre befolkningstäthet etc.

7 Referenser

- [1] - P. Oja, I. Vuori och O. Paronen , *"Daily walking and cycling to work: their utility as health-enhancing physical activity"*, *Patient Education and Counseling Volume 33, Supplement 1, 1 April 1998, Pages S87-S94 Transportation Systems (ITSC)*
- [2] - Tsapakis, Ioannis; Cheng, Tao; Bolbol, Adel; *"Impact of weather conditions on macroscopic urban travel times"*, *Journal of Transport Geography 28 (2013) 204–211*
- [3] - Li Fang Xu *"Klusteranalys"*, (2008)
- [4] - Mahdie H, Arash J and Sahar G.M, *"Developing Models for Matching of Short-term and Long-term Data Collection Sites to Improve the Estimation of Average Annual Daily Bicyclists"*, *Conference: 2018 IEEE International Conference on Intelligent Transportation Systems (ITSC)*
- [5] - Mohamed El.E, Ahmed I.M, Khaled N, *discoveries "Estimation of daily bicycle traffic volumes using sparse data"*, *Computers, Enviroment and Urban systems, Volume 54, Pages 195-203, Ain Shams University, 2015 November 5.*
- [6] - Holmgren Johan, Moltubakk, Gabriel, O'Neill, Jody *"Regression-based evaluation of bicycle flow trend estimates"*, *Procedia computer science [1877-0509] Holmgren, Johan yr:2018 vol:130 pg:518 -525*
- [7] - Holmgren Johan, Aspegren Sebastian, Dahlström Jonas *"Prediction of bicycle counter data using regression"*, *Procedia Computer Science, Volume 113 , Pages 502-507. Malmö University 2017*
- [8] - Holmgren Johan, Aspegren Sebastian, Dahlström Jonas *"En jämförelse av maskininlärningsalgoritmer för uppskattning av cykelflöden baserat på cykelbarometer- och väderdata"*, *27 Maj, 2016 Malmö University*
- [9] - Purnima Bholowalia, Arvind Kumar , *"EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN"*, *International Journal of Computer Applications (0975 – 8887) Volume 105 – No. 9, November 2014*
- [10] - Youguo Li, Haiyan Wu, , *"A Clustering Method Based on K-Means Algorithm"*, *Physics Procedia, 2012, Pages 1104-1109*
- [11] - K M Archana Patel, Prateek Thakral, *"The Best Clustering Algorithms in Data Mining"*, *2016 International Conference on Communication and Signal Processing (ICCSP), Melmaruvathur, India. 24 November 2016*
- [12] - Jacobsen, Dag, Ingvar, *"Vad, hur och varför?: Om metodval i företagsekonomi och andra samhällsvetenskapliga ämnen"*, 2002, *Studentlitteratur AB, Lund. s, 210*
- [13] - Farhad Malik, *"Machine Learning Algorithms Comparison"*, (27 Aug, 2018)

- [14] -Steve Easterbrook, Janice Singer, Margaret-Anne Storey, Daniela Damian , "*Selecting Empirical Methods for Software Engineering Research*", (2008)
- [15] - S. Saad Harous, Maryam Al Harmoodi, Hessa Biri, "*A Comparative Study of Clustering Algorithms for Mixed Datasets*", *College Information Technology, UAE University, Al Ain, UAE 29 (April 2019)*
- [16] - Osvaldo Simeone, "*A Very Brief Introduction to Machine Learning With Applications to Communication Systems*", (November 2018)
- [17] - Bryman, Alan "*Samhällsvetenskapliga metoder.2* " uppl. Malmö: Liber, s, 340, (2011)
- [18] - Miljöförvaltningen, *Malmö.se*,
<http://miljobarometern.malmö.se/trafik/cykling/cykeltrafikutveckling/>
(Hämtat Mars, 2019)
- [19] - Sveriges Meteorologiska och Hydrologiska Institut, *SMHI.se*,
<https://opendata.smhi.se/apidocs/> (Hämtat Mars, 2019)