LINKÖPINGS UNIVERSITET

Prognostisering av serviceärenden

En jämförelse av statistiska metoder för tidsseriemodellering av ärendeflödet i hyresfastigheter

Isabella Roos
Leonard Persson Norblad

Avdelningen för Statistik och maskininlärning
Institutionen för datavetenskap
Linköpings universitet
Vårterminen, 2022

Handledare: Josef Wilzén
Examinator: Linda Wänström

Sammanfattning

Atrium Ljungberg är en fastighetsförvaltare med fastigheter runt om i Sverige, bland annat i Stockholm, Malmö och Göteborg. För att upprätthålla hög standard har deras hyresgäster möjlighet att rapportera fel som kan uppstå i fastigheterna. Genom att fördela personalstyrkan efter behovet att åtgärda ärenden kan risken för över- och underbemanning minskas samt hanteringstiden förkortas. Detta ger förutsättningar för både nöjdare hyresgäster och minimerad arbetsbörda för drifttekniker. Statistiska metoder kan användas som beslutsstöd för personalfördelningen genom att prognostisera antalet serviceärenden som kommer uppkomma i framtiden. Syftet var att skapa statistiska modeller som kan prognostisera ärendeflödet de kommande 14 dagarna för fastigheter i Atrium Ljungbergs driftområden.

Datamaterialet som analyseras består av serviceärenden från företagets nio driftområden mellan november 2015 och december 2021. För att undersöka ärendeflödet över tid har ärenden aggregerats på dag och driftområde. Tre typer av statistiska modeller jämfördes för att hitta den mest optimala och de typer som användes var ARIMA, dynamisk regression och XGBoost. Modellerna jämfördes med måttet RMSE, där den modell med lägst RMSE vid flest prognoshorisonter av de 14 prognoshorisonterna ansågs vara den mest lämpliga. För att därefter göra mer konkreta tolkningar av skattningsförmågan hos de bästa modellerna för respektive driftområde användes måttet MAE, som mäter det genomsnittliga felet.

Resultatet visade att RMSE skiljde sig marginellt mellan de olika modellerna för respektive driftområde, vilket innebar att det var något svårt att utse den bästa modellen. De slutsatser som kunde tas visade att XGBoost var bäst i fem av de nio driftområdena och dynamisk regression med två olika typer av indikatorvariabler presterade bäst i två av driftområdena. Modellerna för de övriga två driftområdena var dynamisk regressionsmodell med en typ av indikatorvariabel och en SARIMA-modell.

Abstract

In order to keep a high standard in real estates owned by Atrium Ljungberg, they give their tenants the possiblity to communicate when there is an issue with the property they are renting. By optimising the staff according to the volume of real estate matters, the processing time can be minimised and the company can avoid over- and understaffing. This could result in more satisfied tenants and less workload for the caretakers. With the help of statistical methods the number of real estates matters could be forecasted. The purpose of this thesis was to build statisitcal models that could forecast the number of real estate matters in different operational areas, for the next 14 days.

The data that were used for analysis contained observations between november 2015 and december 2021 for the nine different operational areas. Three different types of models were compared to find the most optimal one for each operational area. The three types that were used was ARIMA, Dynamic regression models and XGBoost. The models were compared with RMSE and the one with the lowest RMSE for most forecast horizons was chosen as the best one. After the nine different models were chosen they were evaluated with MAE which measures the mean absolute error. This measurement is appropriate to use when studying the average number of errors made by the model.

The outcome showed that there were small differences in RMSE between the models within the different operational areas, which made it somewhat difficult to find the most ideal one. The conclusion was that XGBoost preformed best in five of the nine areas, and the dynamic regression model including two types of dummyvariables performed best in two of the areas. In the two remaining areas the dynamic regression model with one type of dummyvariabel and SARIMA model performed best.

Förord

Ett stort tack till Per Karmteg på Atrium Ljungerg som har bistått med data och agerat som viktig kontaktperson.

Tack till Håkan Sahlberg på WSP som har gett oss tillgång till deras databas och hjälp oss med datahämtning.

Stort tack till vår handledare Josef Wilzén som har varit stöttande i processen och tillfört mycket värdefulla kommentarer.

Innehåll

Figurer

Tabeller

1. Inledning

Hyresgäster har höga krav på service och önskar ofta att problem ska lösas snabbt. Detta innebär att många företag behöver effektivisera sina verksamheter, inte bara för att uppfylla sina kunders krav, utan även för att själva spara tid och resurser.

1.1 Bakgrund

Atrium Ljungberg AB är ett svenskt fastighetsbolag som bildades 2006 genom en sammanslagning av de två bolagen LjungbergGruppen och Atrium fastigheter. Företget består i huvudsak av kontors- och handelsfastigheter och har ett fastighetsbestånd som omfattar cirka 1 miljon kvadratmeter i nio olika driftområden runt om i Sverige, bland annat i Stockholm, Uppsala och Malmö. För att säkerställa att alla fastigheter håller hög kvalité kan Atrium Ljungbergs hyresgäster skapa en felanmälan för fastigheterna om något problem skulle uppstå, detta skapar en trygghet för både kunden och företaget då de kontinuerligt kan underhålla fastigheterna. Som en del av bolagets strategiska fundament ”att göra rätt saker på rätt ställen” (Atrium Ljungberg 2022) är det viktigt att kunna tillgodose personal som kan åtgärda ärendena snabbt för att bibehålla god standard och skapa förtroendefulla relationer till hyresgästerna. Genom att utveckla hanteringen för ärenden kan hanteringstiden förkortas och personalstyrkan kan optimeras, vilket skapar förutsättningar att fördela resurserna i andra delar av verksamheten.

Med hjälp av en statistisk modell som predikterar ärendeflödet kan Atrium Ljungberg skapa en prognos och därefter fördela sin personalstyrka på ett effektivare sätt. Med hjälp av informationen om antalet ärenden som kommer uppstå vid en given plats och tidpunkt kan fastighetsbolaget planera scheman för sina driftteknikeroch på så sätt säkerställa att de har den kompetensen som krävs för att uträtta ett visst ärende. En balanserad personalfördelning innebär att företaget kan fördela sina resurser där det behövs vilket minskar risken för över- och underbemanning. Detta leder både till ökad trivsel hos medarbetarna eftersom arbetsbördan inte blir för hög men också att hanteringstiden kan förkortas vilket resulterar i bättre service och nöjdare kunder. Eventuella följdeffekter av detta kan innebära ökade intäkter och en minimerad personalkostnad.

Tidigare studier som syftar till att prediktera frekvensen av en händelse över tid har undersökt olika statistiska metoders effektivitet. I en studie undersöktes reperationsintervallen för flygplansdelar, i studien undersöktes 6 olika statistiska metoder där Support Vector Regression gav högst prediktiv förmåga (Baptista et al. 2018). En annan studie som undersöker trafikolyckor syftade till att prediktera antalet dödsfall, skador och olyckor har använt sig av data från olika delar av Indien mellan 1970 och 2017 (George 2020). Där användes olika metoder för att skapa en optimerad modell, bland annat Poission regression, AutoRegressive Integrated Moving Average (ARIMA) och Vector Autoregressive model (VAR). ARIMA är en vedertagen typ av modell för att modellera tidsserier. På senare tid har även modeller som använder gradient boosting börjat etablera

sig inom tidsseriemodellering. I en studie som syftar till att förbättra planering för distribution av cyklar i cykelpooler används XGBoost för att prognostisera antalet användare av cyklarna under kommande 36 timmar (Rodrigo 2022). Det visade sig att XGboost var en väl fungerande metod för att skapa prognoser för tidsserier, där modellen förbättrades avsevärt genom att inkludera exogena variabler.

1.2 Syfte och problemformulering

Syftet är att undersöka vilken av metoderna ARIMA, dynamisk regression och XGBoost som är mest lämplig för att prognostisera antalet ärenden per dag de kommande 14 dagarna för de olika driftområdena. Genom att utvärdera flera olika statistiska metoder kan den mest lämpliga av dessa tas fram, där utvärderingsmåttet är Root Mean Squared Error (RMSE). Prognoser kan användas för att optimera bemanning och se till att varken överbemanna vilket ökar personalkostnaden eller underbemanna vilket försämrar servicen.

- Vilken av metoderna ARIMA, dynamisk regression och XGBoost har lägst RMSE för flest prognoshorisonter under en 14 dagars period för respektive driftområde?

1.3 Etiska och samhälleliga aspekter

Genom hela arbetet har hänsyn tagits till Statistikfrämjandets etiska kod (Svenska statistikfrämjandet 2010).

Metoderna som använts för analyser beskrivs tydligt för att säkerställa den vetenskapliga grunden och för att möjliggöra utomstående granskning. Resultat och slutsatser har kvalitetssäkrats och presenterats på ett tydligt sätt. De har inte styrts för att åstadkomma förutbestämda resultat utifrån önskemål eller liknande.

Data är hämtat från en databas som hanteras av företaget WSP där all data har anonymiserats. Detta görs för att säkra risken för bakvägsindentifiering, vilket innebär att ingen koppling kan göras till vem som anmält eller hanterat ett ärende. Data har endast använts för att undersöka rapportens syfte och har hanterats varsamt.

Studien kan bidra till att säkerhetställa företagets kunders verksamhetsflöde. Genom en snabbare ärendehantering avbrotten som felen skapar kortare och stör därför inte kundernas verksamheter i samma utsträckning. En bättre personalfördelning hjälper till att förebygga för hög arbetsbelastning för driftteknikerna, vilket är viktigt för deras välmående.

2. Data

Detta kapitel syftar till att ge inblick i datamaterialet som analyseras i studien. Förklaring av datamaterialet, beskrivande statistik och transformering av variabler kommer vara huvudfokus i detta kapitel.

2.1 Datamaterialet

De ärenden som inkommer lagras i en databas som hanteras av företaget WSP. I databasen finns information om alla ärenden och denna information används sedan av Atrium Ljungberg för att hantera ärenden. Endast ett fåtal variabler plockats ut från databasen för att användas vid analyser, där dessa är de som anses vara rimliga för att besvara syftet och beskriva datamaterialet. För varje ärende finns det information om exempelvis vilken typ den klassas som, i vilket driftområde ärendet skett samt vilket datum som ärendet är registrerat. Data består av 110 949 serviceanmälningar mellan 7 maj 2010 och 1 december 2021.

I tabell 2.1 visas de driftområden som finns i datamaterialet. Dessa områden innehåller fastigheter som tillhör företaget och det är för dessa fastigheter som serviceärenden registreras. Vid tillfället då data extraherades från databasen fanns det totalt nio stycken driftområden och det är dessa som kommer användas i studien.

Tabell 2.1: Datamaterialets driftområden

Driftområde
Göteborg/Lindholmen
Slakthusområdet/Proppen
Slussen
Kista/Sundbyberg
Uppsala
Medborgarplatsen/Liljeholmen
Hagastaden/City
Malmö
Sickla

Tabell 2.2 visar alla olika typer av ärenden som kan registreras i databasen, där det finns 11 olika ärendekategorier.

Tabell 2.2: Datamaterialets ärendekategorier

Ärendekategori
Styr- och Övervakningssystem
Utemiljö
Vitvaror och Tvättstuga
Transportsystem
Elsystem
Byggnad utvändigt
Allmänt
Inneklimat
VVS
Tele, passage och datasystem
Byggnad invändigt

Datamaterialet delades upp i tre mängder med olika syften. Träningsmängden användes för att träna modellerna och hitta optimala parametrar. Valderingsmängden användes för att utvärdera de tränade modellerna och optimera regulariseringsparametrarna. När de slutgiltiga modellerna valts användes testmängden för att skapa prognoser och undersöka hur modellerna presterar på ny data. Genom att dela upp data på detta sätt kan mer pålitliga modeller skapas och det är lättare att undvika både under- och överanpassning.

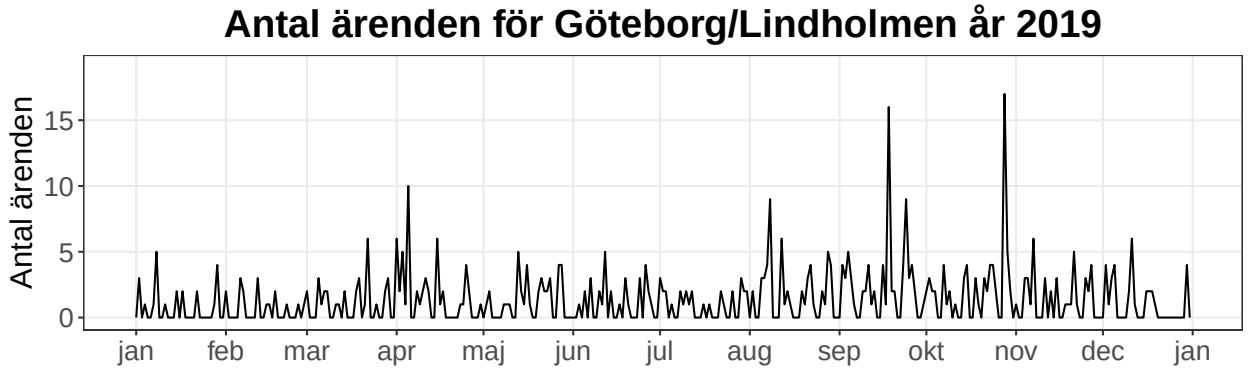
Tabell 2.3: Uppdelning av datamaterialet

Mängd	**Tidsperiod**	**Andel i data**
Träningsmängd	2015-11-02 - 2020-11-30	84%
Valideringsmängd	2020-12-01 - 2021-08-31	12%
Testmängd	2021-09-01 - 2021-12-01	4%

Figur 2.1 och 2.2 visar hur ärendena fördelar sig i Sickla respektive Göteborg/Lindholmen från 1 januari 2019 till 31 december 2019.

Figur 2.1: Antal ärenden för Sickla år 2019

Figur 2.1 visar tidsserien över serviceärenden per dag i Sickla under 2019. Ur figuren går det att se att ärendeflödet är mellan 0 och 30 ärenden per dag. Det ser inte ut att finnas någon ökande trend.



Figur 2.2: Antal ärenden för Göteborg/Lindholmen år 2019

Figur 2.2 visar tidsserien över serviceärenden per dag i Göteborg år 2019. Från figuren går det att se att antal ärenden per dag ligger mellan ungefär 16 och 0 samt att tidsserien inte har någon ökande trend i antal ärenden under året.

2.2 Avgränsningar

Datamaterialet består av registrerade ärenden mellan maj 2010 och december 2021. Mellan dessa år har företaget utökat sitt innehav av fastigheter i fler driftområden, vilket innebär att vissa driftområden inte har registrerad data från 2010. För att få en representativ bild av antalet ärenden i respektive driftområde har avgränsningar genomförts, där data avgränsats till ärenden registrerade mellan 2 november 2015 och 1 december 2021. Detta för att fastighetsbeståndet under denna tid ansågs representera företagets nuvarande.

2.3 Aggregering

De ärenden som registreras i databasen kategoriseras beroende på typ av ärende, där de olika kategorierna som ett ärende kan tilldelas presenterades i tabell 2.2. I detta arbete har inte hänsyn tagits till vilken kategori ett ärende tillhört, utan arbetet fokuserar på det totala antalet ärenden per dag i respektive driftområde. I och med att serviceärendena inte utfärdas av specifik servicepersonal utifrån respektive ärendekategori, ansågs detta inte vara relevant att ta hänsyn till för rapportens syfte. Aggregeringen har även genomförts på grund av det stora antalet nollor i datamaterialet, vilket ofta leder till problem när prognoser skall skapas.

På grund av att ärenden endast registreras på det datum de uppstår behövde nollor implementeras för de datum som inte hade registrerade ärenden. Detta eftersom en komplett tidsserie behövdes för att kunna utföra analyser. Det skapades även datamaterial som aggregerade antalet ärenden utifrån veckodag eller månad, dessa kunde användas för att undersöka antalet ärenden på olika tidsnivåer som i sin tur kunde ge mer insikt i data och se eventuella säsongsmönster.
För att vidare kunna skapa analyser som tog hänsyn till eventuella säsongsmönster i data skapades indikatorvariabler för veckodag och månad. Variabeln antar 1 vid den specifika veckodagen/månaden och 0 annars. För veckodagarna användes måndag som referenskategori och för månaderna används Januari.

2.4 Datahantering för XGBoost modellering

För att använda XGBoost till tidsseriedata förbereddes datamaterialet med laggade antal som förklaringsvariabler för de 28 senaste dagarna. Detta för att ärendeflödet de kommande 14 dagarna antas påverkas av tidigare fyra veckors ärendeflöde. Förutom laggade antal så inkluderades även indikatorvariabler för veckodagar och månader som tidigare beskrevs i kapitel 2.3.

2.5 Beskrivande statistik

I kapitlet kommer beskrivande statistik för data presenteras, detta för att ge en tydligare bild av datamaterialet. Antalet ärenden för varje driftområde undersöks för att få en uppfattning över hur ärendena är fördelade. Tabell 2.4 och figur 2.3 visar fördelningen av ärenden i de olika driftområdena.

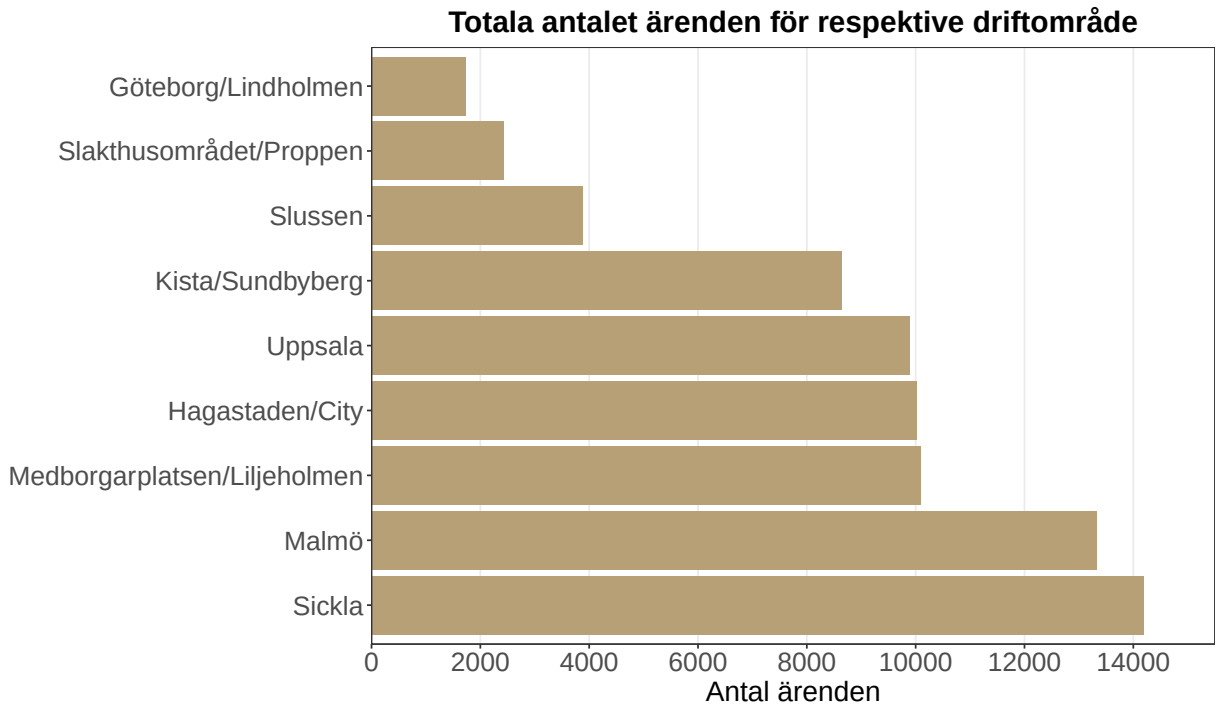
Tabell 2.4 visar antalet samt andelen ärenden i data för respektive driftområde. Ur tabellen går det att se att det finns tre driftområden som har relativt få ärenden, Göteborg/Lindholmen, Slakthusområdet/Proppen

och Slussen utgör var för sig mindre än 10% av datamaterialet. Totalt har datamaterialet 74 206 ärenden mellan november 2015 och december 2021.

Tabell 2.4: Antal ärenden uppdelat på driftområde

Driftområde	**Antal ärenden**	**Andel av data**
Göteborg/Lindholmen	1 730	2.33%
Slakthusområdet/Proppen	2 431	3.28%
Slussen	3 877	5.22%
Kista/Sundbyberg	8 639	11.64%
Uppsala	9 890	13.33%
Hagastaden/City	10 017	13.50%
Medborgarplatsen/Liljeholmen	10 103	13.61%
Malmö	13 330	17.96%
Sickla	14 189	19.12%
Totalt	74 206	100%

Tabell 2.4 har även visualiserats i figur 2.3 för att tydligare visa skillnader i antal ärenden mellan driftområdena.



Figur 2.3: Fördelningen av ärenden i de olika driftområdena

Figur 2.3 visar antalet ärenden för respektive driftområde. Det framgår tydligt att driftområde Sickla har

flest antal anmälda ärenden medan Göteborg/Lindholmen har minst.

Tabell 2.5 visar andelen dagar som saknar ärenden inom respektive driftområde, det vill säga andel dagar med 0 ärenden. Det går exempelvis att se att av alla dagar i datamaterialet saknar driftområdet Göteborg/Lindholmen anmälda ärenden på 58.78% av dagarna. De driftområden med de tre största andelarna av dagar utan ärenden är även de driftområden som är minst antal ärenden i data enligt tabell 2.4.

Tabell 2.5: Andel dagar utan ärenden uppdelat på driftområde

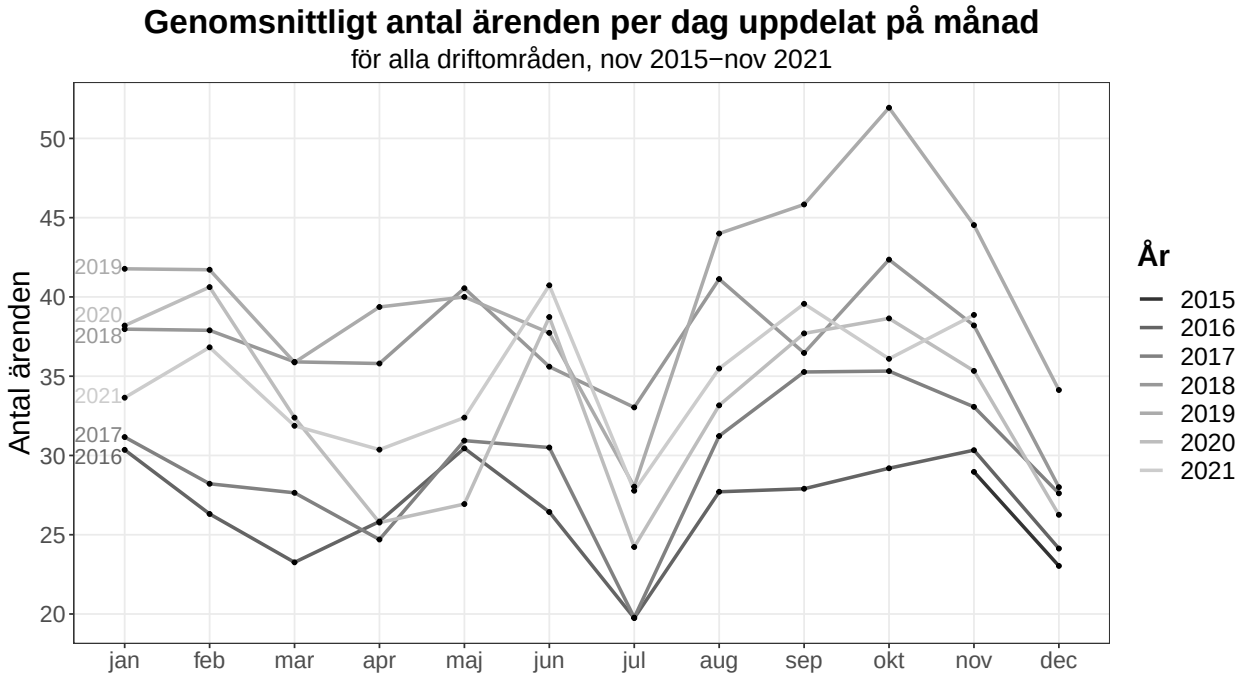
Driftområde	**Andel dagar utan ärenden**
Göteborg/Lindholmen	58.78%
Slakthusområdet/Proppen	51.80%
Slussen	40.14%
Kista/Sundbyberg	26.60%
Hagastaden/City	24.08%
Medborgarplatsen/Liljeholmen	20.97%
Sickla	10.13%
Uppsala	6.80%
Malmö	5.76%

I tabell 2.6 visas ytterligare beskrivande statistik för respektive driftområde. Tabellen visar högsta antal anmälda ärenden på en dag i ett driftområde, medianen samt det genomsnittliga antalet ärenden per dag. Det visas även ett typvärde inom varje område, detta visar hur många antal ärenden per dag som är mest förekommande. Från tabellen kan det utläsas att Sickla har det högsta antalet ärenden per dag samt det största medelvärdet, men har ett typvärde på endast två serviceärenden per dag. Medan Malmö har det största typvärdet på 6 serviceärenden per dag med ett max antal ärenden per dag på 27 stycken. Driftområdet med lägst maximalt antal ärenden på en dag var Slakthusområdet/Proppen med 13 ärenden som mest, där även medelvärdet var en av de lägsta på 1.09 ärenden i genomsnitt.

Tabell 2.6: Förtydligande av ärendefördelning inom driftområden

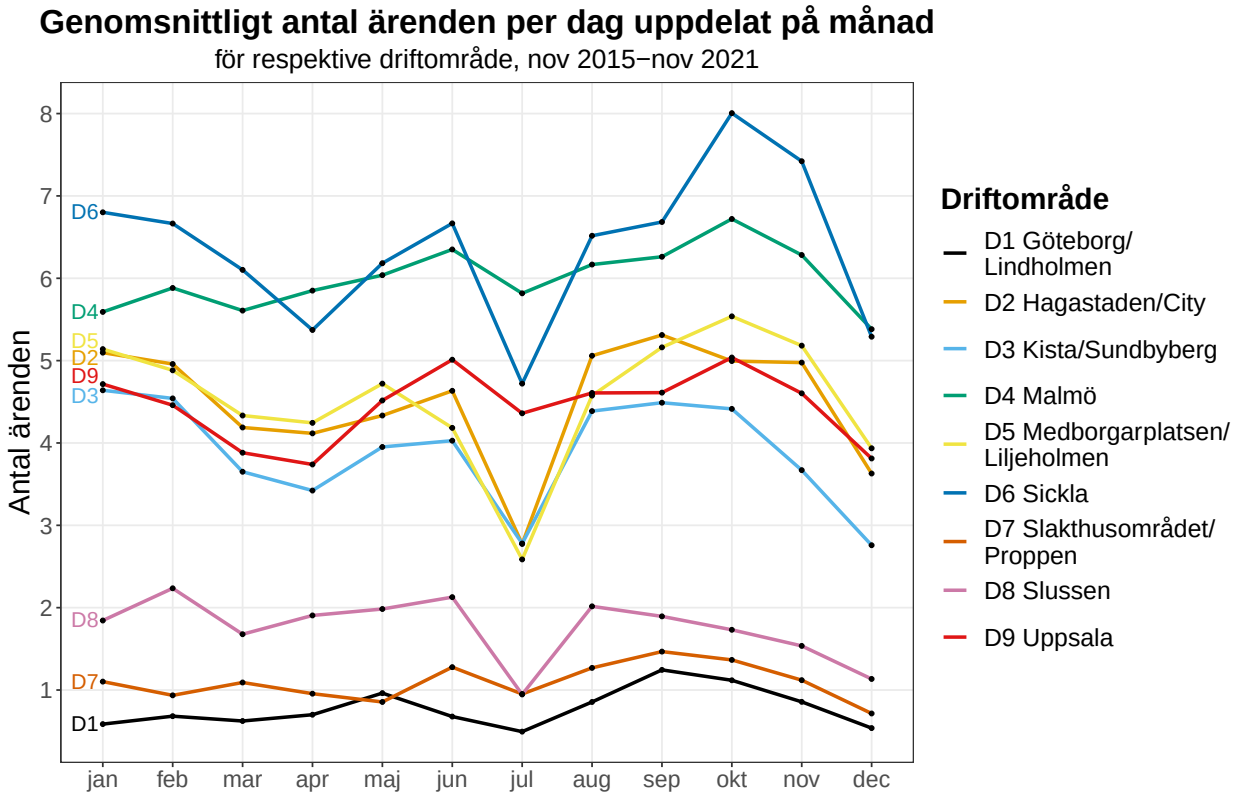
Driftområde	**Max antal**	**Median**	**Medelvärde**	**Typvärde**
Medborgarplatsen/Liljeholmen	25	4	4.55	0
Slakthusområdet/Proppen	13	0	1.09	0
Malmö	27	6	6.00	6
Sickla	44	6	6.39	2
Uppsala	22	4	4.45	3
Kista/Sundbyberg	39	3	3.89	0
Göteborg/Lindholmen	30	0	0.78	0
Hagastaden/City	28	4	4.51	0
Slussen	15	1	1.74	0

Data undersöktes även på olika tidsnivåer, på månadsbasis och veckobasis. Detta användes för att tydligare undersöka eventuella säsongsvariationer i datamaterialet.



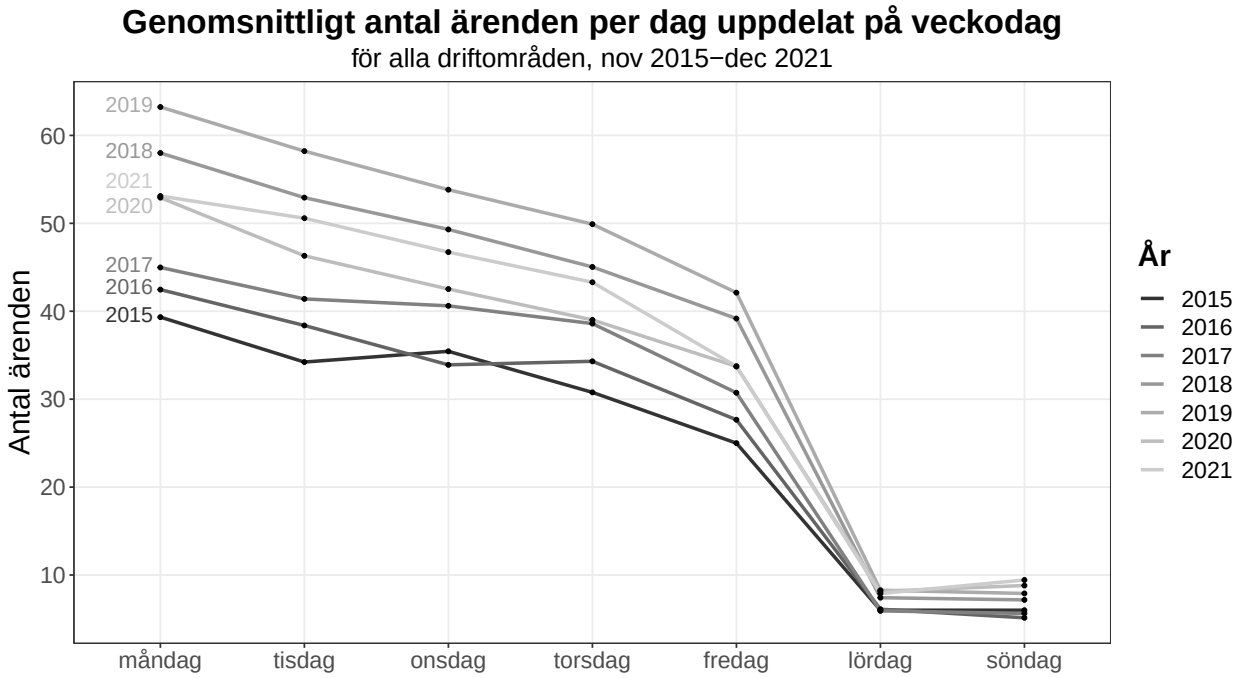
Figur 2.4: Genomsnittligt antal ärenden per dag uppdelat på månad för respektive år

Figur 2.4 visar genomsnittligt antal ärenden per dag för varje hel månad under datamaterialets tidsperiod summerat för alla driftområden. Ur diagrammet går det att se en minskning i ärendeflöde vid juli varje år. Ärendeflödet verkar även öka varje år runt hösten. Både åren 2016 och 2017 har något färre ärenden än de övriga åren dock verkar det inte finnas någon ökande trend av ärendevolymer då 2019 har fler ärenden än både 2020 och 2021.



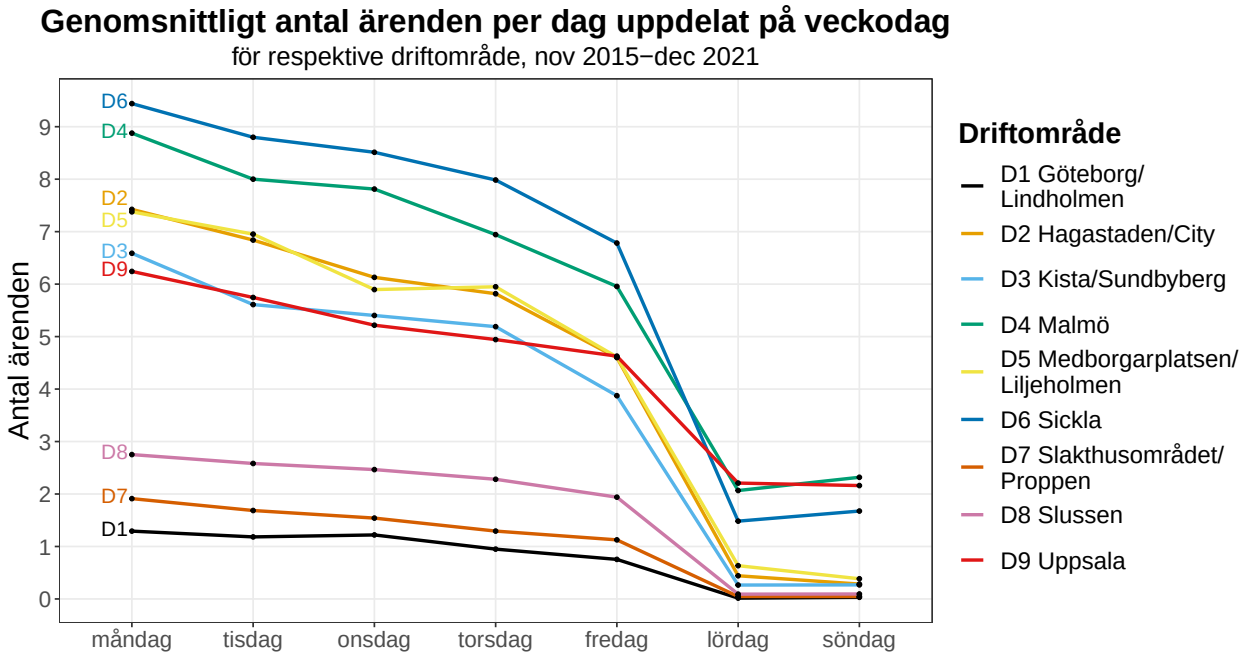
Figur 2.5: Genomsnittligt antal ärenden per dag uppdelat på månad för respektive driftområde

Figur 2.5 visar genomsnittligt antal ärenden per dag för varje hel månad uppdelat på respektive driftområde. Diagrammet visar ett liknande mönster som figur 2.4 där ärendeflödet minskar kraftigt vid juli månad för samtliga driftområden. Ärendevolymen ser även ut att variera mellan olika driftområden där driftområdena kan delas in i tre ärendevolyms grupper. Områdena Göteborg/Lindholmen, Slakthusområdet/Proppen samt Slussen har minst antal ärenden med ett genomsnitt på ca 1-2 ärenden per dag. Kista/Sundbyberg, Uppsala, Hagastaden/City och Medborgarplatsen/Liljeholmen verkar ha liknande ärendevolym med ett medelvärde på ca 3-5 ärenden per dag. Områdena Malmö och Sickla har högst ärendevolym med ca 5-8 ärenden per dag.



Figur 2.6: Genomsnittligt antal ärenden per dag uppdelat på veckodag för respektive år

Figur 2.6 visar genomsnittligt antal ärenden per dag för varje veckodag under datamaterialets tidsperiod, summerat för alla driftområden. Ur diagrammet går det att tyda en fallande trend där veckodagarna senare i veckan har i genomsnitt färre ärenden. Den största skillnaden är mellan vardagar och helger. Genomsnittet för veckodagar är mellan 25 och 65 ärenden och genomsnittet för helger är mellan 5 och 10 ärenden. År 2015 till 2017 har något färre ärenden i genomsnitt samt år 2018 till 2021 har något fler ärenden i genomsnitt.



Figur 2.7: Genomsnittligt antal ärenden per dag uppdelat på veckodag för respektive driftområde

Figur 2.7 visar genomsnittligt antal ärenden per dag för varje veckodag uppdelat på respektive driftområde. Ur diagrammet går det att tyda samma variation mellan veckodagarna som i figur 2.6 där veckodagarna i slutet på veckan har färre ärenden och där den största skillnaden är att vardagarna har betydligt fler ärenden än helgdagarna. Ärendevolymen följer samma grupper som tidigare konstaterades i figur 2.5 där driftområdenas ärendevolym kan delas in tre grupper. Områdena Göteborg/Lindholmen, Slakthusområdet/Proppen samt Slussen har minst antal ärenden med ett genomsnittsnitt på ca 0-3 ärenden per dag. Kista/Sundbyberg, Uppsala, Hagastaden/City och Medborgarplatsen/Liljeholmen verkar ha liknande ärendevolym med ett medelvärde på ca 0-8 ärenden per dag. Områdena Malmö och Sickla har högst ärendevolym med ca 1-10 ärenden per dag. Det som skiljer från grupperna som tidigare konstaterades i figur 2.5 är att på helgdagarna så har inte bara Sickla och Malmö fler ärenden utan även Uppsala, dessa områden har ca 2-3 ärenden per dag. De övriga områdena har betydligt färre ärenden, där alla har mindre än 1 ärende per dag i genomsnitt.

3. Metod

I följande kapitel presenteras de metoder som har använts i uppsatsen.

3.1 Stationäritet och differentiering

För att undersöka stationäriteten i tidsserien finns tre viktiga delar som behöver kontrolleras (Hyndman & Athanasopoulos 2021) En tidsserie är stationär om:

- Konstant väntevärde över tid, bortsätt från säsongsvariation
- Konstant varians över tid
- Autokorrelationen får endast bero på tiden mellan observationer

Detta kan undersökas med hjälp av autokorrelation och partiell autokorrelation. Genom att undersöka dessa grafiskt kan olika mönster stödja antagandet om stationäritet. Autokorrelationskoefficienter som beräknas med autokorrelationsfunktionen visar hur starkt observationerna mellan vissa tidpunkter korrelerar med varandra. Den partiella autokorrelationen undersöker korrelationen mellan observationer som inte förklaras av de tidigare korrelationerna, vilket kan användas för att undersöka ordningen av den autoregressiva delen i modellen. Om en tidsserie ej är stationär kan differentiering tillämpas för att skapa stationäritet. För första ordningens differentiering görs enligt:

$$y'_t = y_t - y_{t-1} \tag{3.1}$$

där

$$\begin{aligned} y_t &= \text{En observation vid tidpunkten t} \\ t &= \text{Tidpunkten t} \end{aligned}$$

Detta kan även göras med en så kallad bakåtskiftsoperator. Denna operator är användbar för att beskriva just differentiering, där den förenklar sättet att skriva komplicerade ekvationer. Ekvation 3.1 kan skrivas om med hjälp av bakåtskiftsoperatorn enligt:

$$y'_t = y_t - y_{t-1} = y_t - By_t = (1 - B)^1 y_t \tag{3.2}$$

Formel 3.2 kan generaliseras för alla ordningar av differentiering och skrivs då enligt:

$$(1 - B)^d y_t \tag{3.3}$$

Differentieringen kan upprepas tills stationäriteten är uppnådd. När tidsserien kan anses vara stationär kan data användas för att skapa modeller.

3.2 ARIMA-modeller

En ARIMA-modell, eller Autoregressive integrated Moving Average, är en typ av modell som ofta används för att prediktera tidsserier. Den kan hantera både stationära och icke stationära tidsserier. Problem med stationäritet kan uppstå i tidsserier som har någon tydlig trend eller säsong vilket ofta gör att kraven för stationäritet inte är uppfyllda. Till skillnad från en klassisk regressionmodell med förklarande variabler, använder sig ARIMA-modellen av tidigare värden från tidsserien för att skapa prediktioner. Den autoregressiva delen av modellen innebär att tidsserien successivt förklarar sig själv genom regression av tidigare observerade värden medan den delen som baseras på glidande medelvärde använder sig av tidigare skattningsfel (Hyndman & Athanasopoulos 2021).

3.2.1 ARIMA-modellens uppbyggnad

ARIMA-modellen är som tidigare nämnt uppbyggd av en autoregressiv del och en del kallad glidande medelvärde. Den består även av en del som beskriver om differentiering har utförts. Detta innebär att modellen består av tre ingående komponenter som kommer beskriva modellen (Hyndman & Athanasopoulos 2021).

$$ARIMA(p, d, q) \tag{3.4}$$

där

$$\begin{aligned} p &= \text{Ordning av autoregressiv del} \\ d &= \text{Antal differentieringar} \\ q &= \text{Ordning av glidande medelvärde} \end{aligned}$$

Värdena på p,d och q är kritiska för att skapa en bra modell. Vid högre värden på p, d, och q blir modellen mer flexibel. För höga värden på dessa parametrar kan därför leda till överanpassning (Hyndman & Athanasopoulos 2021).
ARIMA-modellen formuleras enligt:

$$(1 - \phi_1 B - ... - \phi_p B^p)(1 - B)^d y_t = c + (1 + \theta_1 B + ... + \theta_q B^q)\varepsilon_t \tag{3.5}$$

Den autoregressiva delen i ARIMA formuleras enligt:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \varepsilon_t \tag{3.6}$$

Och glidande medelvärde modelleras enligt:

$$y_t = c + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + ... + \theta_q \varepsilon_{t-q} + \varepsilon_t \tag{3.7}$$

där

$$\begin{aligned} \phi &= \text{Parameter för p} \\ \theta &= \text{Parameter för q} \\ c &= \text{En konstant} \\ \varepsilon_t &\sim N(0, \sigma^2) \\ \sigma &= \text{Standardavvikelse} \end{aligned}$$

Modellen är definierad för att göra prognoser 1 dag fram i tiden, men med hjälp rekurivmetod kan prognoser för fler steg skapas. Detta innebär att för att kunna skapa prognoser över en längre tid skapas prediktioner med hjälp av det senast skattade värdet. För att finna de optimala parametervärdena minimeras maximum likelihood funktionen.

3.2.2 SARIMA-modeller

SARIMA-modellen är en typ av ARIMA-modell men även som tar hänsyn till säsongsvariation i tidsserien. Modellen formas på samma sätt som en ARIMA men lägger även till säsongskomponenter som betecknas med versaler (Hyndman & Athanasopoulos 2021).

$$ARIMA(p,d,q)(P,D,Q)_m \tag{3.8}$$

där

$$\begin{aligned} p &= \text{Ordning av autoregressiv del} \\ d &= \text{Antal differentieringar} \\ q &= \text{Antal glidande medelvärdes termer} \\ P &= \text{Ordning av autoregressiv del för säsong} \\ D &= \text{Antal differentieringar för säsong} \\ Q &= \text{Antal glidande medelvärdes termer för säsong} \\ m &= \text{Säsongperioden/Säsongskomponent} \end{aligned}$$

Införs säsongskomponenterna i formel 3.5 skapas en SARIMA-modell och ekvationen skrivs enligt:

$$\begin{aligned} &(1-\phi_1 B - ... - \phi_p B^p)(1-\Phi_1 B - ... - \Phi_P B^{P,m})(1-B)^d(1-B^m)^d y_t = \\ &(1+\theta_1 B + ... + \theta_q B^q)(1+\Theta_1 B + ... + \Theta_Q B^{Q,m})\varepsilon_t \end{aligned} \tag{3.9}$$

där

$$\begin{aligned} \Phi &= \text{Parameter för P} \\ \Theta &= \text{Parameter för Q} \end{aligned}$$

3.2.3 Dynamiska regressionsmodeller

En dynamisk regressionsmodell skapas genom en kombination av en ARIMA-modell och en regressionsmodell. Först anpassas regressionsparametrar till data och sedan modelleras den resterande variationen av en ARIMA-modell. Dessa modeller används för att kunna använda viktig information från förklarande variabler samtidigt som hänsyn tas till komplexiteten i tidsserier (Hyndman & Athanasopoulos 2021). Denna metod är användbar om fler variabler utöver tid antas påverka data.

En regressionsmodell modelleras som en linjärfunktion enligt:

$$y_t = \beta_0 + \beta_1 x_{1,t} + ... + \beta_k x_{k,t} + \varepsilon_t \tag{3.10}$$

där

$$\begin{aligned}
\beta &= \text{Regressionparametrar} \\
x &= \text{Observerade värden för förklaringsvariablerna vid tidpunkt t} \\
k &= \text{Förklaringsvariabel nummer k} \\
\varepsilon_t &\sim N(0, \sigma^2)
\end{aligned}$$

ε_t antas vara icke korrelerat vitt brus, alltså variation som inte kan förklaras av regressionsmodellen. I den dynamiska regressionmodellen tillåts istället feltermen innehålla autokorrelation och modelleras med hjälp av en ARIMA/SARIMA-modell, där denna modellen istället innehåller det vita bruset.

Den dynamiska regressionmodellen modelleras då enligt:

$$y_t = \beta_0 + \beta_1 x_{1,t} + ... + \beta_k x_{k,t} + \eta_t \tag{3.11}$$

där både x och y kan vara differentierade enligt 3.3. η_t är feltermer från regressionsmodellen och modelleras då enligt:

$$\begin{aligned}
&(1 - \phi_1 B - ... - \phi_p B^p)(1 - \Phi_1 B - ... - \Phi_P B^{P,m})(1 - B)^d (1 - B^m)^d \eta_t = \\
&(1 + \theta_1 B + ... + \theta_q B^q)(1 + \Theta_1 B + ... + \Theta_Q B^{Q,m})\varepsilon_t
\end{aligned} \tag{3.12}$$

För att bestämma alla parametrar i modellen används minsta kvadratmetoden för att minimera ε_t (Hyndman & Athanasopoulos 2021).

3.2.4 Utvärderingsmått för ARIMA och test för parametrar

För att hitta parametrarna p, d, q, P, D och Q som ger bäst modellanpassningar används olika typer av mått och tester beroende på parameter. AIC används för att optimera värden på p, q, P och Q. Canova-Hansen test och KPSS unit-root test används för att bestämma ordningen av differentiering, alltså värden på d och D (Hyndman & Khandaka 2008). Dessa tester är inbyggda i funktionen som används i R-studio.

AIC

AIC används för att finna optimala värden för parametrarna p, q, P och Q i ARIMA/SARIMA-modeller, där ett lågt AIC är önskvärt. För att finna parametrarna som ger lägst AIC kan en sökalgoritm bilda en mängd modeller och de parametrarna i den modell som ger lägst AIC anses vara de optimala (Hyndman & Khandakar 2008). AIC beräknas enligt:

$$AIC = -2log(L) + 2(p + q + P + Q + k) \tag{3.13}$$

där

L = Maximum likelihood för modellen anpassad till differensierad data
$k = 1$ om $c \neq 0$, 0 annars

Canova-Hansen test

Canova-Hansen test används för att hitta D för data med säsong, där testet undersöker om säsongsmönstret är stabilt över tid. Nollhypotesen som testas är om den det valda D har skapat en stabil säsong i data över tid. Detta innebär att en nollhypotes som inte kan förkastas innebär att ett lämpligt D har valts (Hyndman, R. J., & Khandakar, Y. 2008).

KPSS unit-root test

KPSS unit-root test används för att bestämma d för data utan säsongsvariation. Unit-root testet undersöker stationäritet utifrån bestämda d och D. Detta innebär att nollhypotesen som testas säger att data är stationär, vilket leder till att en nollhypotes som motbevisas innebär att data behöver differentieras. Om p-värdet är lägre än α förkastas nollhypotesen och data antas inte vara stationär. (Hyndman, R. J., & Khandakar, Y. 2008).

3.2.5 Ljung-Box test

För att utvärdera om modellen är lämplig för data används residualanalys. Genom att göra ett Ljung-Box test kan det undersökas om residualerna kan antas vara vitt brus (Bowerman, O'Connell & Koehler. 2005) . Teststatistikan Q^* beräknas enligt formel 3.14 och hypoteserna som testas är:

H_0: Det finns inte autokorrelation i residualerna i laggarna 1 till K
H_a: Det finns autokorrelation i residualerna i minst en av laggarna 1 till K

$$Q^* = n'(n'+2)\sum_{l=1}^{K}(n'-l)^{-1}r_l^2(\hat{a}) \tag{3.14}$$

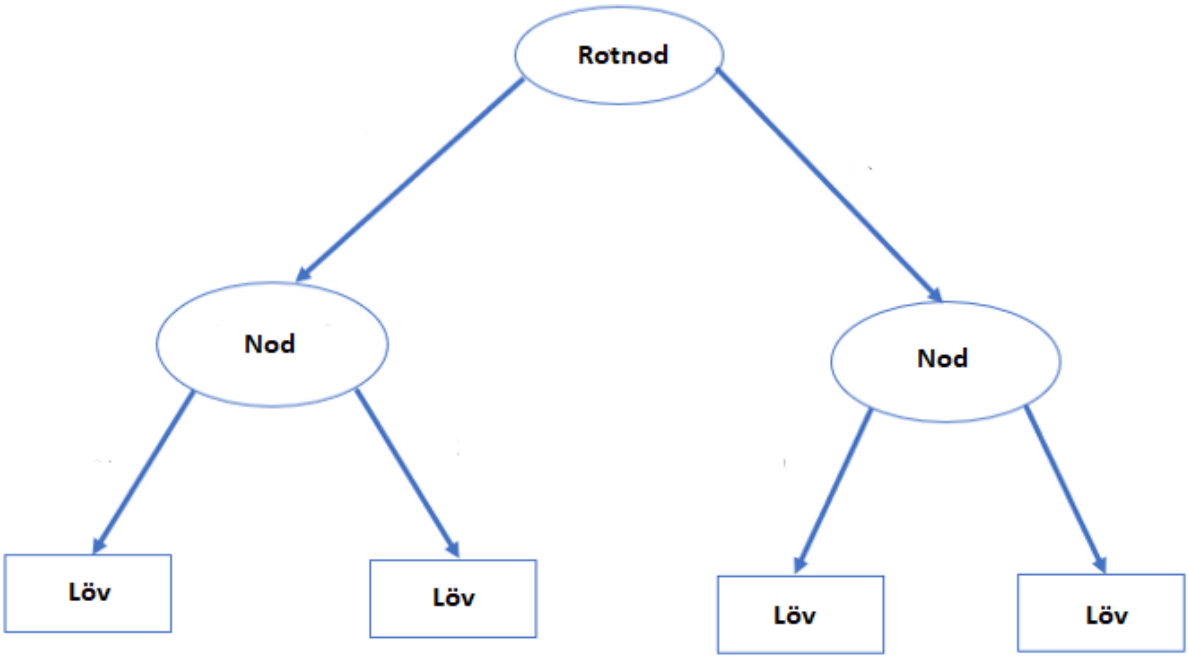
där

n = Antalet observationer i orginaltidsserien
$n' = n - d$
l = Antal lag
$r_l^2(\hat{a})$ = Kvadrerade autokorrelationen av residualerna vid lag l

H_0 kan förkastas om $Q^* > \chi_\alpha^2(K - n_c)$ där n_c är antalet parametrar i modellen. Testresultatet kan också fastställas genom att studera testets p-värde. Om p-värdet är mindre än α antas modellens residualer vara autokorrelerade, där p-värdet är arean till höger om $\chi_\alpha^2(K - n_c)$. K kan väljas beroende på α och tidsseriens längd (Hassani Reza Yeganegi 2020).

3.3 XGBoost

XGBoost är en typ av maskininlärningsmetod som bygger på trädmodeller. Trädmodeller konstrueras med förklaringsvariablerna i data och modellerar därefter olika uppdelningar utifrån dessa. Modellen startar med alla observationer i rotnoden och delar sedan data i två delar utifrån något bestämt villkor. Dessa uppdelningar fortsätter att ske tills något bestämt kriterium uppnåtts. Exempel på dessa villkor och kriterium finns i avsnitt 3.5.3. En visualisering av ett enkelt beslutsträd visas i figur 3.1.



Figur 3.1: Visualisering av ett beslutsträd

XGBoost är en metod som är användbar för att hantera stora och komplicerade datamängder och kan användas för exempelvis regression och klassificering. Den stora fördelen med XGBoost är hur omfattande den undersöker olika scenarion, som kortfattat kan beskrivas som att många olika trädmodeller skattas och vägs samman för att skapa en optimal modell (Chen Guestrin 2016). Detta kallas för Gradiet Boosting eftersom att nya träd skattas med hjälp av tidigare träds residualer. Ytterligare en fördel med XGBoost är att metoden inte ställer samma krav på data som andra traditionella statistiska metoder, så som normalfördelade residualer. Formel 3.15 visar ekvationen för att prediktera med hjälp av additiva funktioner.

$$\hat{y}_i = \sum_{k=1}^{s} f_k(x_i), f_k \in \mathcal{F} \tag{3.15}$$

där

$$\begin{aligned}
\mathcal{F} &= \{f(x) = w_{b(w)}\}(b : \mathbb{R}^m \to T, w \in \mathbb{R}^T) \\
s &= \text{Antal additiva steg} \\
x &= \text{Förklaringsvariabel} \\
f(x) &= \text{En trädmodell med struktur b och vikt w} \\
w &= \text{Vikt för ett löv i trädmodellen} \\
b &= \text{Strukturen för varje trädmodell} \\
T &= \text{Antalet löv i en trädmodell}
\end{aligned}$$

F kan ses som rymden för alla regressionträd. De olika funktionerna i modellen behöver tränas för att skapa en bra prediktion, detta görs genom att minimera det regulariserande kostnadsfunktion enligt:

$$\mathcal{L}(\phi) = \sum_i u(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{3.16}$$

där

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda ||w||^2 \tag{3.17}$$

$$\begin{aligned}
u &= \text{En differentierbar konvex förlustfunktion} \\
\Omega &= \text{Straffterm för komplexiteten för funktionerna i modellen} \\
\gamma &= \text{Regulariseringsparameter} \\
\lambda &= \text{Regulariseringsparameter för lövvikterna} \\
i &= \text{Observation nr i}
\end{aligned}$$

Då det ofta är omöjligt att undersöka alla trädstrukturer q, används en metod som utvärderar varje uppdelning i trädet. Detta går ut på att trädet utgår från rotnoden, vilket är den nod som innehåller alla observationer, därefter delas observationerna upp i noder utefter den uppdelning av en förklaringsvariabel som bäst kan förklara responsvariabeln. Processen sker upprepade gånger för att tillslut dela in observationerna i löv. Detta kallas för en greedy search eftersom modellen väljer den uppdelning som bäst förklarar responsvariabeln och därmed minimerar felet mest vid varje uppdelning.

$$L_{split} = \frac{1}{2}\left[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma \tag{3.18}$$

där

$$\begin{aligned}
g_i &= \partial_{\hat{y}^(s-1)} u(y_i, \hat{y}^{(s-1)}) \\
h_i &= \partial^2_{\hat{y}^(s-1)} u(y_i, \hat{y}^{(s-1)})
\end{aligned}$$

där

$$I_L = \text{Vänstra noden efter uppdelning}$$
$$I_R = \text{Högra noden efter uppdelningen}$$
$$I = I_L \cup I_R$$

Ekvation 3.18 mäter förlustminskning efter en uppdelning och kan därför användas för att jämföra olika uppdelningar, där ett större värde innebär en större förlustminskning och där med en bättre noduppdelning.

De kritiska momenten för att skapa en bra anpassad XGBoost-modell är att välja rätt hyperparametrar och optimera deras värden. De olika parametrarna har olika funktion för modellen och det finns många olika att välja mellan, där de även kan anta en mängd olika värden. För att hitta de mest lämpliga värdena på parametrarna används parametertuning, som söker igenom en mängd olika värden för parametrarna och bestämmer de bästa värdena utifrån ett bestämt utvärderingsmått. Parametertuning kan ske med olika typer av metoder, till exempel korsvalidering eller gridsökning.

3.4 Utvärderingsmått för att jämföra mellan modeller

RMSE är ett mått som jämför skattade värden med observerade värden, detta mått är användbart för att jämföra resultat mellan olika typer av modeller. Ett lågt värde på RMSE innebär att modellen skapar mer säkrare skattningar, då skillnaden mellan skattatvärde och observerat värde kommer minimeras (Hyndman, R.J., Athanasopoulos, G. 2021).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \tag{3.19}$$

Med MAE kan genomsnittliga absoluta felet undersökas. Eftersom MAE ger en konkret tolkning av modellens prestanda så är det ett lämpligt mått för att undersöka hur bra den valda modellen är. Ett lågt MAE innebär att modellen i snitt skattar genomför en skattning närmre det observerade värdet (Hyndman, R.J., Athanasopoulos, G. 2021).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i^2\right| \tag{3.20}$$

3.5 Programvaror

För att hantera, visualisera och analysera data har kod skrivits i programmeringsspråket R. Med hjälp av RStudio och R-paket kan mängder av statistiska uppgifter utföras. För databearbetningen har paketet fpp3 använts, detta paket innehåller ett flertal andra paket bland annat paketet fable som används för att skatta ARIMA, SARIMA och dynamiska regressionmodeller (O'Hara-Wild, Hyndman & Wang 2022). Detta paket genomför skattningar och väljer parametrar till modellerna automatiskt inuti funktionen. För att skatta XGBoost modeller har paketet xgboost använts (Chen et al. 2022). För visualisering av data har ggplot2 (Wickham et al. 2021) används.

3.6 Metodanpassningar

Detta kapitel syftar till att presentera de specifika anpassningar som har gjorts till de valda metoderna.

3.6.1 ARIMA/SARIMA

För att finna de mest optimala parametrarna har en sökning av alla möjliga kombinationer av p, d, q, P, D och Q som lyder under följande kriterium:

$$p + q + P + Q \leq 6 \tag{3.21}$$

$$c + d + D \leq 2 \tag{3.22}$$

Den parameter kombination med lägst AIC ansågs vara mest lämplig. Eftersom ARIMA och SARIMA inte är anpassad för positiva heltal avrundades prognoserna samt skattningarna innan beräkningarna av RMSE och MAE till närmaste heltal och de negativa skattningarna sätts till 0. För Ljung-Box testet kommer en signifikansnivå på 5% att användas.

3.6.2 Dynamisk regression

Då de förklarande variablerna veckodag och månad är statiska prediktorer är de alltid kända för ett specifikt datum och därför behövs ingen prognosmodell för dessa variabler. Parameterarna p, d, q, P, D och Q väljs enligt formlerna 3.21 och 3.22. Likt ARIMA och SARIMA-modelleringen är inte den dynamiska regressionen anpassad för positiva heltal och därför avrundas prognoserna samt skattningarna innan beräkningarna av RMSE och MAE till närmsta heltal och de negativa skattningarna sätts till 0. För Ljung-Box testet kommer en signifikansnivå på 5% att användas.

3.6.3 XGBoost

Genom att använda laggade observationer som förklaringsvariabler till XGBoost bildas en icke-linjär AR-modell. Detta eftersom XGBoost kan modellera icke-linjär data samt att modellen får information om ärendeflödet från de tidigare dagarna. Prognoser för tidpunkten t tas fram genom direkta prognoser. Detta innebär att en modell skapas för varje prognoshorisont istället för att använda skattade värden för att skapa prognoser längre än ett steg fram i tiden. I den observerade tidsserien omvandlas data till ny data med en responsvariabel och förklaringsvariabler. Detta innebär att tidsserieproblemet hanteras som ett regressionsproblem som sedan modelleras med XGBoost-modellen.

För att optimera XGBoost modellens hyperparametrar användes grid-sökning som parametertuning.Detta innebar att en grid med olika kombinationer av förutbestämda värden på parametrarna skapades och därefter skattades modeller med varje kombination. RMSE användes som utvärderingsmått för att välja vilken kombination som optimerade modellen. De hyperparametrar som användes i modellen var följande:

- $\boldsymbol{\eta}$ - Kontrollerar inlärningstakten för modellen i träningsfasen.
- $\boldsymbol{\gamma}$ - Anger den minsta reduceringen av förlustfunktionen för att genomföra ytterligare en partition av ett löv i trädet.
- $\boldsymbol{\lambda}$ - En regluariseringsparameter som utgår från L2 regularisering.
- **Maxdjup** - Bestämmer maximalt antal uppdelningar i trädet.

Tabell 3.1 visar de värden som hyperparametrarna kunde anta i modellerna, detta innebar att 864 olika kombinationer testades för varje modell för respektive prognoshorisont och driftområde.

Tabell 3.1: Parametervärden för parametertuning för XGBoost-modeller

Hyperparameter	**Parametervärden**					
η	0	0.2	0.4	0.6	0.8	1
γ	2	4	6	8	10	12
λ	0	2	4	6	8	10
Max träddjup	3	5	10	20		

Då responsvariabeln endast antar heltal i denna tidsserie ansågs det rimligt att använda en förlustfunktion som utgår från en poission fördelning (Lucas et al. 2019)

$$u(\hat{y}_t, y_t) = \frac{1}{N}\sum_{t=0}^{N}(\hat{y}_t - y_t log(\hat{y}_t)) \tag{3.23}$$

där

N = Antalet observationer i tidsserien
$\hat{y}_t$ = Skattat värde för tidpunkten t

4. Resultat

Följande kapitel redovisas resultatet från de olika modellerna. Syftet är att ge inblick kring hur olika typer av modeller har anpassats utifrån data, samt ge grund till de analyser som genomförs i senare kapitel. Kapitlet innehåller delkapitel som är uppdelade efter modelltyp.

4.1 ARIMA och SARIMA

I detta kapitel presenteras resultat från ARIMA & SARIMA-modeller. Dessa modeller innehöll inga förklarande variabler utan beror endast på tidigare observationer för antal ärenden. För varje driftområde skattades det flera modeller med olika kombinationer av p, d, q, P, D och Q, där sedan den med lägst AIC valdes för respektive område. Tabell 4.1 visar de slutgiltiga modellerna med tillhörande val av p, q, d, P, D och Q och säsongskomponent för respektive driftområde. Tabellen visar att alla driftområden har en säsongskomponent på sju, vilket innebär att den antar ett säsongsmönster på sju dagar och detta innebär att alla modeller är skattade som SARIMA. Det går även att urskilja att alla modeller har en komponent q för glidande medelvärde i den delen av modellen som ej tar hänsyn till säsong.

Tabell 4.1: Skattade modeller utan indikatorvariabler

Driftområde	**ARIMA(p,d,q)(P,D,Q)$_m$**
Göteborg/Lindholmen	ARIMA(3,1,1)(2,0,0)$_7$
Hagastaden/City	ARIMA(0,0,2)(0,1,1)$_7$
Kista/Sundbyberg	ARIMA(1,0,2)(0,1,1)$_7$
Malmö	ARIMA(3,1,2)(0,0,1)$_7$
Medborgarplatsen/Liljeholmen	ARIMA(0,0,3)(0,1,1)$_7$
Sickla	ARIMA(3,1,1)(0,0,2)$_7$
Slakthusområdet/Proppen	ARIMA(3,1,1)(2,0,0)$_7$
Slussen	ARIMA(0,1,2)(2,0,0)$_7$
Uppsala	ARIMA(2,1,1)(0,0,2)$_7$

Tabell 4.2 visar resultatet från Ljung-Box test för 50 lag. Det är endast residualerna för Kista/sundbyberg som kan antas vara vitt brus på 5% signifikansnivå då resterande p-värden är under 0.05.

Tabell 4.2: Resultat från Ljung-Box test tillhörande modellerna i tabell 4.1

Driftområde	**p-värde**
Göteborg/Lindholmen	0.00
Hagastaden/City	0.00
Kista/Sundbyberg	0.86
Malmö	0.00
Medborgarplatsen/Liljeholmen	0.00
Sickla	0.00
Slakthusområdet/Proppen	0.00
Slussen	0.00
Uppsala	0.00

4.2 Dynamiska regressionsmodeller

Vidare skall resultat från de dynamiska regressionmodellerna presenteras. Dessa modeller har skattats med olika indikatorvariabler.

4.2.1 Dynamiska regressionsmodeller med indikatorvariabler för veckodagar

De dynamiska modeller som presenteras i detta delkapitel är alla strukturerade enligt formel 4.1, där indikatorvariabler för dagar är implementerade.

$$y_t = \beta_0 + \beta_{\text{tis}} x_{1,t} + \beta_{\text{ons}} x_{2,t} + \beta_{\text{tors}} x_{3,t} + \beta_{\text{fre}} x_{4,t} + \beta_{\text{lör}} x_{5,t} + \beta_{\text{sön}} x_{6,t} + \eta_t \tag{4.1}$$

Tabell 4.3 visar hur feltermen blivit modellerad för varje driftområde. Det går att se att alla modeller har en säsongskomponent på sju, vilket innebär att alla feltermer är modellerade med SARIMA. Varje modell är differentierade med ett och har en komponent för glidande medelvärde i den delen av modellen som inte tar hänsyn till säsong.

Tabell 4.3: Dynamiska regressionsmodeller med indikatorvariabler för veckodagar

Driftområde	**Skattad modell för feltermen η_t**
Göteborg/Lindholmen	ARIMA(0,1,2)(0,0,2)$_7$
Hagastaden/City	ARIMA(1,1,3)(2,0,0)$_7$
Kista/Sundbyberg	ARIMA(2,1,2)(1,0,1)$_7$
Malmö	ARIMA(1,1,1)(0,0,2)$_7$
Medborgarplatsen/Liljeholmen	ARIMA(0,1,4)(0,0,2)$_7$
Sickla	ARIMA(1,1,3)(0,0,1)$_7$
Slakthusområdet/Proppen	ARIMA(2,1,1)(0,0,2)$_7$
Slussen	ARIMA(0,1,2)(2,0,0)$_7$
Uppsala	ARIMA(0,1,2)(0,0,1)$_7$

Tabell 4.4 visar Ljung-Box testet för 50 lag. För driftområdena Malmö, Medborgarplatsen/Liljeholmen och Sickla är p-värdet över 0.05 vilket visar på att H_0 inte kan förkastas och residualerna kan då antas vara vitt brus. För de övriga områdena är p-värdet under 0.05 vilket innebär att H_0 kan förkastas och residualerna för dessa driftområden kan därför inte antas vara vitt brus.

Tabell 4.4: Resultat från Ljung-Box test tillhörande modellerna i tabell 4.3

Driftområde	**p-värde**
Göteborg/Lindholmen	0.01
Hagastaden/City	0.00
Kista/Sundbyberg	0.00
Malmö	0.26
Medborgarplatsen/Liljeholmen	0.25
Sickla	0.23
Slakthusområdet/Proppen	0.00
Slussen	0.00
Uppsala	0.08

4.2.2 Dynamiska regressionsmodeller med indikatorvariabler för månader

I detta delkapitel redovisas resultatet från de dynamiska modeller som har indikatorvariabler för månader. Dessa modeller har en regressionstruktur enligt formel 4.2, där η_t är modellerad med ARIMA eller SARIMA.

$$\begin{aligned} y_t = \beta_0 + \beta_{\text{feb}} x_{1,t} + \beta_{\text{mar}} x_{2,t} + \beta_{\text{apr}} x_{3,t} + \beta_{\text{maj}} x_{4,t} + \beta_{\text{jun}} x_{5,t} + \beta_{\text{jul}} x_{6,t} + \\ \beta_{\text{aug}} x_{7,t} + \beta_{\text{sep}} x_{8,t} + \beta_{\text{okt}} x_{9,t} + \beta_{\text{nov}} x_{10,t} + \beta_{\text{dec}} x_{11,t} + \eta_t \end{aligned} \tag{4.2}$$

Från tabell 4.5 går det att urskilja en felterm som är modellerad med ARIMA och detta är för driftområdet Malmö, resterande modeller har säsongskomponenter. Utöver detta har alla modeller en komponent q i den delen av modellen som inte tar hänsyn till säsong.

Tabell 4.5: Dynamiska regressionsmodeller med indikatorvariabler för månader

Driftområde	**Skattad modell för feltermen η_t**
Göteborg/Lindholmen	ARIMA(3,1,1)(2,0,0)$_7$
Hagastaden/City	ARIMA(0,0,1)(2,1,2)$_7$
Kista/Sundbyberg	ARIMA(0,0,3)(0,1,2)$_7$
Malmö	ARIMA(5,1,1)(0,0,0)
Medborgarplatsen/Liljeholmen	ARIMA(0,0,1)(0,1,1)$_7$
Sickla	ARIMA(2,1,2)(0,0,1)$_7$
Slakthusområdet/Proppen	ARIMA(3,1,1)(2,0,0)$_7$
Slussen	ARIMA(0,1,2)(2,0,0)$_7$
Uppsala	ARIMA(2,1,1)(0,0,2)$_7$

Tabell 4.6 visar resultatet från Ljung-Box testet för 50 lag. Endast residualerna för Kista/Sundbyberg kan antas vara vitt brus då p-värdet är över 0.05. För de övriga driftområdena är p-värdet mindre än 0.05 och residualrna kan därför inte antas vara vitt brus.

Tabell 4.6: Resultat från Ljung-Box test tillhörande modellerna i tabell 4.5

Driftområde	**p-värde**
Göteborg/Lindholmen	0.00
Hagastaden/City	0.00
Kista/Sundbyberg	0.24
Malmö	0.00
Medborgarplatsen/Liljeholmen	0.00
Sickla	0.00
Slakthusområdet/Proppen	0.00
Slussen	0.00
Uppsala	0.00

4.2.3 Dynamiska regressionsmodeller med indikatorvariabler för veckodagar och månader

De dynamiska modellerna i detta delkapitel har inkluderat indikatorvariabler både för dagar och månader. Detta innebär att alla modeller följer strukturen från formel 4.3, där η_t modelleras med ARIMA eller SARIMA. Resultatet för hur η_t har skattats presenteras i tabell 4.7.

$$\begin{aligned} y_t = \beta_0 + \beta_{\text{tis}} x_{1,t} + \beta_{\text{ons}} x_{2,t} + \beta_{\text{tors}} x_{3,t} + \beta_{\text{fre}} x_{4,t} + \beta_{\text{lör}} x_{5,t} + \beta_{\text{sön}} x_{6,t} + \\ \beta_{\text{feb}} x_{7,t} + \beta_{\text{mar}} x_{8,t} + \beta_{\text{apr}} x_{9,t} + \beta_{\text{maj}} x_{10,t} + \beta_{\text{jun}} x_{11,t} + \beta_{\text{jul}} x_{12,t} + \\ \beta_{\text{aug}} x_{13,t} + \beta_{\text{sep}} x_{14,t} + \beta_{\text{okt}} x_{15,t} + \beta_{\text{nov}} x_{16,t} + \beta_{\text{dec}} x_{17,t} + \eta_t \end{aligned} \tag{4.3}$$

I tabell 4.7 visar att det finns två feltermer som inte har någon säsongskomponent och är därför modellerade med ARIMA, dessa modeller tillhör driftområde Malmö och Slussen. Alla modeller innehåller komponenten q och har differentierats med ett.

Tabell 4.7: Dynamiska regressionsmodellermed indikatorvariabler för veckodagar och månader

Driftområde	**Skattad modell för feltermen η_t**
Göteborg/Lindholmen	ARIMA(0,1,1)$(0,0,2)_7$
Hagastaden/City	ARIMA(2,1,3)$(1,0,0)_7$
Kista/Sundbyberg	ARIMA(0,1,2)$(0,0,2)_7$
Malmö	ARIMA(0,1,2)(0,0,0)
Medborgarplatsen/Liljeholmen	ARIMA(0,1,4)$(0,0,2)_7$
Sickla	ARIMA(1,1,3)$(0,0,1)_7$
Slakthusområdet/Proppen	ARIMA(3,1,1)$(0,0,2)_7$
Slussen	ARIMA(2,1,4)(0,0,0)
Uppsala	ARIMA(0,1,2)$(0,0,1)_7$

Tabell 4.8 visar resultatet från Ljung-Box testet för 50 lag. Endast för Medborgarplatsen/Liljeholmen är p-värdet över 0.05 vilket innebär att residualerna kan antas vara vitt brus. För de övriga områdena är p-värdet under 0.05 vilket innebär att residualerna inte kan antas vara vitt brus.

Tabell 4.8: Resultat från Ljung-Box test tillhörande modellerna i tabell 4.7

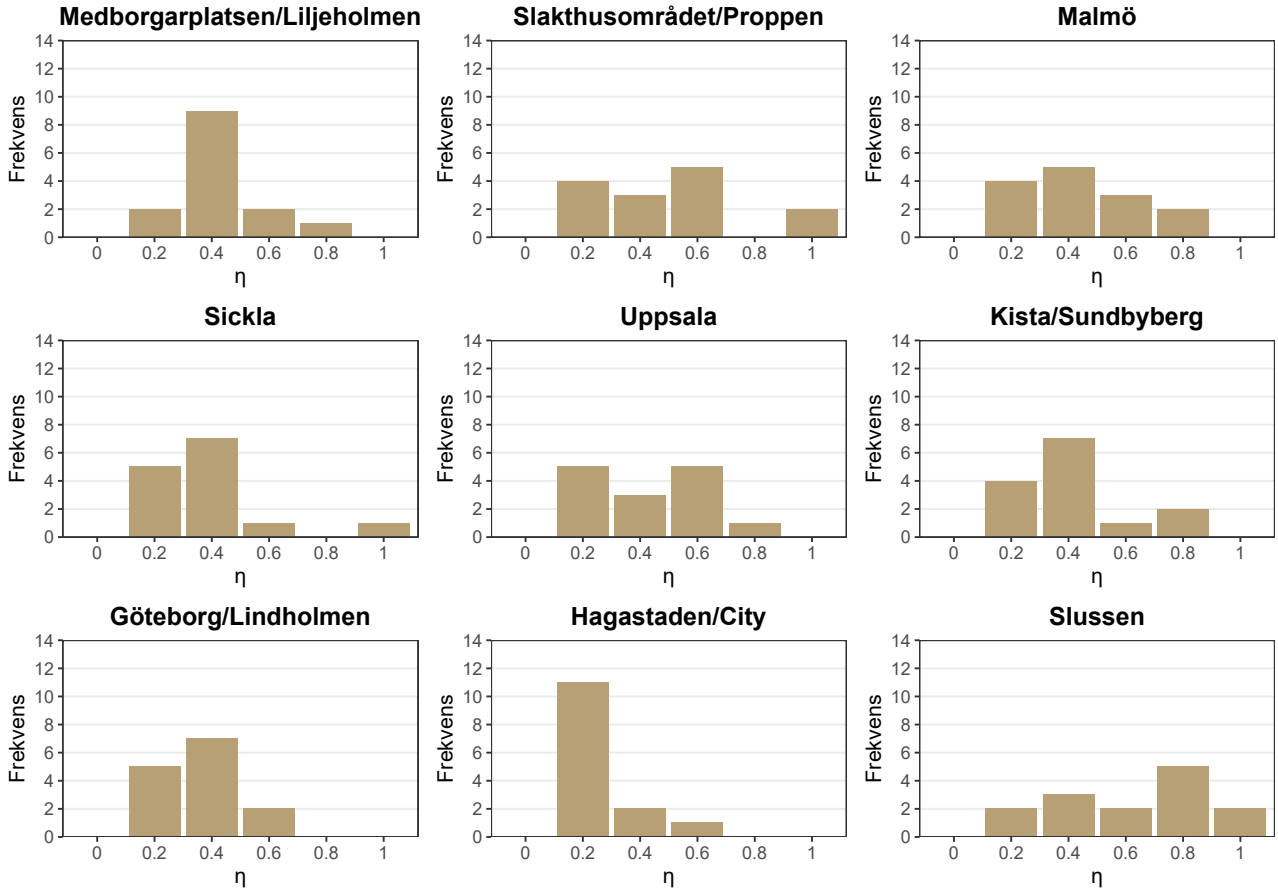
Driftområde	**p-värde**
Göteborg/Lindholmen	0.00
Hagastaden/City	0.00
Kista/Sundbyberg	0.00
Malmö	0.01
Medborgarplatsen/Liljeholmen	0.07
Sickla	0.01
Slakthusområdet/Proppen	0.00
Slussen	0.00
Uppsala	0.01

4.3 XGBoost

I detta kapitel presenteras resultat från XGBoost-modeller. Det skattades modeller för varje specifik prognoshorisont och driftområde, vilket medför att det totalt skapades 126 XGBoost-modeller. Data som används för dessa modeller består av laggade y samt indikatorvariabler för veckodagar och månader. Denna datahantering förklarades mer ingående i avsnitt 2.4.

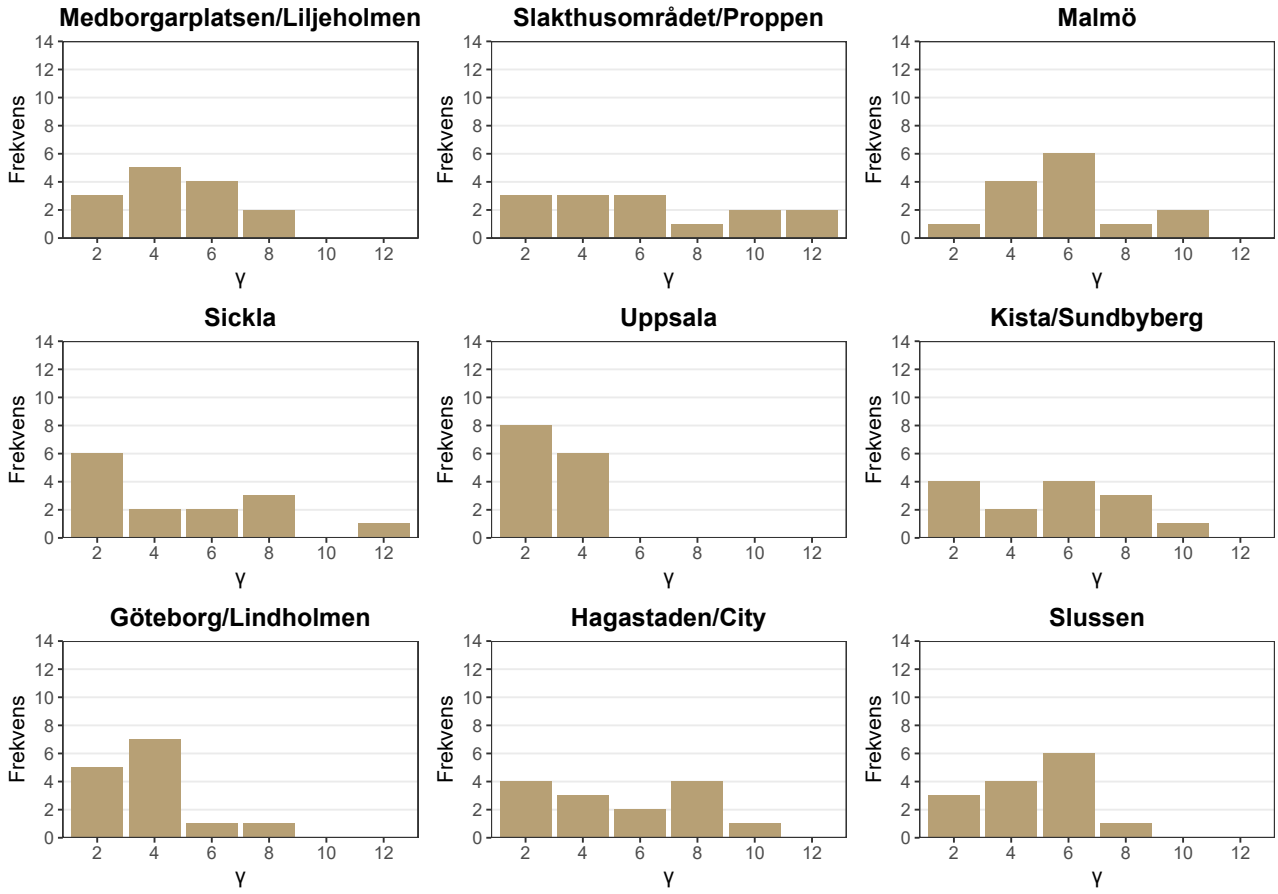
4.3.1 Parametertuning

För varje driftområde bildas en modell för varje prognoshorisont, detta resulterar i 126 olika modeller. Fördelningen av optimal hyperparameter för respektive område visas i följande diagram.



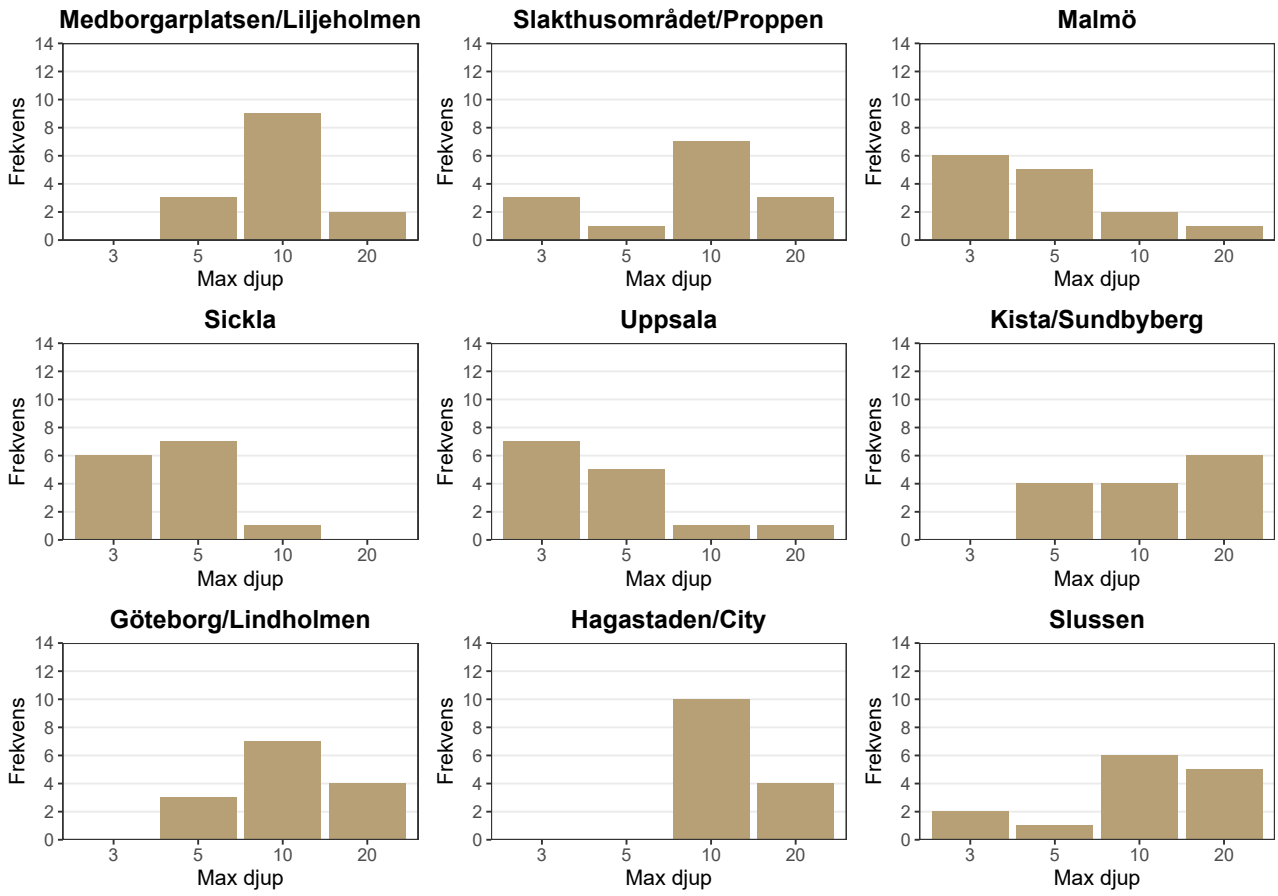
Figur 4.1: Fördelningen för skattningar av η

Figur 4.1 visar fördelningen för hyperparametervalet för η i de olika driftområdena. Ur diagrammen går det att tyda att det vanligaste värdet för η är mellan 0.2 till 0.6 samt att i inget av driftområdena är 0 det optimala värdet för η. Det går även att se att väldigt få modeller har valt värden 1 för hyperparametern.



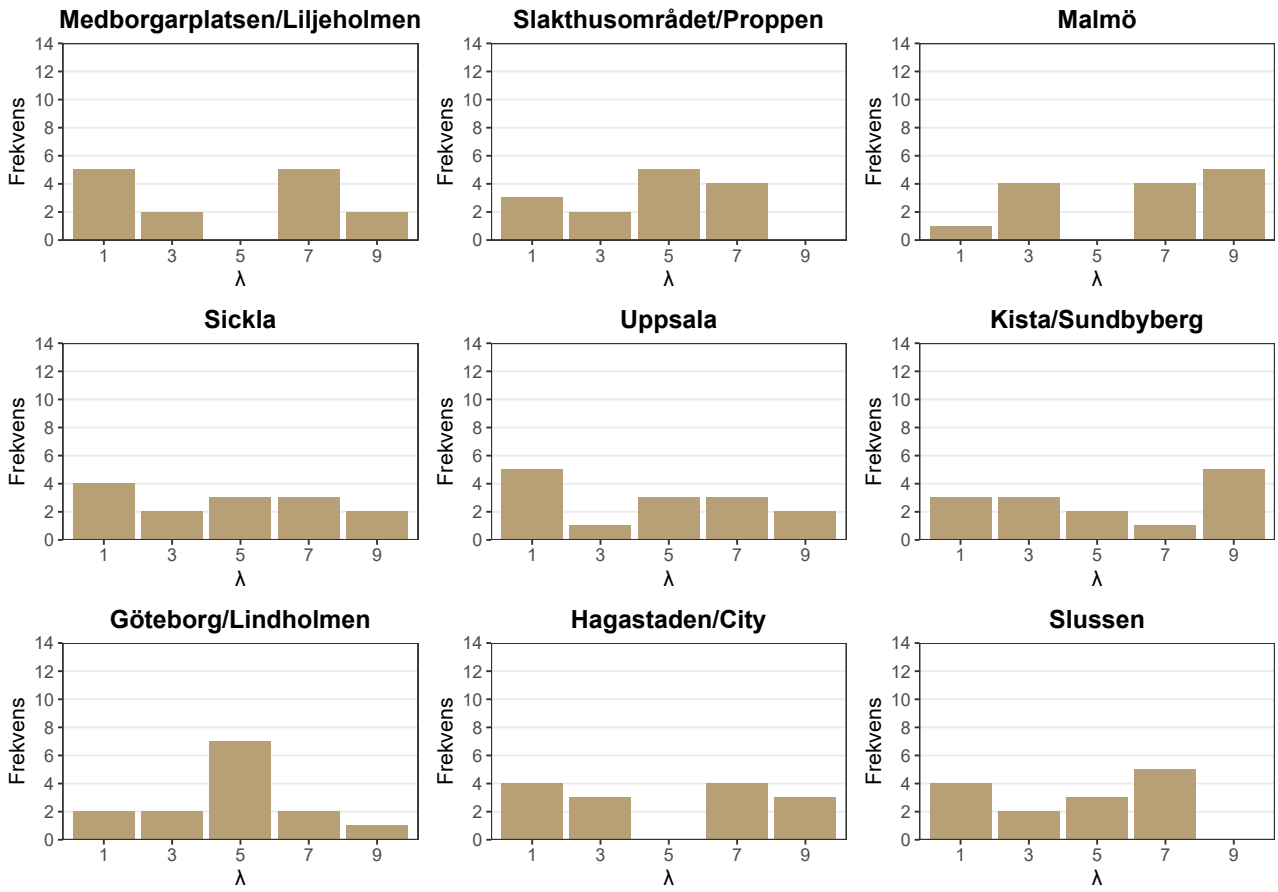
Figur 4.2: Fördelningen för skattningar av γ

Figur 4.2 visar hur valet av hyperparametern γ såg ut inom respektive driftområde. Diagrammen visar att γ varierade mellan olika prognoshorisonter, där det ofta var låga värden på parametern som ansågs vara mest optimala. Det var endast 3 modeller som valde $\gamma = 12$ och 6 modeller som valde $\gamma = 10$. Fördelningen av parametervärden skiljer sig något mellan driftområdena, där exempelvis Uppsalas parametervärden endast var värden på 2 och 4 medan det i Slakthusområdet/Proppen var en jämn fördelning mellan de olika värdena.



Figur 4.3: Fördelningen för skattningar av maxdjup

Figur 4.3 visar hur värdena för maxdjupet fördelar sig i de olika driftområdena. Det optimala parametervärdet varierade mellan både prognoshorisont och driftområde, detta visas i figurerna då fördelningarna av max djupet skiljer sig.



Figur 4.4: Fördelningen för skattningar av λ

Figur 4.3 visar fördelningen för hyperparametervalet för λ. Det är en relativt jämn fördelning av de olika parametervärdena i de flesta driftområden, där ett undantag är Göteborg/Lindholmen där flest modeller använde $\lambda = 5$ som det mest optimala. Att värdena varierade i driftområdena innebär att det varierade mellan olika prognoshorisonter.

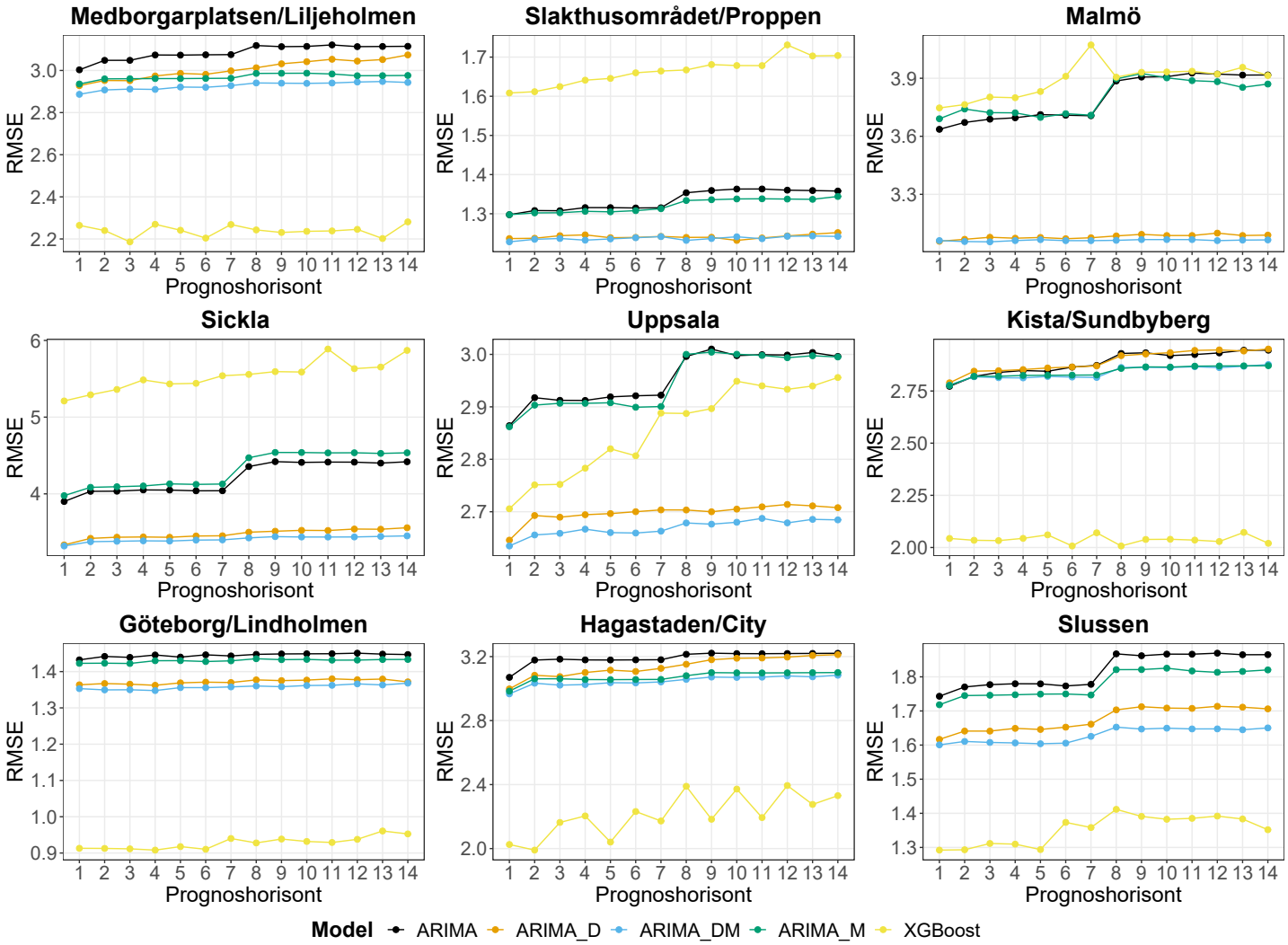
4.4 Modelljämförelse

I detta kapitel kommer de olika modellerna jämföras med varandra för att senare kunna välja den mest optimala modellen för varje specifikt driftområde. Detta görs genom att jämföra RMSE mellan de olika modellerna, där ett lågt RMSE innebär att modellen skapar skattningar nära det observerade värdet. RMSE har beräknats med anpassningar enligt 3.6.1. Tabell 4.9 presenterar ett förtydligande av vilka modeller som representeras i legenden för figur 4.5 och figur 4.6.

Tabell 4.9: Förtydligande till legenden i figur 4.6 och 4.5

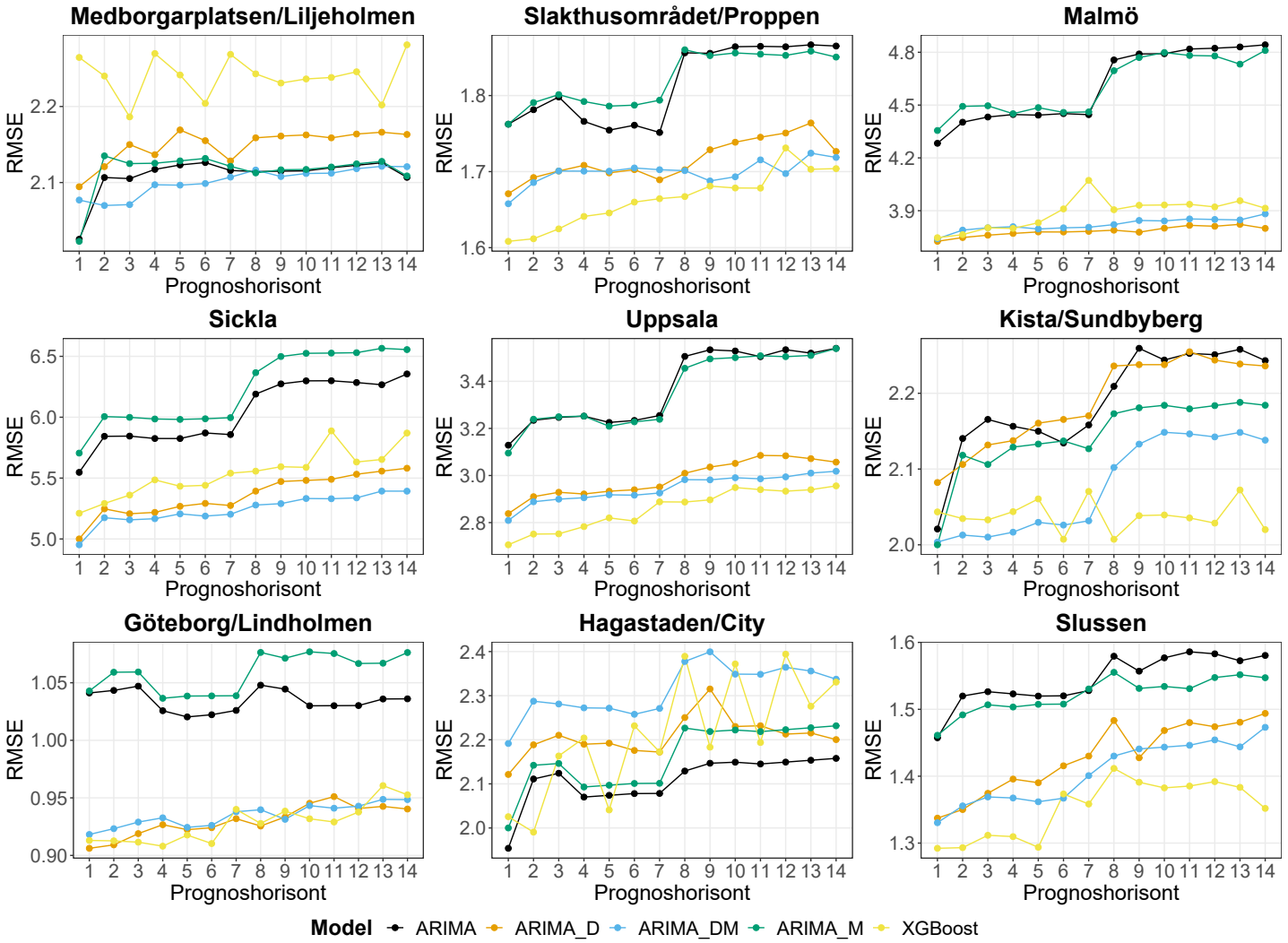
Legendförklaring	**Modell**
ARIMA	ARIMA eller SARIMA enligt formel 3.5
ARIMA_D	Dynamisk regressionsmodell enligt formel 4.1
ARIMA_{DM}	Dynamisk regressionsmodell enligt formel 4.3
ARIMA_M	Dynamisk regressionsmodell enligt formel 4.2
XGBoost	XGBoost modell enligt formel 3.15

För att simplifiera jämförelser av de olika modellerna för respektive driftområde skapades diagram som visar RMSE över prognoshorisonterna. Diagrammen ger en även tydligare bild av hur RMSE varierar mellan prognoshorisonterna för de enskilda modellerna och underlättar därför att se förändring i RMSE. Figur 4.5 visar beräknat RMSE i träningsdata för alla modeller i respektive område.



Figur 4.5: Genomsnittligt RMSE för varje prognoshorisont uppdelat på driftområde med träningsdata

I figur 4.5 går det att urskilja att att XGBoost modellen har lägst RMSE för alla prognoshorisonter för fem av de nio driftområdena, dessa är Medborgarplatsen/Liljeholmen, Kista/Sundbyberg, Göteborg/Lindholmen, Hagastaden/City och Slussen. För de övriga driftområdena har de dynamiska modellerna med indikatorvariabler för veckodagar eller veckodagar och månader lägst RMSE. För de flesta modeller och driftområden ligger RMSE på ett relativt konstant värde över prognoshorisonterna. Dock går det att se att i exempelvis Sickla, Uppsala och Hagastaden/City ökar RMSE för XGBoost med längre prognoshorisont. Eftersom att driftområdena har olika genomsnittligt antal ärenden per dag, skiljer sig skalan på RMSE.



Figur 4.6: Genomsnittligt RMSE för varje prognoshorisont uppdelat på driftområde med valideringsdata

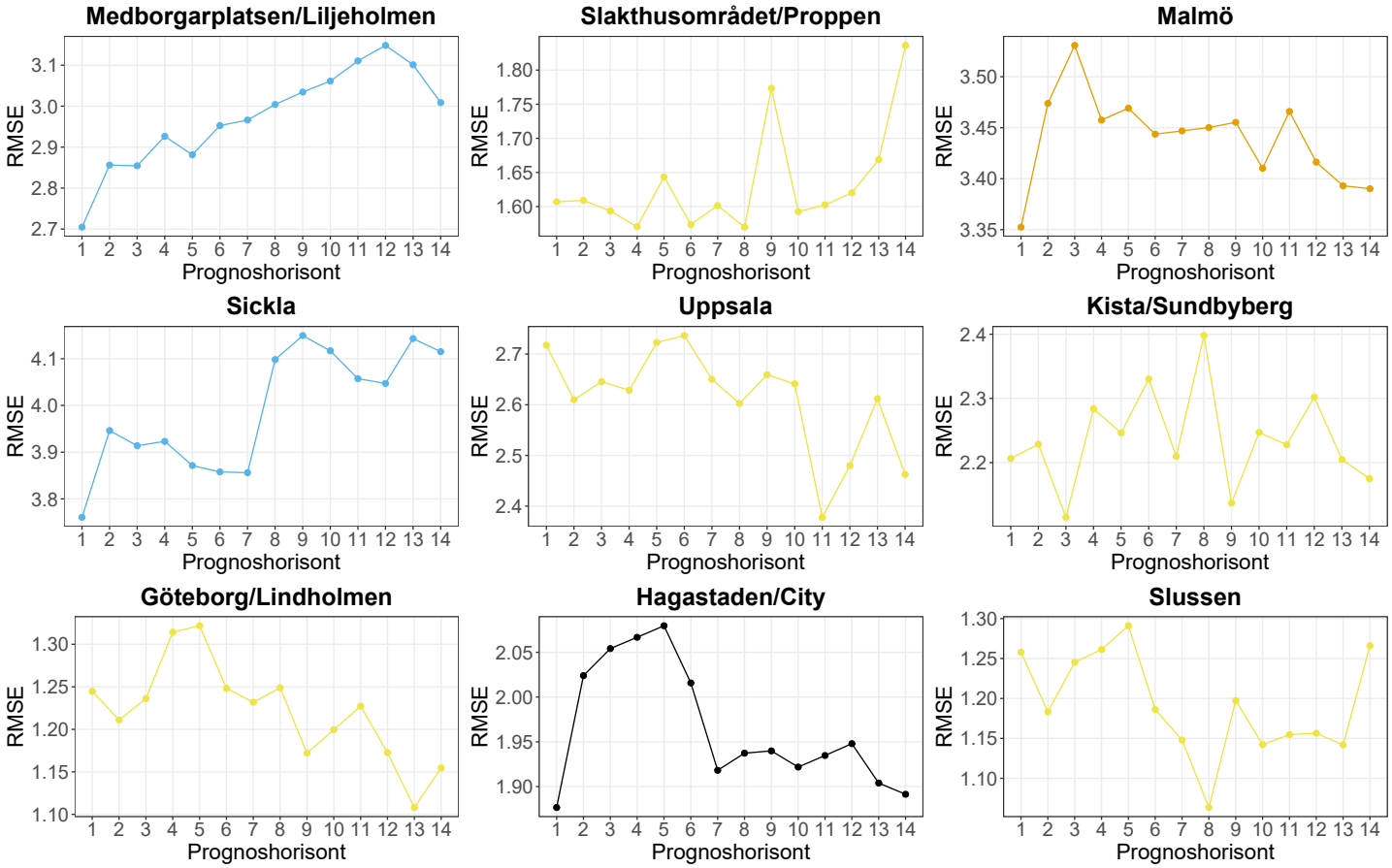
Figur 4.6 visar RMSE för samtliga driftområden beräknat på valideringsdata. Ur diagrammet går det att se ett ökande RMSE vid längre prognoshorisonter för de flesta driftområden. Det går även att se att RMSE varierar mellan de olika områdena på grund av den varierande ärendevolymen. I de flesta driftområdena kan även ett starkt ökande RMSE mellan prognoshorisont 7 och 8 för ARIMA och ARIMA_M tydas. ARIMA_{DM}, ARIMA_D och XGBoost verkar alla få ett snarlikt RMSE i de flesta driftområdena, där de områden som skiljer sig från detta är Medborgarplatsen/Liljeholmen och Hagastaden/City. I Hagastaden/City är RMSE för XGBoost väldigt oregelbunden till skillnad från hur den rör sig i de andra driftområdena. Modellerna med lägst RMSE i respektive driftområde visas i tabell 4.10.

Tabell 4.10 presenterar de modeller som har lägst RMSE för flest prognoshorisonter i figur 4.6. Det går att se att XGBoost presterar bäst utifrån RMSE i majoriteten av driftområdena och därefter de dynamiska regressionsmodellerna och sist enkel SARIMA.

Tabell 4.10: Modeller med lägst RMSE för respektive driftområde med valideringsdata

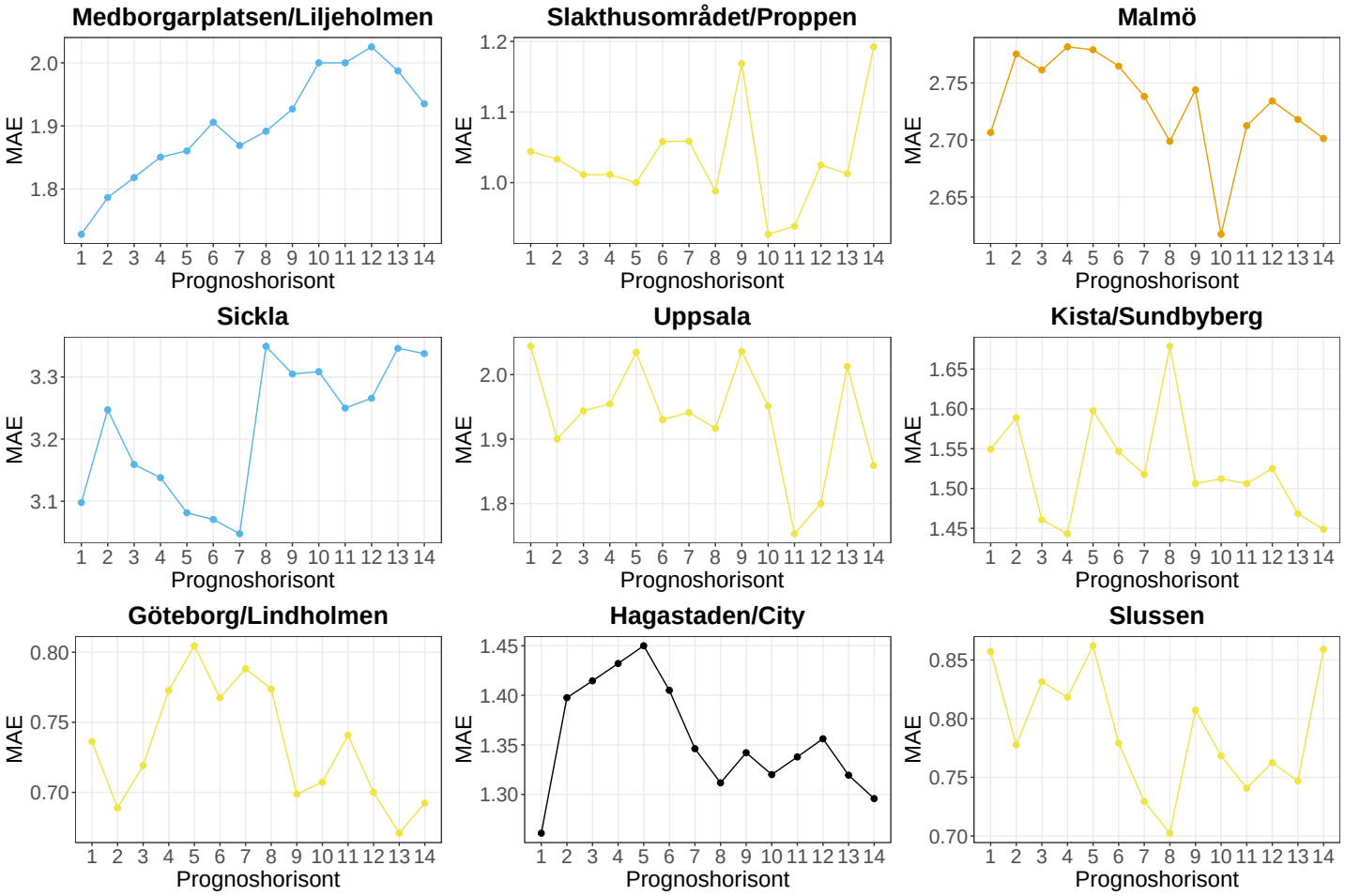
Driftområde	**Modell med lägst RMSE**
Göteborg/Lindholmen	XGBoost
Hagastaden/City	SARIMA
Kista/Sundbyberg	XGBoost
Malmö	Dynamisk regression med indikatorvariabler för veckodagar
Medborgarplatsen/Liljeholmen	Dynamisk regression med indikatorvariabler för veckodagar och månader
Sickla	Dynamisk regression med indikatorvariabler för veckodagar och månader
Slakthusområdet/Proppen	XGBoost
Slussen	XGBoost
Uppsala	XGBoost

För att vidare utföra tydligare tolkningar av hur felskattningarna ser ut för de valda modellerna användes måttet MAE. MAE visar det genomsnittliga absoluta felet för modellers skattningar. I figur 4.7 visas beräknat RMSE för testmängden och i figur 4.8 visas det beräknade MAE för varje prognoshorisont i respektive driftområde.



Figur 4.7: Beräknat RMSE för testmängden med modellerna från tabell 4.10

Figur 4.7 visar hur RMSE fördelar sig i de olika driftområdena beroende på prognoshorisont. Det går att se att RMSE skiljer sig något ifrån det RMSE som beräknades med valideringsmängden. RMSE varierar även mer mellan prognoshorisonter än vad det gjorde i figur 4.6. För majoriteten av driftområdena har XGBoost lägst RMSE för flest prognoshorisonter.



Figur 4.8: Beräknat MAE för testmängden med modellerna från tabell 4.10

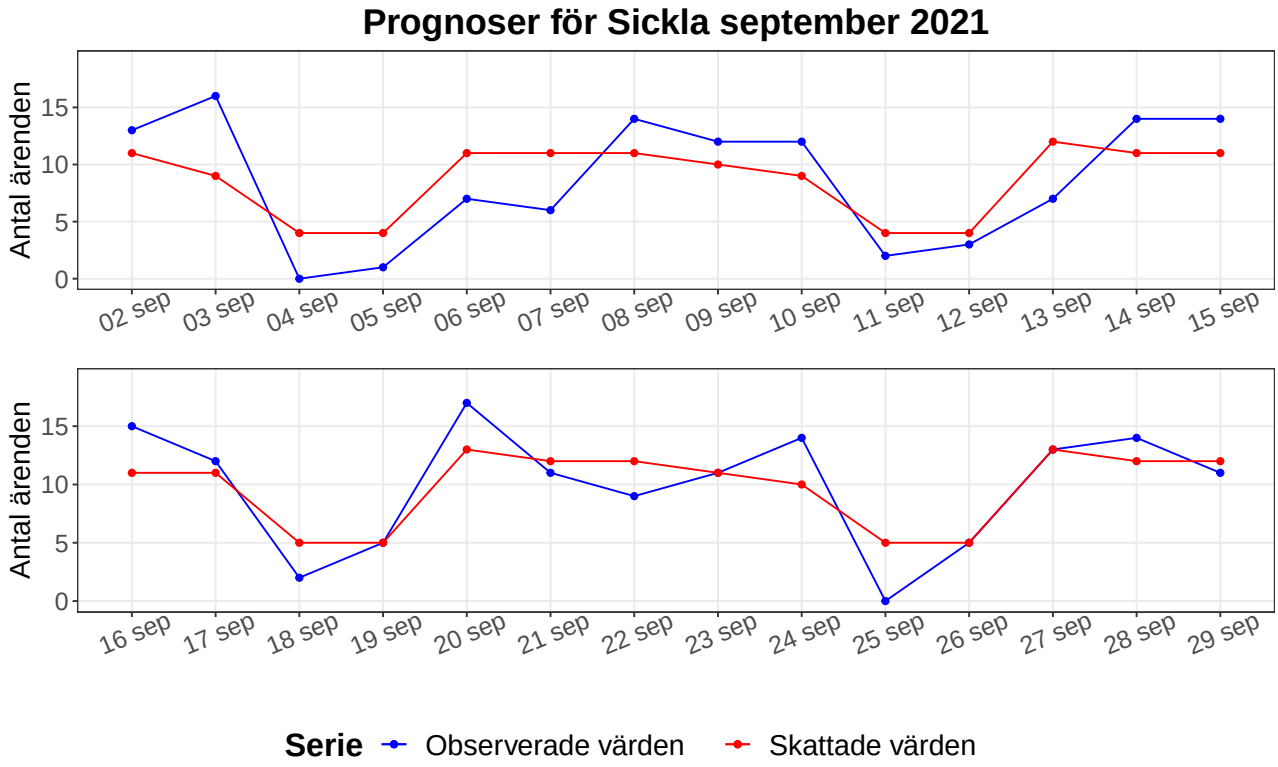
Figur 4.8 visar MAE beräknat med testdata för de modellerna med lägst RMSE för valideringsdata. Det går att se tydligt att MAE varierar kraftigt mellan de olika driftområdena. För Medborgarplatsen/Liljeholmen där ärendevolymen i genomsnitt är 5.52 ärenden per dag så predikterar den dynamiska regressionsmodellen med indikatorvariabler för veckodagar och månader i genomsnitt mellan 1.72 och 2.02 ärenden fel. För Slakthusområdet/Proppen där ärendevolymen i genomsnitt är 1.92 ärenden per dag så skattar XGBoost i genomsnitt mellan 0.92 och 1.19 ärenden fel. För Malmö där ärendevolymen i genomsnitt är 7.33 ärenden per dag så skattar den dynamiska regressionsmodellen med indikatorvariabler för veckodagar i genomsnitt mellan 2.74 och 2.83 ärenden fel. För Sickla där ärendevolymen i genomsnitt är 9.05 ärenden per dag så skattar den dynamiska regressionsmodellen med indikatorvariabler för veckodagar och månader i genomsnitt mellan 3.05 och 3.5 ärenden fel. För Uppsala där ärendevolymen i genomsnitt är 5.59 ärenden per dag så skattar XGBoost modellen mellan 1.75 och 2.04 ärenden fel. För Kista/Sundbyberg där ärendevolymen i genomsnitt är 2.66 ärenden per dag så skattar XGBoost modellen mellan 1.44 och 1.68 ärenden fel. För Göteborg/Lindholmen där ärendevolymen i genomsnitt är 1.03 ärenden per dag så skattar XGBoost modellen i genomsnitt mellan

0.67 och 0.80 ärenden fel. I Hagastaden/City där ärendevolymen i genomsnitt är 2.93 ärenden per dag så skattar ARIMA-modellen mellan 1.26 och 1.45 ärenden fel. Samt i Slussen där ärendevolymen i genomsnitt är 1.49 så skattar XGBoost modellen mellan 0.7 och 0.86 ärenden fel.

4.5 Prognoser

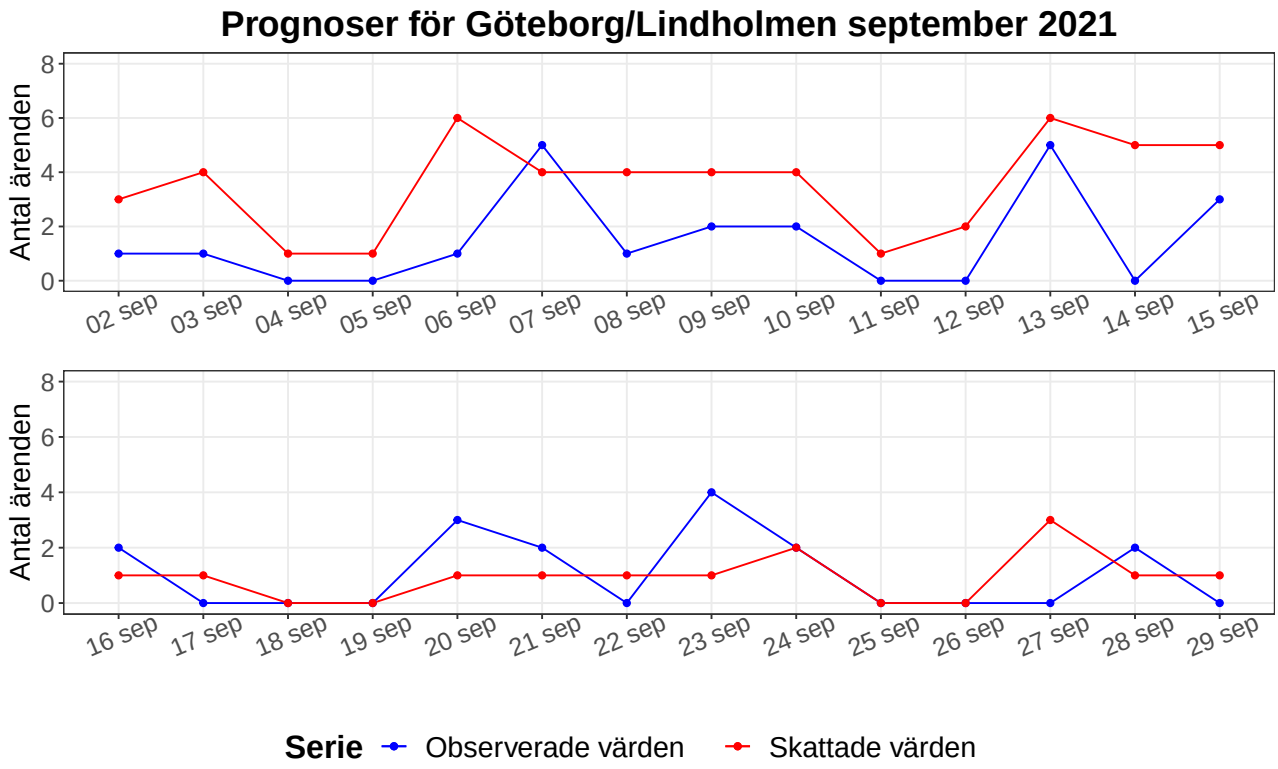
I detta kapitel visas ett två exempel på prognoser som har skapats med de skattade modellerna. Resterande prognoser visas i bilaga 2.

När de mest lämpliga modellerna för respektive driftområde valts används dessa för att skapa prognoser på testmängden, detta för att visa exempel på hur en prognos kan se ut i praktiken. Prognoser har skapats för två perioder som innehåller 14 dagar vardera, där den första perioden har startdatum 1 september 2021 och den andra perioden har startdatum 15 september 2021.



Figur 4.9: Prognoser för Sickla under september 2021 med dynamisk regression innehållande indikatorvariabler för veckodagar och månader

Figur 4.9 visar observationerna och hur den dynamiska regressions modellen med indikatorvariabler för veckodagar och månader har skapat prognoser i Sickla för de två perioderna. Denna modell skapar prediktioner relativt nära det observerade värdet för de flesta dagar, men har svårt att fånga de kraftiga förändringarna.



Figur 4.10: Prognoser för Göteborg/Lindholmen under september 2021 med XGBoost

Figur 4.10 visar prognoserna och observationerna för antalet ärenden i Göteborg/Lindholmen under september 2019. Under perioden 2 september till 15 september i nästan alla fall skapar prognoser högre än de faktiska antalen. Medan modellen under perioden 16 september till 29 september prognostiserat mer spritt.

5. Analys

Detta kapitel syftar till att analysera resultatet som sedan kan användas för att stötta slutsatser.

Genom att jämföra RMSE för modellerna som skapades för varje driftområde kunde den mest lämpliga modellen utses. Den mest lämpliga modellen valdes genom att undersöka vilken av de fem modellerna som hade lägst RMSE för flest prognoshorisonter i valideringsmängden, där alla RMSE presenterades i figur 4.6. Modellerna som valts för respektive driftområde presenterades i tabell 4.10. Tabell 4.10 visade att XGBoost har presterat bäst i de flesta fall och därefter den dynamiska regressionsmodellen med två typer av indikatorvariabler. Det var endast ett driftområde som hade dynamisk regression med indikatorvariabler för veckodagar som den mest lämpliga modellen och ett driftområde som hade en enkel SARIMA-modell som mest lämpliga modell.

Vid jämförelser av RMSE för de olika modellerna blir det tydligt att skillnaderna är marginella. I driftområdet Medborgarplatsen/Liljeholmen skiljde det endast 0.19 mellan det högsta och lägsta RMSE. Även i Sickla skiljde det inte mer än 1.04 i RMSE mellan den modell med lägst och högst värde. Eftersom det visade sig vara marginella skillnader mellan de olika typerna av modeller skulle de flesta modeller fungera för att skapa prognoser för serviceärenden. Genom att jämföra graferna i figur 4.5 och figur 4.6 kunde överanpassning undersökas. I figur 4.5 har modellerna använt sig av träningsmängden för att skapa prognoser, som också är den mängd modellerna byggdes med, och i figur 4.6 har modellerna använt sig av valideringsmängden för att skapa prognoser. Om ett kraftigt ökande RMSE sker från prognoser med träningsmängden till prognoser med valideringsmängden kan överanpassning misstänkas. Ingen av modellerna verkar överanpassat sig till träningsmängden, dock går det att se att XGBoost-modellen är den enda som inte har märkbara förändringar i RMSE mellan de två datamängderna. Detta tyder på att modellen antingen skapar bra skattningar eller är något underanpassad vilket gör att den eventuellt skulle kunna förbättras.

I figur 4.8 visades beräknat MAE för testmängden med de modeller som ansågs vara bäst lämpad för respektive driftområde. Detta mått ska ge förståelse för hur många ärenden i genomsnitt som modellerna skattar fel. Graferna hade ett varierande mönster vilket beror på att modellerna hade olika god skattningsförmåga för olika prognoshorisonter. Detta kan antingen bero på modellernas skattningsförmåga eller hur datamaterialet i testmängden såg ut. Dock bör det uppmärksammas att det tydligt varierande mönstret i MAE också beror på de smala intervallet av värden som MAE sträcker sig mellan.

Modellerna som ansågs vara mest optimala för Sickla och Hagastaden/City hade ett Ljung-Box test som ej var signifikant vilket innebär att residualerna inte kan antas vara vitt brus. Detta kan betyda att det fortfarande finns en del variation som modellerna inte lyckas förklara.

6. Diskussion

I detta kapitel kommer studiens metod, tillvägagångsätt och resultat att diskuteras och utvärderas.

Metodval

Eftersom att en ARIMA-modell antar att data har en kontinuerlig responsvariabel kan detta skapa problematik när en icke-kontinuerlig variabel användas. Ett förbättringsområde skulle kunna vara att undersöka andra typer av modeller som kan ta bättre hänsyn till detta. Exempelvis Poisson ARIMA, även kallat Integer-valued Autoregressive Model (INAR). Denna modell tar hänsyn till att data endast antar heltal, vilket inte en vanlig ARIMA-modell gör. Det finns fler vetenskapliga artiklar, studier och övrig litteratur som fokuserar på denna typ av modell. Ett exempel på detta är en artikel skriven av Du Jin-Guan och Li Yuan, som diskuterar likheterna mellan INAR modellen och den ursprungliga AR modellen (Jin-Guan & Yuan, 1991). Även andra maskininlärningsmetoder skulle också kunna användas. Ett exempel på detta är Convolutional Neural Network (CNN) som är mest förekommande vid bildklassificering men som också har visat sig vara en bra metod för att kunna prognostisera tidsserier (Géron 2019).

Då RMSE endast varierade marginellt mellan de olika modellerna kan det diskuteras kring relevansen att välja ut en specifik modell som den bästa. Modellerna som undersöktes i denna studie lyckades prognostisera ärendeflödet någorlunda likvärdigt. En rimlig slutsats skulle därför kunna vara att valet av modell inte har så stor betydelse för prognosernas precision. Detta kan också kopplas till att eventuellt andra utvärderingsmått kan implementeras för att jämföra modellerna på flera nivåer. Huruvida modellerna är användbara för företaget i praktiken är svår att fastställa då det är upp till användarens krav på modellens skattningsförmåga. Därför skulle mer mer kunskap om företagets personalfördelning behövas. Först efter det kan kan ledningen ta ett beslut om modellerna ska användas i verksamheten.

Autokorrelerade residualer

Eftersom residualerna i modellen för Sickla respektive Hagastaden/City inte kunde antas vara vitt brus kan det vara aktuellt att välja andra modeller som mest lämpliga för dessa områden. Dock presterar dessa modeller fortfarande bäst utifrån RMSE vilket gör det rimligt att välja dessa. Detta problem uppstår förmodligen på grund av den stora andelen 0:or i data, samt att de låga observerade antalen. När Ljung-box testen genomfördes valdes det att undersöka 50 laggar. Detta kan vara en anledning till att många test inte blev signifikanta eftersom ett stort antal laggar inkluderades i testet.

Transformering av datamaterialet

Antalet ärenden varierade både i antal och på olika sätt mellan de olika driftområdena. Även andelen 0:or skiljde sig mycket mellan driftområden, där det stora antalet 0:or i datamaterialet kan försvåra skapandet av en bra modell. Alternativa sätt att hantera detta diskuterades, exempelvis relevansen i att prediktera 0:or. Där det skulle kunna ses som oanvändbart att få information kring de dagar då det inte sker ett ärenden eftersom företaget förmodligen har personal på plats ändå. Ett tillvägagångsätt för att hantera detta hade varit att öka alla observerade antal med ett ärende för avlägsna alla 0:or. Skulle det även anses vara icke relevant att veta exakt hur många ärenden som inträffar på en specifik dag hade data kunna grupperats i olika intervall. Problemet med att data inte är kontinuerlig hade då undvikits och data hade hanterats som ett klassificeringsproblem istället. Om det skulle vara känt hur många drifttekniker som krävs för att hantera en viss mängd ärenden så skulle intervallen kunna anpassas till detta.

Fler förklarande variabler

Den lilla mängden förklarande variabler som inkluderades i modellerna kan även vara en anledning till att de inte lyckas fånga upp all variation i data. För att förbättra modellernas precision och bättre lyckas förklara ärendeflödet skulle fler förklaringsvariabler kunna inkluderas. Variabler som skulle kunna vara relevanta är olika variabler som beskriver vädret så som vind, temperatur och nederbörd. Vädret skulle kunna orsaka skador och problem runt fastigheterna vilket hade ökat antalet anmälda ärenden. Vid mildare väder hade eventuellt färre ärenden uppstått vilket bidrar till att vädervariablerna hade kunnat förklara ärendeflödet. Detta diskuterades även av i studien kring cykelanvändning i cykelpooler av Joaquín Amat Rodrigo, där det visade sig att XGBoost presterade markant bättre med exogena variabler (Rodrigo 2022).

Andra variabler som också hade kunnat öka precisionen är fastigheternas uthyrningsgrad. Om fastigheterna står tomma är det sannolikt att få ärenden uppkommer, till skillnad från då de är fullt uthyrda. Högre uthyrdningsgrad medför förmodligen mer slitage på byggnaderna vilket i sin tur skulle kunna öka ärendeflödet. Fastigheternas storlek skulle också kunna påverka antalet anmälda ärenden. Det skulle vara rimligt att de fastigheter som är större också får fler ärenden då det finns en större yta som kan behöva service.
Att skapa en variabel som förklarar när det senast utfördes en reparation eller service på en fastighet skulle tillföra information om hur många serviceärenden som inkommer i framtiden. Om en fastighet nyligen genomgått en större reparation, mindre service eller liknande kan det minska att ytterligare problem uppstår inom en snar framtid. Detta skulle kunna vara en faktor till att ärenden sjunker eller ökar under vissa tidsperioder.

Optimering av XGBoost

För XGBoost finns en stor mängd parametrar som kan optimera modellen. Vilka parametar som skall inkluderas i modellen och hur dess värden skall väljas är väldigt tidskrävande, dock är dessa parametrar avgörande för att skapa en bra modell. I denna studie användes fyra stycken parametrar som sedan parametertunings applicerades på för att hitta de mest optimala värdena. En förbättringspotential ligger i att eventuellt byta ut eller lägga till parametrar till XGBoost modellen, men även att söka igenom ett större antal tillåtna värden för parametrarna. Detta skulle ge modellen fler antal kombinationer av parametrar att söka igenom och detta skulle eventuellt resultera i bättre skattningar. Problematiken med detta är att det är väldigt tidskrävande

då det finns ett väldigt stort val av hyperparametrar, en annan typ av metod för parametertuning hade då kunnat användas. I denna rapport användes gridsökning vid parametertuning men det finns även andra typer av optimering för hyperparametrar som kan användas, bland annat randomiserad sökning, Bayes sökning och Halving gridsökning (Pedregosa et al. 2011).

Ett annat kritiskt moment för att optimera XGBoost modeller är att välja rätt förlustfunktion. Eftersom modellen tränas genom att minimera förlustfunktionen är det viktigt att välja en funktion som är anpassad efter vilket typ av problem som skall lösas. Eftersom endast en förlustfunktion som baseras på poissonfördelningen undersöktes i denna uppsats är det svårt att fastställa om någon en annan förlustfunktion hade varit mer lämplig, vilket gör det till en intressant frågeställning att undersöka i framtida studier.

7. Slutsatser

Detta kapitel besvaras uppsatsens frågeställningar genom att presentera slutsatserna.

- Vilken av metoderna ARIMA, dynamisk regression och XGBoost har lägst RMSE för flest prognoshorisonter under en 14 dagars period för respektive driftområde?

I resultatkapitlet presenteras tabell 4.10 som visar den valda modellen för respektive driftområde. Tabellen visar att XGBoost har lägst RMSE för flest prognoshorisonter i Göteborg/Lindholmen, Kista/Sundbyberg, Slakthusområdet/Proppen, Slussen och Uppsala. Den dynamiska regressionsmnodellen med indikatorvariabler för veckodagar och månader har lägst RMSE för flest prognoshorisonter i Medborgarplatsen/Liljeholmen och Sickla. Den dynamiska regressionsmodellen med indikatorvariabler för veckodagar har lägst RMSE för flest prognoshorisonter i Malmö. Samt SARIMA har lägst RMSE för flest prognoshorisonter i Hagastaden/City.

Referenser

Atrium Ljungberg (2022) *Affärsmodell, mål & strategier*
https://www.al.se/om-oss/affarsmodell-mal-strategier/ [2022-03-03]

Bowerman, L.B., O'connel, T.R., Koehler, B.A. (2005). *Forecasting, Time Series, and Regression.* Cengage Learning.

Chen, T. Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* ss. 785-794. https://doi.org/10.1145/2939672.2939785

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., Yuan, J. (2022). Package xgboost. https://cran.r-project.org/web/packages/xgboost/xgboost.pdf

George, M. (2020). *Development of a forecasting model of Indian road traffic scenario to predict road user share, injuries and fatalities.* Masteruppsats. Statistics and Machine Learning. Linköping: Linköping Universitet. http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-166130

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems.* O'Reilly Media, Inc.

Hassani. H., Reza Yeganegi M (2020). Selecting optimal lag order in Ljung–Box test. *Physica A: Statistical Mechanics and its Applications* Vol 541. https://doi.org/10.1016/j.physa.2019.123700

Hyndman, R.J., Anthanasopoulos G. (2021). *Forecasting: Principles and Practice.* 3 uppl., OTexts: Melbourne, Australia. https://OTexts.com/fpp3

Hyndman, J.R., Khandakar Y. (2008). Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Softwate*, 27(3),1-22. https://doi.org/10.18637/jss.v027.i03

Jin-Guan D. & Yuan L. THE INTEGER-VALUED AUTOREGRESSIVE (INAR(p)) MODEL. Journal of Time Series Analysis. Vol 12, ss. 129-142 (1991). https://doi.org/10.1111/j.1467-9892.1991.tb00073.x

Lucas A, Tomlinson T, Rohani N, Chowdhury R, Solla SA, Katsaggelos AK and Miller LE (2019) Neural Networks for Modeling Neural Spiking in S1 Cortex. Front. Syst. Neurosci. 13:13. doi: https://doi.org/10.3389/fnsys.2019.00013

Marcia Baptista, Shankar Sankararaman, Ivo. P. de Medeiros, Cairo Nascimento, Helmut Prendinger, Elsa M.P. Henriques (2018). Forecasting fault events for predictive maintenance using data-driven techniques and ARMA modeling. *Computers Industrial Engineering*, 115, ss. 41-53.
https://doi.org/10.1016/j.cie.2017.10.033.

O'Hara-Wild M., Hyndman R., Wang E,.
https://fable.tidyverts.org/index.html

Svenska statistikfrämjandet (2010) *Svenska statistikfrämjandets etiska kod för statistiker och statistisk verksamhet* https://statistikframjandet.se/wp-content/uploads/2010/12/etisk_kod_final.pdf

Forecasting time series with gradient boosting: Skforecast, XGBoost, LightGBM y CatBoost by Joaquín Amat Rodrigo, available under a Attribution 4.0 International (CC BY 4.0) at
https://www.cienciadedatos.net/documentos/py39-forecasting-time-series-with-skforecast-xgboost-lightgbm-catboost.html

Pedregosa, F. and Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011). Journal of Machine Learning Research, Vol 12. ss. 2825-2830.
https://scikit-learn.org/stable/modules/grid_search.html

Wickham, H., Chang, W., Henry L., Lin Pedersen T., Takahashi, K., Wilke, C., Woo, K., Yutani, H., Dunnington, D., (2021). https://ggplot2.tidyverse.org/

Bilaga 1

Valda XGBoost parametrar för Medborgarplatsen/Liljeholmen

Medborgarplatsen/Liljeholmen				
Prognoshorisont	**Eta**	**Gamma**	**Max djup**	**Lambda**
1	0.80	2	10	9
2	0.20	6	10	1
3	0.40	2	5	9
4	0.60	4	10	1
5	0.20	4	10	3
6	0.40	2	20	7
7	0.40	4	20	1
8	0.40	6	10	7
9	0.40	6	10	7
10	0.40	4	5	7
11	0.40	4	5	7
12	0.60	8	10	3
13	0.40	6	10	1
14	0.40	8	10	1

Valda XGBoost parametrar för Slakthusområdet/Proppen

Slakthusområdet/Proppen				
Prognoshorisont	**Eta**	**Gamma**	**Max djup**	**Lambda**
1	0.40	10	5	7
2	0.60	6	20	3
3	0.20	4	10	1
4	0.20	2	20	7
5	0.40	8	10	1
6	0.60	4	3	7
7	0.40	10	3	1
8	0.20	6	20	7
9	0.60	2	10	5
10	0.60	12	10	5
11	0.60	12	10	5
12	1.00	2	3	5
13	0.20	4	10	5
14	1.00	6	10	3

Valda XGBoost parametrar för Malmö

Malmö				
Prognoshorisont	**Eta**	**Gamma**	**Max djup**	**Lambda**
1	0.40	6	10	1
2	0.40	6	5	9
3	0.40	6	10	7
4	0.20	4	5	3
5	0.40	6	5	9
6	0.60	4	3	3
7	0.80	4	3	7
8	0.40	8	20	9
9	0.80	10	3	9
10	0.20	6	3	3
11	0.20	6	3	3
12	0.20	2	5	7
13	0.60	10	5	7
14	0.60	4	3	9

Valda XGBoost parametrar för Sickla

Sickla				
Prognoshorisont	**Eta**	**Gamma**	**Max djup**	**Lambda**
1	0.40	8	5	1
2	0.40	8	5	1
3	0.20	2	3	5
4	0.20	2	3	9
5	0.20	6	10	1
6	0.40	4	5	3
7	0.40	2	3	7
8	0.40	8	3	3
9	0.20	2	3	1
10	0.40	6	5	7
11	1.00	2	5	9
12	0.60	12	5	7
13	0.40	4	5	5
14	0.20	2	3	5

Valda XGBoost parametrar för Uppsala

Uppsala				
Prognoshorisont	**Eta**	**Gamma**	**Max djup**	**Lambda**
1	0.40	2	5	1
2	0.40	2	5	1
3	0.60	2	3	3
4	0.20	2	3	5
5	0.60	2	3	7
6	0.80	4	3	9
7	0.60	2	3	5
8	0.20	2	5	7
9	0.60	4	3	7
10	0.60	2	3	1
11	0.20	4	5	1
12	0.20	4	20	1
13	0.40	4	10	5
14	0.20	4	5	9

Valda XGBoost parametrar för Kista/Sundbyberg

Kista/Sundbyberg				
Prognoshorisont	**Eta**	**Gamma**	**Max djup**	**Lambda**
1	0.40	6	20	3
2	0.40	6	10	5
3	0.80	10	10	3
4	0.80	2	5	9
5	0.60	4	10	1
6	0.40	2	5	9
7	0.40	8	5	5
8	0.40	2	20	1
9	0.20	6	20	9
10	0.20	8	20	9
11	0.20	6	20	9
12	0.40	2	5	1
13	0.20	8	20	7
14	0.40	4	10	3

Valda XGBoost parametrar för Göteborg/Lindholmen

Göteborg/Lindholmen				
Prognoshorisont	**Eta**	**Gamma**	**Max djup**	**Lambda**
1	0.20	8	20	7
2	0.40	2	10	5
3	0.20	2	10	5
4	0.40	4	10	3
5	0.40	2	20	9
6	0.20	2	20	3
7	0.60	4	10	5
8	0.60	4	10	5
9	0.20	4	10	1
10	0.40	4	5	5
11	0.40	4	5	5
12	0.40	2	10	1
13	0.20	6	20	5
14	0.40	4	5	7

Valda XGBoost parametrar för Hagastaden/City

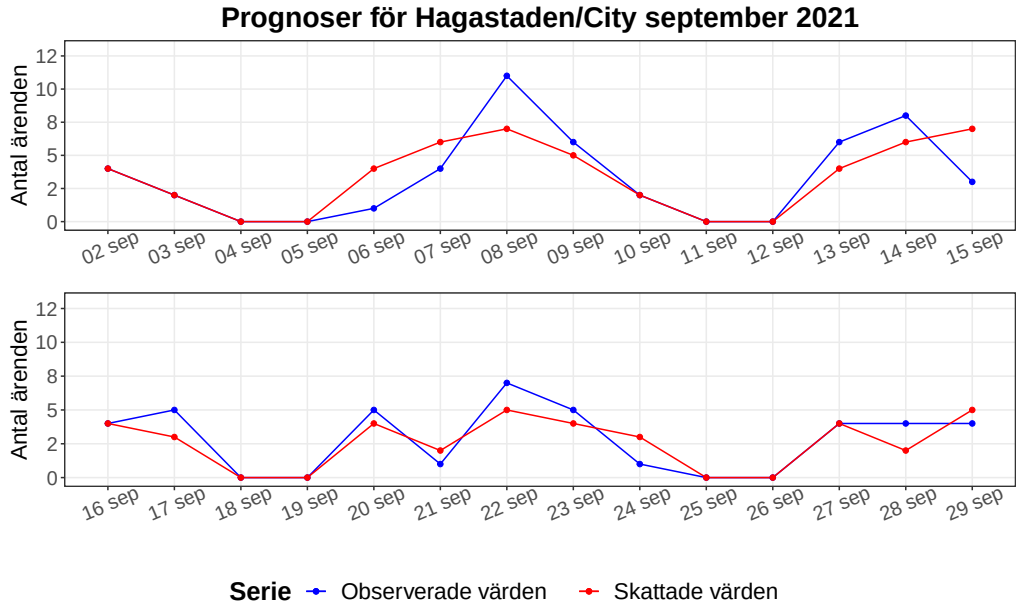
Hagastaden/City				
Prognoshorisont	**Eta**	**Gamma**	**Max djup**	**Lambda**
1	0.20	2	10	1
2	0.20	4	20	7
3	0.20	6	10	7
4	0.20	8	10	1
5	0.40	2	10	3
6	0.20	6	10	7
7	0.40	8	10	9
8	0.20	4	20	1
9	0.20	2	20	3
10	0.60	10	10	3
11	0.20	2	10	7
12	0.20	8	10	9
13	0.20	4	10	9
14	0.20	8	20	1

Valda XGBoost parametrar för Slussen

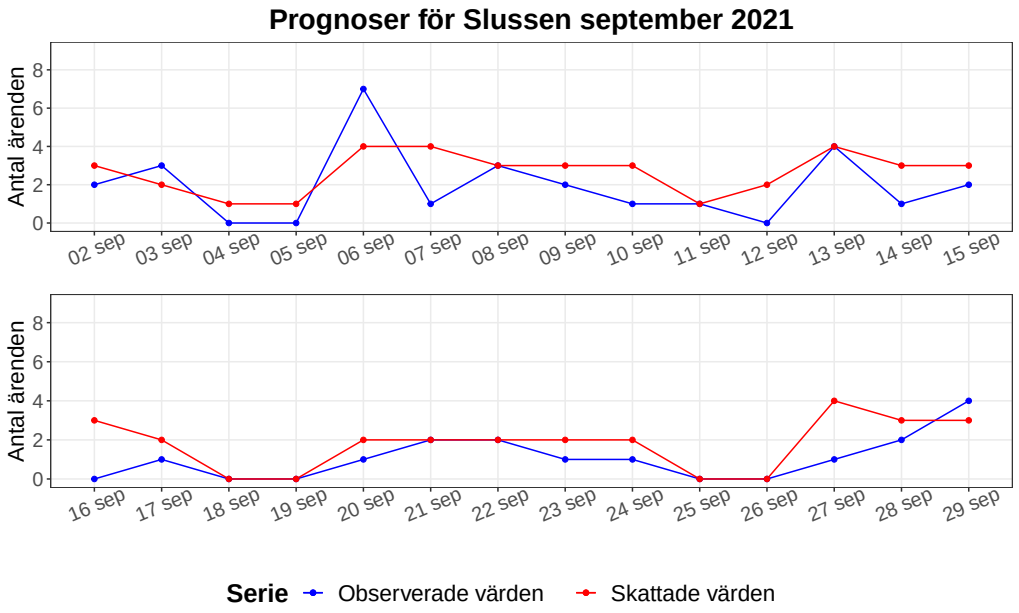
Slussen				
Prognoshorisont	**Eta**	**Gamma**	**Max djup**	**Lambda**
1	0.20	2	5	5
2	0.80	6	10	7
3	0.80	6	20	7
4	0.40	4	10	7
5	0.80	6	20	3
6	0.80	4	10	7
7	1.00	2	3	5
8	0.20	6	20	1
9	0.40	6	10	5
10	0.60	6	10	7
11	0.40	4	10	3
12	0.80	4	20	1
13	0.60	8	20	1
14	1.00	2	3	1

Bilaga 2

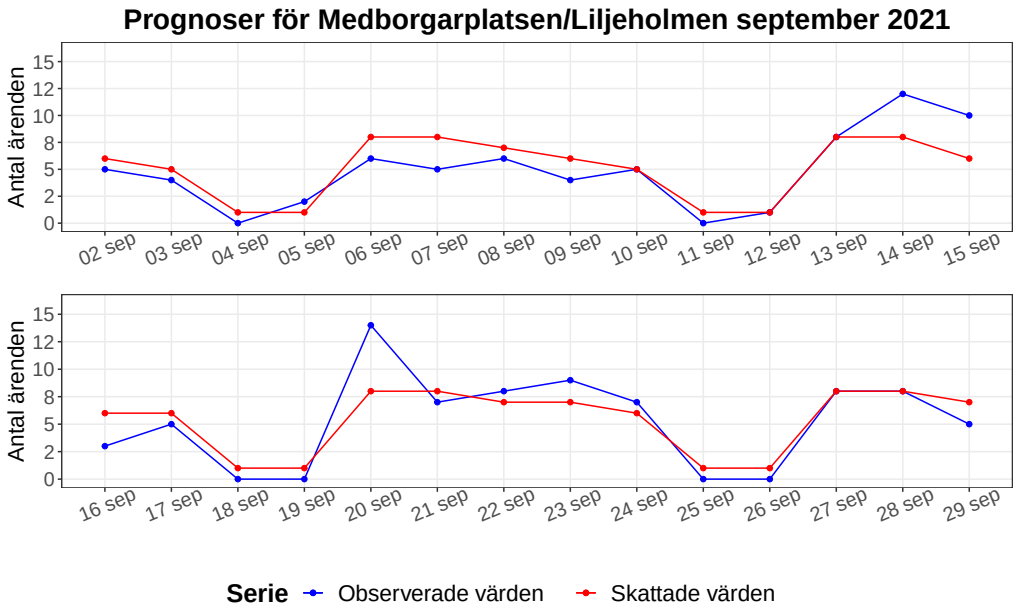
Bilaga 2 presenterar prognoser under september 2021 för de resterande driftområdena som inte presenterades i resultatkapitlet.



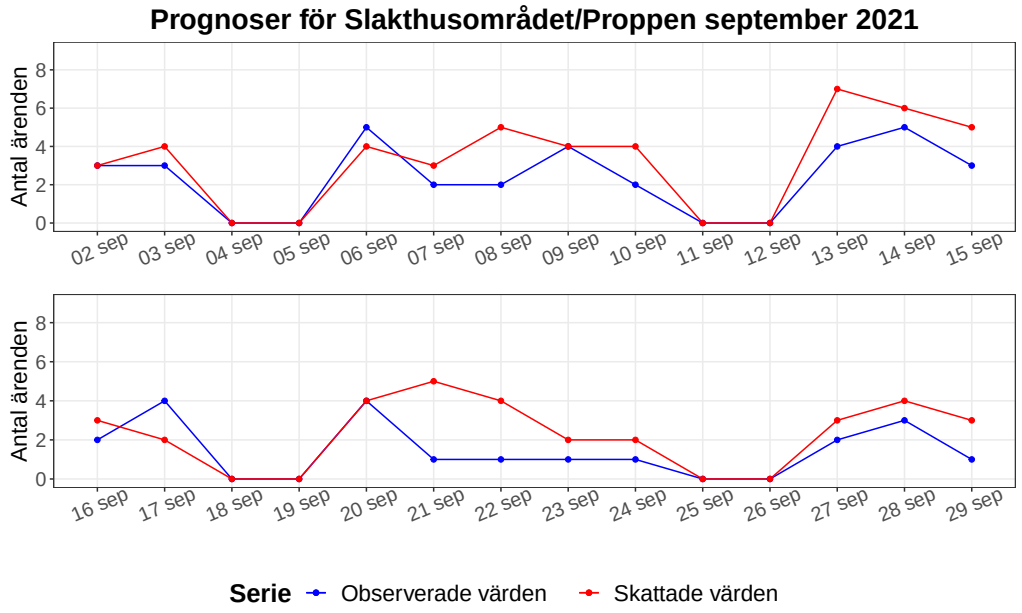
Figur 1: Prognoser för Hagastaden/City under september 2021 med SARIMA



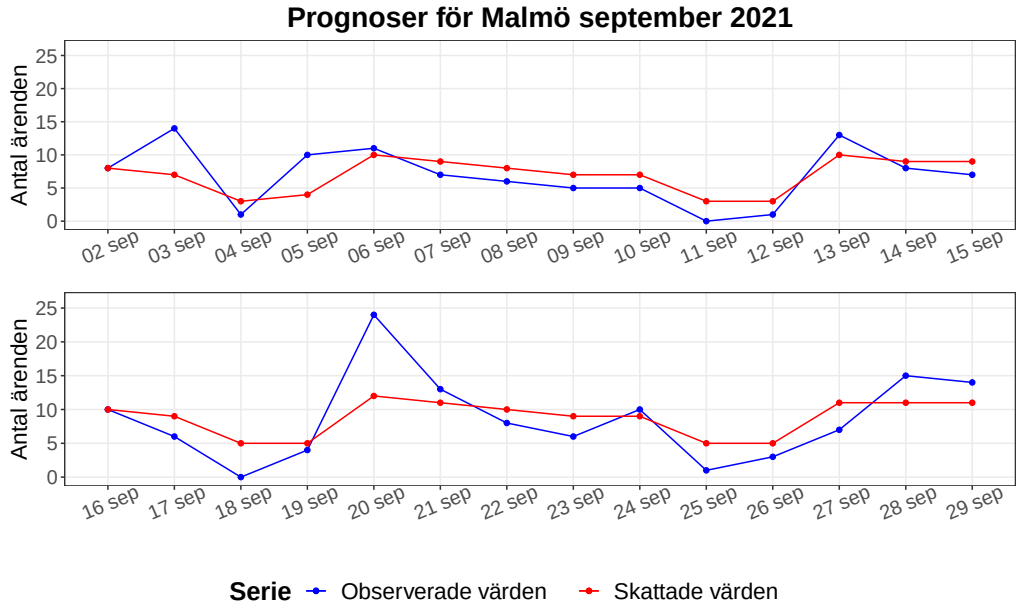
Figur 2: Prognoser för Slussen under september 2021 med XGBoost



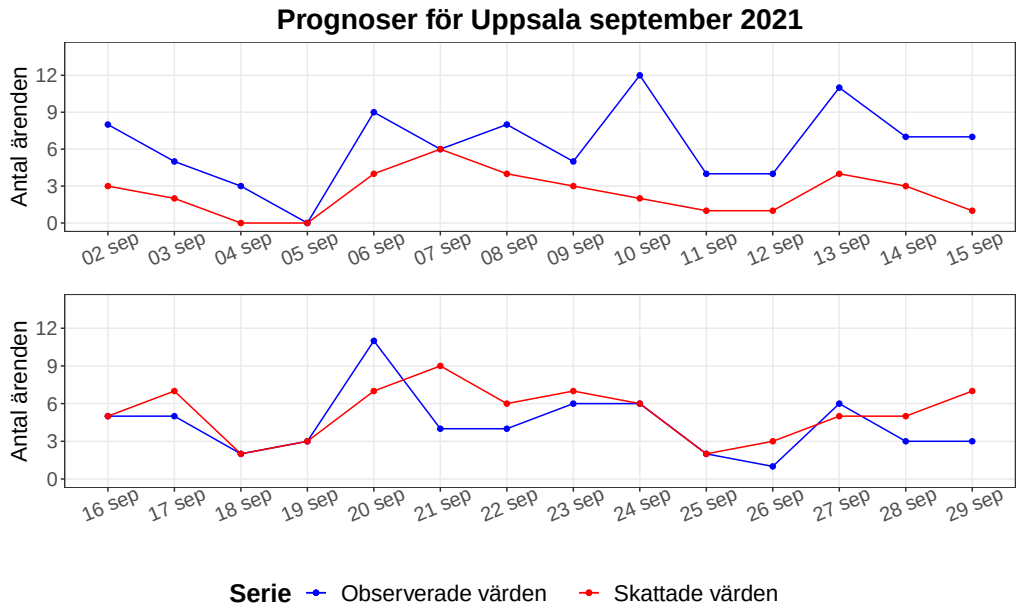
Figur 3: Prognoser för Medborgarplatsen/Liljeholmen under september 2021 med den dynamiska regressionsmodellen med indikatorvariabler för veckodagar och månader



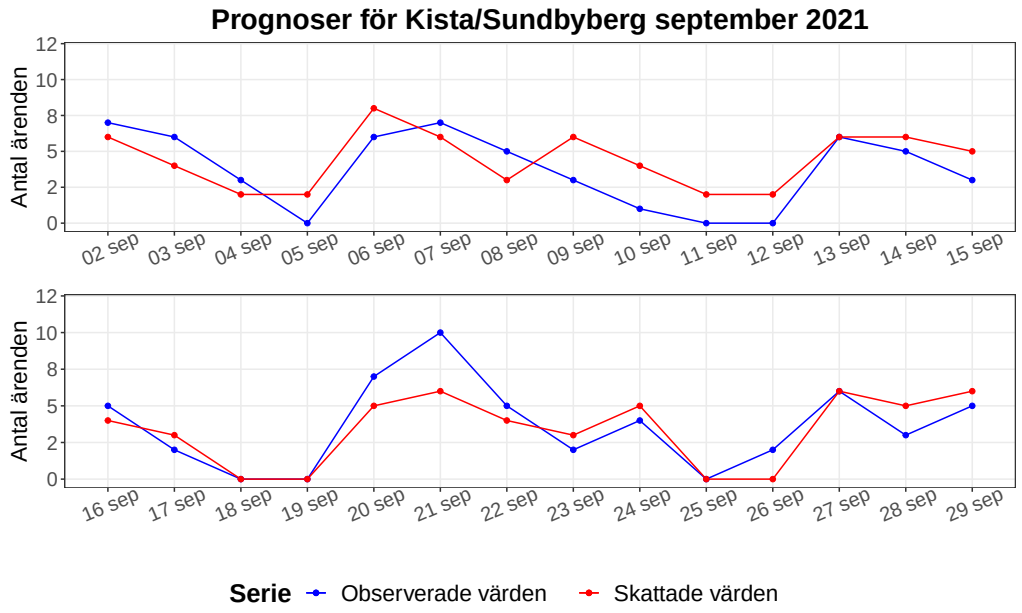
Figur 4: Prognoser för Slakthusområdet/Proppen under september 2021 med XGBoost



Figur 5: Prognoser för Malmö under september 2021 med den dynamiska regressionsmodllen med indikatorvariabler för veckodagar



Figur 6: Prognoser för Uppsala under september 2021 med XGBoost



Figur 7: Prognoser för Kista/Sundbyberg under september 2021 med XGBoost