CHALMERS

GÖTEBORGS UNIVERSITET

# Matematiska modeller för smittspridning av covid-19

En jämförelse av statistiska och mekanistiska modeller

# Mathematical Modeling of Disease Transmission of COVID-19

*Kandidatarbete inom civilingenjörsutbildningen vid Chalmers*


Dan Johansson
Erman Kulaglic
William Nilsen
Zackarias Olofsson
Isabella Simonsson



Institutionen för Matematiska vetenskaper
CHALMERS TEKNISKA HÖGSKOLA
GÖTEBORGS UNIVERSITET
Göteborg, Sverige 2022

# Matematiska modeller för smittspridning av covid-19

En jämförelse av statistiska och mekanistiska modeller

*Kandidatarbete i matematik inom civilingenjörsprogrammet Teknisk matematik vid Chalmers*

Erman Kulaglic  Zackarias Olofsson

*Kandidatarbete i matematik inom civilingenjörsprogrammet Bioteknik vid Chalmers*

Dan Johansson  William Nilsen  Isabella Simonsson

Handledare:  Philip Gerlee

Institutionen för Matematiska vetenskaper
CHALMERS TEKNISKA HÖGSKOLA
GÖTEBORGS UNIVERSITET
Göteborg, Sverige 2022

# Förord

Vi vill rikta ett stort tack till vår handledare Philip Gerlee som har bidragit med sin expertis inom ämnet. Vi vill även tacka våra examinatorer Maria Roginskaya och Ulla Dinger som hjälpt till med det administrativa i kursen och svarat på de frågor vi haft om det. Till sist vill vi tacka Hans Malmström som har bidragit med sin fackspråkliga kompetens och svarat på de fackspråkliga frågor vi har haft.

## Gruppens arbetsgång

Vi började med att diskutera med vår handledare över vad vi eventuellt vill göra i detta projekt. Vi kom ganska snabbt fram till att dela upp arbetet i två, där den ena gruppen tittar på statistiska modeller och den andra gruppen tittar på mekanistiska modeller. Vid eventuella problem med arbetet så vände vi oss först och främst till varandra och därefter till våra möten med vår handledare som vi haft varje vecka på torsdagar.

För att inkludera alla i båda sidor utav arbetet, hölls gruppmöten efter handledningsmötet på torsdagar där vi gick igenom vad vi har gjort samt eventuella problem och hur alla i gruppen mår. Vid behov så har extrainsatta gruppmöten skett på tisdagar.

Vid skrivprocessen har interna deadlines satts för att skicka utkast till handledare för att få återkoppling. Vid eventuella återgärder har Trello använts för att veta vad som har åtgärdats samt vad som fortfarande behövts åtgärdas.

Projektets arbete har införts i en loggbok där varje gruppmedlem har skrivit vad de har gjort och hur lång tid det har tagit. Dagbok har skrivits varje vecka, där ansvaret har cirkulerat inom gruppen. Samma person har även varit mötessekreterare den veckan. Nedan följer en beskrivning av hur vi har arbetat individuellt, samt vad den individen har skrivit i rapporten i ett första utkast. Texterna har sedan bearbetats av samtliga i gruppen.

## Dan Johansson

Under projektets gång så har jag arbetat med SIR-modellerna. Eftersom jag har tidigare kunskaper och ett stort intresse för programmering så tog jag på mig det ansvaret att skriva majoriteten av kodbasen för SIR-modellerna samt plottning av de producerade resultaten. På grund av detta har mitt fokus för skrivandet varit på metod, resultat samt tolkningen av resultatet i diskussionen. I rapporten har jag skrivit på sektionerna:

- Metod 3, Mobilitetsbaserad SIR-modell 3.6, Konfidens- och prediktionsintervall för SIR-modeller 3.6.1, Global parameteroptimering 3.6.2, Resultat 4, Diskussion 5-5.4, Slutsats 7.

## Erman Kulaglic

Under projektets gång så har jag arbetat med de statistiska modellerna. Inledningsvis började jag med att läsa på om teorin för både linjär och ickelinjär regression. Jag tittade även närmare på vilka modeller vi ville studera. Modellerna implementerades sedan i Python och samtidigt försökte jag även skriva på rapporten. I rapporten så har jag bidragit med att skriva på sektionerna:

- Förord, Gruppens arbetsgång, Populärvetenskaplig presentation, Sammanfattning, 2.5. Root mean square error, 2.5.2 Mean absolute percentage error, 4.1, 4.3 Resultat, 5 Diskussion

## William Nilsen

Då programmering var ett nytt område för mig inledde jag projektets arbetsgång med att sätta mig in i grunderna för att använda Python som programspråk. Under projektets gång jobbade jag med de mekanistiska SIR-modellerna. Där bidrog jag mest med att söka och tolka litteratur,

men även med att implementera en del data i modellen, samt input vid utveckling av modellerna. Jag skrev abstract, stycket om samhälleliga och etiska aspekter och bidrog till diskussionen. Jag införde även figurer och en del av härledningen till SIR-modellen.

- Abstract, 6. Samhälleliga och etiska aspekter, 5.3-5.4 Diskussion, 2.4.1 En del av härledning av SIR i teori

## Zackarias Olofsson

Under projektets gång så jobbade jag med de logistiska modellerna. Detta arbetet innefattade att läsa på om modellerna och teorin för regression, implementera lösningarna i Python, generera grafer på resultaten samt tolka resultaten. Utöver de logistiska modellerna jobbade jag även en hel del på en autoregresiv modell som tyvärr inte kom med i rapporten. I rapporten har jag skrivit på sektionerna:

- 2.1 Regression, 2.2 Linjär regression, 2.3 Ickelinjär regression, 2.5 Utvärderingsmetoder för modellerna, 3 inledande stycke till metod, 3.2-3.5 metod för de logistiska modellerna, 3.7 Jämförelse av modellerna, 5.1 Analys MAPE, 5.2 Analys $IS_\alpha$, 5.3 Analys totalt antal inskrivningar.

## Isabella Simonsson

Inledningsvis försökte jag lära mig att programmera i Python, då det var nytt för mig. Efter uppdelningen av modellerna arbetade jag med SIR-modellerna, och började med att läsa in mig på dem. Jag insåg ganska snabbt att mina programmeringskunskaper inte var tillräckliga för att kunna bidra till arbetet med kodning av modellerna. I stället försökte jag hjälpa Dan så gott jag kunde med inputs kring vad som var relevant att ta med, och fokuserade mest på teorin och själva texten i rapporten. Utöver de avsnitt som står nedan, har jag lagt mycket tid på rapporten som helhet. Jag har kontinuerligt läst igenom vad alla har skrivit i de olika avsnitten och rättat grammatiska fel, särskrivningar, och så vidare. Jag har försökt göra vår text mer enhetlig, samt skapa ett flow i texten med formuleringar som gör den lätt att förstå, även om man inte är van vid programmering och matematisk modellering.

- Sammanfattning, 1 Inledning, 2.4 Mekanistiska modeller, 3.1 Inskrivningar och 3.4 Jämförelse av modellerna.

# Populärvetenskaplig presentation

En sjukdom som sprider sig väldigt fort får ofta stora konsekvenser för ett samhälle. När en sådan sjukdom spridit sig till angränsande geografiska områden kallas det epidemi, och när en epidemi har spridit sig till en eller flera världsdelar så kallas det pandemi. Genom människans utveckling har vi gått igenom flera olika pandemier och epidemier som har dödat väldigt många människor. Några som vi har kämpat oss igenom är digerdöden, spanska sjukan och nyligen covid-19.

För att underlätta hanteringen av pandemier kan prognoser göras, vilket matematiken kan hjälpa oss med. För att designa och analysera ett simulerat händelseförlopp kan matematisk modellering användas.

Målet med arbetet är att titta närmare på vilka matematiska modeller som kan representera smittspridningen av covid-19. Matematiska modeller för smittspridning kan ha många former. I denna studie kommer vi endast beakta statistiska och mekanistiska modeller.

De statistiska modellerna bygger på matematiska ekvationer med ett variande antal parameterar vars värden bestäms genom anpassning till ett verkligt händelseförlopp. Totalt beaktas fyra statistiska modeller som alla är baserade på den logistiska ekvationen. Alla fyra modeller beskriver hur många som blir inskrivna på sjukhus, och anpassas efter inskrivningsdata från Sahlgrenska Universitetssjukhus. Den första baseras bara på den logistiska ekvationen. Den andra har en begränsning på hur många som kommer bli inskrivna totalt. Den tredje baseras på den logistiska ekvationen men använder sig av mobilitetsdata för att beskriva hur inskrivningarna sker. Den fjärde och sista har både en begränsning på hur många som blir inskrivna totalt och använder sig av mobilitetsdatan.

De mekanistiska SIR-modellerna bygger på kopplade differentialekvationer som tar hänsyn till mänsklig kontakt, smittans effektiva överföringstakt samt tiden det tar för befolkningen att insjukna och tillfriskna. Från grunden byggdes dessa modeller utifrån ett statiskt perspektiv där både befolkningens rörelsemönster och tidsramen för smittans spridning ignorerades. Detta utvecklades under arbetets gång för att till slut koppla grundparametrarna samman med mobilitetsfunktioner där två modeller byggdes på två olika dataset.

Modellerna som byggs anpassas till en stigande mängd träningsdata. Utifrån dessa modeller jämförs de med det verkliga händelseförloppet för att se hur väl prediktionerna stämmer överens med det verkliga utfallet. Detta görs med hjälp av olika jämförelsemetoder för att ge en så tydlig bild som möjligt om vilka modeller som är bäst i olika perspektiv.

Den största nyttan med att hitta bra modeller är att man med dessa kan planera sjukdomen effektivare. På så vis kan resurserna placeras på rätt ställe så att konsekvenserna i samhället kan minskas. Utöver detta kan vi också eventuellt få en större förståelse hur människan bör ändra sitt beteende i en framtida pandemi. Smittspridningen kan eventuellt bromsas och de stora konsekvenserna i samhället kan minskas.

Vid jämförelse av modellerna visar det sig att olika jämförelsemetoder ger olika resultat. Däremot konstateras det att de logistiska modellerna ger ett bättre resultat så länge infektionen endast ger en våg och om taket är känt på förhand. Vid längre pandemier som ger fler vågor så lämpar sig en mekanistisk modell som SIR-modellen bättre.

## Sammanfattning

De senaste åren har världen präglats av covid-19-pandemin. För att förutspå dess utveckling har matematisk modellering använts. En matematisk modell är en förenklad beskrivning av ett verkligt fenomen, som ger en djupare förståelse om vilka mekanismer som styr ett system. I detta arbete har logistiska modeller och SIR-modeller använts för att beskriva smittspridningen av covid-19 under pandemins första våg, det vill säga mellan 1 mars 2020 och 31 juli 2020. Syftet med arbetet är att konstruera och jämföra olika modeller för inskrivningar på Sahlgrenska Universitetssjukhus till följd av covid-19. Detta görs både kvantitativt och kvalitativt. De utvärderingsmetoder som används vid jämförelsen är: Mean absolute percentage error (MAPE), $IS_\alpha$, samt prediktioner av totalt antal inskrivningar vid ett givet slutdatum.

Modellerna baseras huvudsakligen på två dataset: antalet inskrivningar per dag på Sahlgrenska Universitetssjukhus i Västra götalandsregionen, till följd av covid-19, samt mobilitetsdata från Västtrafik och Google. De sistnämnda används som parametrar för att hjälpa modellerna att prediktera antalet inskrivningar, som sedan jämförs med det första datasetet. De logistiska modellerna använder linjär och ickelinjär regression för att prediktera antalet inskrivningar, medan SIR-modellerna även modellerar smittspridningen. Utifrån den kan antalet inskrivningar predikteras 21 dagar framåt i tiden, till följd av en tidsfördröjning.

Under antagandet att infektionen endast ger en våg inskrivningar visar våra resultat att de logistiska modellerna ger ett bättre resultat än SIR-modellerna. Resultaten är bäst då maxtaket på antalet inskrivningar har begränsats. Vid längre pandemier, som ger fler vågor, visar våra resultat att SIR-modellerna är bättre då de modellerar den dynamiska infektiviteten bättre.



## Abstract

The COVID-19 pandemic has for the last few years impacted all corners of the world. In many attempts to predict the development of the disease, mathematical modelling has been a key asset. Mathematical models are simplified descriptions of real life events and can be used both as tools to give deeper understanding about mechanisms in larger systems, and in order to make predictions about future events.

This report deals with SIR models and logistic regression models to describe the hospital admissions due to COVID-19 during the first wave of the pandemic in Gothenburg. The main purpose of the project was to compare and evaluate these different models of disease transmission.

The models are based mainly on two data sets: the number of admitted COVID-19 patients at Sahlgrenska University Hospital aswell as a mobility data sets from the local municipal trafic, Västtrafik, and Google mobility data. The mobility data sets are used as parameters, helping the models to predict admissions which in turn can be compared to the first data set. The logistical models utilize linear and non-linear regression to predict hospital admissions. The SIR models also predict the disease transmission as a three week forcast due to an implemented time shift.

Assuming that the transmission of COVID-19 only leads to one wave of hospital admissions, the logistical models provides a preferable result, particullarly when the total number of admission is assumed to be fixed. When predicting a long lasting pandemic with several waves of infection we argue that the SIR models are superior, due to the ability to better model the dynamics och disease transmission.

# Innehåll

# 1 Inledning

Få saker genom människans historia har skördat så många liv, och format våra samhällen så mycket som pandemier [1]. Hundratals miljoner människor har dött som konsekvens av olika smittsamma sjukdomar såsom pesten, polio och spanska sjukan [2]. Med modern vetenskap och läkemedel blir vi bättre och bättre på att undvika dem, men de senaste åren har världen präglats av ännu en pandemi: covid-19-pandemin. I slutet av december 2019 meddelades WHO av kinesiska myndigheter om utbrott av lunginflammation av okänt ursprung i Wuhan, Kina [3]. I januari 2020 identifierades ett coronavirus som orsaken, och i mars meddelade WHO att utbrottet skulle klassas som en pandemi. Därefter dröjde det inte länge innan viruset spridit sig till alla världsdelar och stora delar av världen stängdes ned.

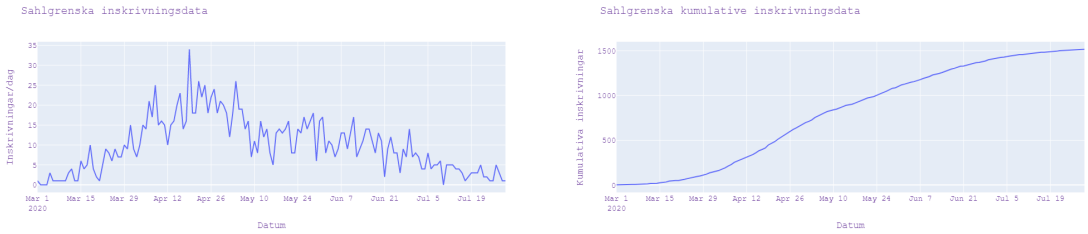
Covid-19 är en så kallad droppsmitta, som huvudsakligen sprids via nära kontakter mellan människor genom små och stora droppar från luftvägarna [4]. Trots att viruset inte är en så kallad luftburen smitta, kan de små dropparna finnas kvar i luften under en tid, och vistelse i trånga utrymmen med dålig ventilation tillsammans med andra människor kan därför leda till infektion. Smittan kan ta sig in i kroppen både via inandning och genom att röra ögon eller slemhinnor i näsa och mun med kontaminerade händer. Genom att hålla avstånd från varandra samt tvätta händerna kan smitta undvikas.

Sedan covid-19-pandemins början har Folkhälsomyndigheten använt matematisk modellering för att försöka förutspå utvecklingen av pandemin. Modellerna har bland annat använts för att studera och prediktera smittspridningen, sjukdomsfall och dödsfall i befolkningen, samt behovet av sjukvårdsresurser [5]. Till detta kan olika modelltyper användas, varav några vanligt förekommande är:

- Statistiska modeller. Dessa skapar prediktioner med antagandet att framtida data kommer följa ett visst mönster i tid, till exempel antalet dagliga fall.
- Fackmodeller, eller tillståndsmodeller. Dessa är en typ av mekanistiska modeller, där populationen delas in i olika fack eller hälsotillstånd. Individer flyttas sedan mellan facken med en viss takt.
- Agentbaserade modeller. Dessa är också en typ av mekanistiska modeller, som liknar tillståndsmodeller. Skillnaden är att varje individ beskrivs för sig, vilket kan ge en mer detaljerad bild av demografin hos populationen.

## 1.1 Modeller för smittspridning

En matematisk modell är en abstrakt och förenklad beskrivning av ett verkligt fenomen eller problem. Dessa modeller gör det möjligt att lösa komplexa problem numeriskt eller ge djupare förståelse om vilka mekanismer som styr ett system [6]. I detta projekt används matematisk modellering för att beskriva smittspridingen av covid-19, samt inskrivningar på Sahlgrenska Universitetssjukhus i Västra götalandsregionen till följd av covid-19. Inskrivningsdatan visas i Figur 1. Modelleringen har valts att begränsas till första vågen, det vill säga från 1 mars 2020 till 31 juli 2020. Detta på grund av att ingen sedan tidigare var vaccinerad eller haft viruset, vilket förenklar modelleringen avsevärt. Modelleringen görs med hjälp av statistiska och mekanistiska modeller, som implementeras i Python. De statistiska modellerna modellerar endast antalet inskrivningar, medan de mekanistiska modellerna modellerar både antalet inskrivningar och smittspridningen.




(a) (b)

Figur 1: (a) Inskrivningsdata från Sahlgrenska Universitetssjukhus per dag. (b) Kumulativ inskrivningsdata.

De statistiska modellerna som används är logistiska modeller, som huvudsakligen baseras på två dataset: antalet inskrivningar per dag på Sahlgrenska Universitetssjukhus i början av pandemin samt mobilitetsdata från Västtrafik. Mobilitetsdatan används som en parameter för att prediktera antalet inskrivningar, som sedan jämförs med det dataset som är hämtat från Sahlgrenska Universitetssjukhus. Utöver det kommer extern kunskap användas om hur högt taket på antalet inskrivna ska sättas till. Därefter används statistiska metoder och kurvanpassning för att förutspå utvecklingen av antalet inskrivningar. För modelleringen används linjär och ickelinjär regression, där koefficienterna anpassas till att passa kurvan för antalet inskrivningar.

De mekanistiska modellerna som används är SIR-modeller, vilka är en typ av fackmodeller. Dessa utgörs av ett system av kopplade differentialekvationer som förutspår antalet mottagliga (**S**useptible), infekterade (**I**nfectious) och återhämtade (**R**ecovered) individer vid en viss tidpunkt. Med hjälp av modellerna kan antalet inskrivningar uppskattas. Detta genom att anta att en andel $p$ av antalet infekterade individer vid samma tidpunkt kommer skrivas in på sjukhus efter $t_a$ veckor. Tidsfördröjningen $t_a$ är tiden från det att en person blir smittad tills denne blir inskriven [7]. Även dessa modeller baseras på data för antalet inskrivningar på Sahlgrenska Universitetssjukhus respektive mobilitetsdata från Västtrafik samt Google. Mobiliteten används som en parameter då den antas påverka infektiviteten av viruset, eftersom smittöverföring sker i nära kontakt med andra människor.

Det genomgående syftet med denna rapport är att konstruera och jämföra olika statistiska och mekanistiska modeller för smittspridningen under den första vågen av covid-19. Jämförelsen kommer göras på två sätt:

- Kvantitativt, genom konfidens- och prediktionsintervall, samt jämförelser av modellernas prediktioner med det verkliga utfallet. Det görs med hjälp av mått som *Root mean square error*, *Mean absolute percentage error* och $\text{IS}_\alpha$.
- Kvalitativt, genom för- och nackdelar med diverse användningsområden, samt en diskussion kring samhälleliga och etiska aspekter.

Rapporten inleds härefter med ett teoriavsnitt för de statistiska respektive mekanisktiska modellerna. Därefter följer ett metodavsnitt som beskriver hur teorin implementerats för att konstruera modellerna, samt hur modellerna ska jämföras. Sedan följer ett resultatavsnitt med figurer som visar prediktionen för de olika modellerna, samt prediktionsfelen. Efter det kommer ett diskussionsavsnitt med diskussion kring resultaten för respektive modelltyp, samt en jämförelse mellan modellerna. Därefter följer en diskussion kring samhälleliga och etiska aspekter. Rapporten avslutas sedan med vår slutsats.

# 2 Teori

I detta avsnitt hanteras teorin som ligger till grund för projektet. Teoriavsnittet inleds med att behandla matematiska begrepp, viktiga satser och matematiska verktyg så som linjär och ickelinjär

regression. Dessa presenteras på djupet för att visa vilka metoder som använts för att optimera och anpassa modellerna. Därefter introduceras grunderna för konstruktion av de mekanistiska samt statistiska modeller som använts under projektets gång. Avslutningsvis presenteras de verktyg som använts vid jämförelsen och konfidens- och prediktionsintervall för modellerna.

## 2.1 Regression

I detta arbete kommer sex olika modeller genereras. Alla dessa modeller kommer beskriva det kumulativa antalet inskrivna, $y$ från 2020-03-01 till dag $t$. Dessa modeller kommer baseras på ett par olika dataset. För att anpassa dessa modeller till datan så kommer en metod som kallas regression användas.

Regression är en metod med målet att anpassa en funktion utefter observerad data, oftast för att få felet mellan funktionen och den observerade datan så litet som möjligt. Vid regression tänker vi oss att de observerade datapunkterna $y_i$ följer en funktion $h$ exakt men att det vid mätningen av $y_i$ uppstår ett additivt fel $\epsilon_i$. Skriver vi ut det har vi att

$$y_i = h(x_i, \boldsymbol{\beta}) + \epsilon_i, \tag{1}$$

där $x_i$ är kända tillståndsvariabler och $\boldsymbol{\beta}$ är okända parametrar. Vi antar även att felen är oberoende och att $\epsilon_i \sim N(0, \sigma^2)$ med en okänd standardavvikelse $\sigma$. Med denna notationen blir problemet att välja en funktion $h$ och hitta dess parametrar $\boldsymbol{\beta}$ så att felet mellan $h$ och de observerade datapunkterna blir så litet som möjligt. Exakt vilken metod som används för att bestämma $\boldsymbol{\beta}$ beror på funktionen $h$ [8].

## 2.2 Linjär Regression

Då $h$ beror linjärt på $x_i$ i (1), det vill säga

$$y_i = h(x_i, \beta) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \epsilon_i,$$

så har vi så kallad linjär regression. I detta fallet kan vi skriva om (1) på matrisform

$$\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2}$$

där

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ 1 & x_{31} & x_{32} & \ldots & x_{3k} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_k \end{pmatrix},$$

$n$ är antalet observationer, $k$ är antalet variabler och $\boldsymbol{\epsilon} \sim N(0, \mathbf{I}\sigma^2)$. En vanlig metod för att skatta $\boldsymbol{\beta}$ vid linjär regression är minsta kvadratmetoden [9].

### 2.2.1 Minsta kvadratmetoden

Syftet med minstakvadratmetoden är att minimera den totala kvadratiska avvikelsen,

$$\sum_{i=1}^{n} (y_i - h(x_i, \boldsymbol{\beta}))^2$$

vilket i det linjära fallet blir

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_k x_{ik})^2.$$

Minimerar vi detta får vi skattningen

$$\boldsymbol{\beta} \approx \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left( \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_k x_{ik})^2 \right) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[ (\mathbf{y} - \boldsymbol{X\beta})^T (\mathbf{y} - \boldsymbol{X\beta}) \right] \text{[9]}.$$

Minsta kvadrat skattningen, $\boldsymbol{\beta}$, hittar vi genom att att lösa ekvationen

$$\frac{\partial}{\partial \boldsymbol{\beta}} \left[ (\mathbf{y} - \boldsymbol{X\beta})^T (\mathbf{y} - \boldsymbol{X\beta}) \right] = \frac{\partial}{\partial \boldsymbol{\beta}} \left[ (\boldsymbol{y}^T \boldsymbol{y} - 2\boldsymbol{y}^T \boldsymbol{X\beta} + \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{X\beta}) \right] = 0.$$

Partialderiverar vi sedan med avsende på varje parameter av $\boldsymbol{\beta}$ får vi normalekvationen

$$\boldsymbol{X}^T \boldsymbol{X\beta} = \boldsymbol{X}^T \boldsymbol{y}.$$

Om matrisen $(\mathbf{X}^T\mathbf{X})$ sedan är inverterbar, är $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ minsta kvadrat-skattningen av $\boldsymbol{\beta}$ till $\mathbf{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ [9, Theorem 3.12]. Skattningen $\hat{\boldsymbol{\beta}}$ kommer vara normalfördelad med väntevärde $\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ och varians $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$, där $\sigma$ är standardavvikelsen för alla $\epsilon_i$. Denna standardavvikelse är obestämbar och måste därför skattas. En vanlig skattning är

$$s^2 = \sum_{i=1}^{n} \frac{r_i^2}{n - (k+1)}, \; r_i = y_i - (\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}). \tag{3}$$

Med denna skattning så får vi istället att $\hat{\boldsymbol{\beta}}$ är $t$-fördelad med $n - (k+1)$ frihetsgrader [9].

### 2.2.2 Konfidens- och prediktionsintervall

Givet en godtycklig datapunkt $\boldsymbol{w} = (1, w_1, w_2, ..., w_k)$ så har vi nu en skattning $\hat{\boldsymbol{y}} = \boldsymbol{w}^T\hat{\boldsymbol{\beta}}$. Även denna skattning kommer att vara normalfördelad men med väntevärde $\mathbb{E}[\hat{\boldsymbol{y}}] = \boldsymbol{w}^T\boldsymbol{\beta}$ och varians $\operatorname{Var}(\hat{\boldsymbol{y}}) = \boldsymbol{w}^T\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{w}$. Skattar vi $\sigma$ med (3) får vi ett konfidensintervall

$$\boldsymbol{y} = \boldsymbol{w}^T\boldsymbol{\beta} \pm a \cdot s\sqrt{\boldsymbol{w}^T(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{w}},$$

där $a$ kommer från $t$-distributionen med $n - (k+1)$ frihetsgrader och bestämmer hur stort konfidensintervallet ska vara. Vi beräknar $a$ genom att först observera att

$$\frac{\boldsymbol{w}^T\hat{\boldsymbol{\beta}} - \boldsymbol{w}^T\boldsymbol{\beta}}{s\sqrt{\boldsymbol{w}^T(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{w}}} \sim t_{n-(k+1)}.$$

Sedan för att få ett $1 - \alpha$ % konfidensintervall beräknar vi $a$ så att

$$\mathbf{P}\left( -a < \frac{\boldsymbol{w}^T\hat{\boldsymbol{\beta}} - \boldsymbol{w}^T\boldsymbol{\beta}}{s\sqrt{\boldsymbol{w}^T(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{w}}} < a \right) = 1 - \alpha.$$

För att få fram ett prediktionsintervall tar vi även hänsyn till mätfelet $\boldsymbol{\epsilon}$ [9]. Detta ger oss en extra term $\sigma^2$ i variansen för vår skattning. Skattar vi återigen $\sigma$ med (3) får vi prediktionsintervallet

$$\boldsymbol{y} = \boldsymbol{w}^T\boldsymbol{\beta} \pm a \cdot s\sqrt{\boldsymbol{w}^T(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{w} + 1}\,.$$

## 2.3 Ickelinjär regression

Då $h$ i (1) inte är linjär så har vi så kallad ickelinjär regression. För ickelinjära $h$ finns det inte alltid något analytiskt sätt att skatta $\boldsymbol{\beta}$ som det finns vid linjär regression. Då måste man använda rent numeriska metoder. För minsta kvadratmetoden beskrivet ovan måste man då använda numeriska minimeringsmetoder. Även konfidens- och prediktionsintervall kan behövas skattas numeriskt. Ibland går det dock att linjärisera problemen för att kunna använda teorin nämnd ovan [8].

### 2.3.1 Linjär approximation

Då det finns ett analytiskt uttryck för $h$, samt då $h$ är deriverbar med avseende på $\boldsymbol{\beta}$, kan vi linjärisera $h$ genom Taylorutveckling med avsende på $\boldsymbol{\beta}$. Vi får då att

$$h(x, \hat{\boldsymbol{\beta}}) \approx h(x, \boldsymbol{\beta}) + \nabla_{\boldsymbol{\beta}} h(x, \boldsymbol{\beta})^T (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \tag{4}$$

där $\hat{\boldsymbol{\beta}}$ är vår normalfördelade skattning och $\boldsymbol{\beta}$ är väntevärdet av denna skattning [8]. Med hjälp av (4) kan vi approximera variansen av vår skattning,

$$\begin{aligned} \mathrm{Var}\left(h(x, \hat{\boldsymbol{\beta}})\right) &= \mathrm{Var}\left(h(x, \boldsymbol{\beta})\right) + \mathrm{Var}\left(\nabla_{\boldsymbol{\beta}} h(x, \boldsymbol{\beta})^T \hat{\boldsymbol{\beta}}\right) + \mathrm{Var}\left(\nabla_{\boldsymbol{\beta}} h(x, \boldsymbol{\beta})^T \boldsymbol{\beta}\right) \\ &= \nabla_{\boldsymbol{\beta}} h(x, \boldsymbol{\beta})^T \mathrm{Var}(\hat{\boldsymbol{\beta}}) \nabla_{\boldsymbol{\beta}} h(x, \boldsymbol{\beta}), \end{aligned} \tag{5}$$

enligt räknereglerna för varians. I (5) antas $\boldsymbol{\beta}$ och $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ vara kända sedan tidigare. Det som återstår för att beräkna konfidens- och prediktionsintervall är att hitta väntevärdet av vår skattning samt att skatta $\sigma$. Väntevärdet är likt vid den linjära regressionen $h(x, \boldsymbol{\beta})$ och $\sigma$ skattas på samma sätt som i (3). Detta ger oss konfidensintervallet

$$\boldsymbol{y} = h(x, \boldsymbol{\beta}) \pm a\sqrt{\nabla_{\boldsymbol{\beta}} h(x, \boldsymbol{\beta})^T \mathrm{Var}(\hat{\boldsymbol{\beta}}) \nabla_{\boldsymbol{\beta}} h(x, \boldsymbol{\beta})}$$

och prediktionsintervallet

$$\boldsymbol{y} = h(x, \boldsymbol{\beta}) \pm a\sqrt{\nabla_{\boldsymbol{\beta}} h(x, \boldsymbol{\beta})^T \mathrm{Var}(\hat{\boldsymbol{\beta}}) \nabla_{\boldsymbol{\beta}} h(x, \boldsymbol{\beta}) + s^2}.$$

## 2.4 Mekanistiska modeller

En mekanistisk modell är en matematisk beskrivning av de element som formar ett system, samt deras interaktioner med varandra respektive miljön [10]. Statistiska metoder används för att skatta parametrar i modellen, vilket gör det möjligt att förutsäga beteendet hos systemet. Vid modellering av epidemier används ofta SIR-modellen, som utvecklades av Kermack and McKendrick år 1927 [11]. Modellen är en fackmodell, vilket innebär att vid spridning av en infektion kan populationen delas in i tre fack:

- Friska individer som kan bli smittade. Kallas för mottagliga individer (susceptible individuals), och betecknas med **S**.
- Smittade individer, som även antas vara smittsamma. Kallas för infekterade individer (infected individuals), och betecknas med **I**.
- Återhämtade individer, som även antas vara immuna, samt borttagna individer. Kallas för återhämtade/borttagna individer (recovered/removed individuals), och betcknas med **R**.

Individer flyttas mellan de olika facken, och därav varierar deras storlek med tiden. En förutsättning för modellen är att den totala populationsstorleken $N$ antas vara konstant, och kan därav definieras som summan av varje fack:

$$N = S(t) + I(t) + R(t).$$

Det finns dock SIR-modeller där födslar tas hänsyn till, i vilka även $N$ är en funktion av tiden $N(t)$. Hur snabbt individer flyttas mellan facken beror på infektiviteten, $\beta$, respektive återhämtningstakten, $\gamma$. Modellen kan beskrivas schematiskt med ett flödesschema enligt Figur 2.


Figur 2: Flödesscchema över hur individer flyttar mellan olika fack i SIR-modellen [12].

### 2.4.1 Härledning av SIR-modellen

SIR-modellen består av ett system av kopplade differentialekvationer som varierar med tiden, med givna initialvillkor $S(0)$, $I(0)$ och $R(0)$ [11]. Vi antar att en infekterad individ har $\kappa$ kontakter per tidsenhet, där $\kappa$ är oberoende av populationsstorleken. Antalet kontakter med mottagliga individer blir då $\kappa S/N$. Med överförbarheten $\tau$, det vill säga andelen kontakter som resulterar i smittöverföring, får vi att varje infekterad individ smittar $\kappa\tau S/N$ mottagliga individer per tidsenhet. Vidare får vi att infektiviteten $\beta = \kappa\tau/N = b/N$ per tidsenhet. För antalet mottagliga individer får vi följande differentialekvation:

$$\frac{dS}{dt} = -\beta SI,$$

som minskar med antalet individer som blir infekterade. Dessa ökar dock samtidigt i belopp, eftersom det är en funktion av $I$, som blir större desto fler som blir infekterade. Individer flyttas från fack $S$ till $I$, varefter de individer som återhämtats eller dött flyttas vidare till fack $R$. Detta sker med en återhämtningstakt $\gamma$ per tidsenhet, som följer av infektionens varaktighet $D = 1/\gamma$ tidsenheter. Av detta fås följande differentialekvationer:

$$\frac{dI}{dt} = \beta SI - \gamma I$$
$$\frac{dR}{dt} = \gamma I.$$

Således bildar följande system av differentialfunktioner SIR-modellen [13]:

$$\frac{dS}{dt} = -\beta SI$$
$$\frac{dI}{dt} = \beta SI - \gamma I$$
$$\frac{dR}{dt} = \gamma I$$

## 2.5 Utvärderingsmetoder för modellerna

För att utvärdera hur väl prediktionerna stämmer överens med inskrivningsdatan använder vi tre olika metoder: Root mean square error, Mean absolute percentage error, samt $\text{IS}_\alpha$ scoring rule. Root mean square error och mean absolute percentage error används för att se hur väl modellerna stämmer överens med inskrivningsdatan medan $\text{IS}_\alpha$ är ett mått som beaktar både träffsäkerheten och osäkerhet i prediktionerna.

### 2.5.1 Root mean square error

Root mean square error, eller RMSE, är en metod för att få fram en siffra som beskriver hur bra en modell överensstämmer med det verkliga utfallet. RMSE beräknas som

$$\textbf{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2} \tag{6}$$

där $n$ är antalet datapunkter, $y_i$ är modellens predikterade värde och $x_i$ är det verkliga värdet [14].

### 2.5.2 Mean absolute percentage error

Mean absolute percentage error, eller MAPE, är ett ytterliggare mått på hur modellen överensstämmer med det verkliga utfallet och fås i procent. MAPE beräknas som

$$\textbf{MAPE} = \frac{1}{n}\sum_{i=1}^{n} \mid \frac{x_i - y_i}{x_i} \mid \tag{7}$$

där n är antalet datapunkter, $y_i$ är modellens predikterade värde och $x_i$ är det verkliga värdet [15].

### 2.5.3 $\text{IS}_\alpha$ scoring rule

För att ta hänsyn till osäkerheten i prediktionerna kan en scoring rule användas. Vi har valt att använda $\text{IS}_\alpha$, som beräknas enligt

$$\text{IS}_\alpha(F, y) = (u - l) + \frac{2}{\alpha} \cdot (l - y) \cdot \mathbf{1}(y < l) + \frac{2}{\alpha} \cdot (y - u) \cdot \mathbf{1}(y > u) \tag{8}$$

där $F$ är modellen, $y$ är den observerade datan vid tidpunkt $t$, $u$ och $l$ är den övre respektive undre gränsen på det $(1 - \alpha) \cdot 100\%$ prediktionsintervallet av $F$ vid tidpunkt $t$ och $\mathbf{1}(y < l) = 1$ om $y > l$ och 0 annars. Den första termen i (8) bestämmer bredden på prediktionsintervallet. De andra två termerna är strafftermer som adderas om den observerade datan ligger utanför prediktionsintervallen [16, sida 3-4].

# 3 Metod

Det primära verktyget som använts i projektet är programmeringsspråket Python, vilket genomför alla beräkningar och hjälper oss visualisera resultaten. Utöver Pythons baskod har vi importerat och använt olika bibliotek som innehåller viktiga funktioner, för att underlätta arbetet. Några av de viktigaste biblioteken som använts är SciPy, Plotly, Numpy och Pandas. SciPy innehåller metoder och funktioner som används brett inom naturvetenskap, men i vårt fall för att lösa differentialekvationssystem samt optimering av parametrar. Plotly används frekvent genom projektet, eftersom det erbjuder många olika sätt att visualisera och presentera data och resultat. Nympy och Pandas används för importering och hantering av de olika dataset som framkommer i projektet. All kod, inklusive datan, finns att hämta på projektets GitHub. Eftersom datan som används i detta projekt är statisk fungerar även projektets Github som en databas för de dataset som används.

Till de modeller som togs fram användes tre olika dataset. Det första var antalet inskrivningar på Sahlgrenska Universitetssjukhus per dag mellan datumen 2020-03-01 och 2020-07-31. I detta dataset saknades data för vissa datum, där det antogs vara 0 inskrivningar. Det andra var mobilitetsdata från Västtrafiks kundräkningssystem, beskrivet i veckovis procentuell skillnad i det totala resandet jämfört med vecka 9 2020. Här avser den totala skillnaden i resandet, skillnaden i resandet med buss, tåg och spårvagn. Det tredje var mobilitetsdata från Google, som innefattar en procentändring av populationens rörelsemönster i kollektivtrafiken från ett basvärde som hämtades från första datumet, (2020-03-01).

## 3.1 Inskrivningar

I denna rapport definieras en inskrivning som en person som skrivs in på sjukhus till följd av covid-19. Detta inkluderar även de som inte vårdats på IVA-avdelningen. Antalet inskrivningar kan beskrivas med hjälp av SIR-modellen genom att anta att en andel $p = 0.013$ av antalet infekterade personer vid en viss tidpunkt kommer bli inskrivna efter $t_a = 21$ dagar. Detta beror på tidsfördröjningen från att en person blir smittad tills denne blir inskriven [7]. Dessa värden är hämtade från en studie som gjort liknande prediktioner som i detta arbete. De är dock beräknade för alla regioner i Sverige, och inte specifikt Sahlgrenska Universitetssjukhus upptagningsområde i Västra götalandsregionen, från vilket inskrivningsdatan som används är hämtad.

## 3.2 Logistiska modellen

Den första statistiska modellen som anpassades till datan var den så kallade logistiska ekvationen

$$y = h(t, \boldsymbol{\beta}) = h(t, L, k, t_0) = \frac{L}{1 + e^{-k(t - t_0)}}. \tag{9}$$

Denna modell beskriver det kumulativa antalet inskrivningar fram till dag $t$. För att anpassa denna modell till inskrivningsdatan summerades därför inskrivningsdatan så att den beskrev det kumulativa antalet från 2020-03-01 till dag $t$ istället för antalet inskrivningar vid dag $t$. För att beräkna $\boldsymbol{\beta} = (L, k, t_0)$ samt covariansmatrisen för $\hat{\boldsymbol{\beta}}$, $\text{Var}(\hat{\boldsymbol{\beta}})$, användes **scipy.optimize.curve_fit()** som

använder sig av en ickelinjär minstakvadratmetod. För att **curve_fit()** skulle kunna hitta en lösning sattes en startgissning av $\boldsymbol{\beta}$ till $(100, 1, 1)$. Valet av dessa siffror var ganska godtyckligt, alla sattes till positiva tal då modellen inte riktigt fungerar för negativa värden.

För att beräkna konfidens- och prediktionsintervallen användes metoden beskriven i 2.3.1 där gradienten av $h$ med avseende på parametrarna beräknades till

$$\nabla_{\boldsymbol{\beta}} h = \begin{pmatrix} \frac{1}{1+e^{-k(t-t_0)}} \\ \frac{L(t-t_0)e^{-k(t-t_0)}}{1+e^{-k(t-t_0)}} \\ \frac{-kLe^{-k(t-t_0)}}{1+e^{-k(t-t_0)}} \end{pmatrix}.$$

## 3.3 Logistiska modellen med begränsat tak

Den andra statistiska modellen var också den logistiska ekvationen (9) men denna gången begränsades maxtaket på antalet inskrivna $L$ till att ligga mellan 1450 och 1550. Valet av 1450 och 1550 baserades på inskrivningsdatan och ska representera möjligheten att skatta taket i förväg baserat på data från andra länder som ligger tidigare i smittspridningen. Detta gjordes genom att sätta gränser i **curve_fit()**. I **curve_fit()** måste antingen alla eller inga parametrar ha gränser för att det ska fungera. Gränserna för $k$ och $t_0$ sattes till 0 och $\infty$. Startgissningen ändrades till $(1500, 1, 1)$ för att vara innanför gränserna. Resterande löstes på samma sätt som för den logistiska ekvationen utan begränsat tak.

## 3.4 Logistiska modellen med mobilitetsdata

Den tredje statistiska modellen var en anpassad logistisk ekvation som även tog hänsyn till mobilitetsdatan, $m(t)$, från Västtrafik

$$y = y(t, \boldsymbol{\beta}) = h(t, L, k, t_0, d) = \frac{L}{1 + e^{-k(1+d \cdot m(t-21))(t-t_0)}}. \tag{10}$$

Skillnaden mellan (9) och (10) är en faktor $1+d \cdot m(t-21)$ i exponenten. Denna faktor ska representera att hastigheten på antalet som skrivs in vid dag $t$ påverkas av mobiliteten av populationen 21 dagar innan dag $t$. Valet av 21 dagar baserades på en studie som visade att mobilitetrestriktioner påverkade antalet inläggningar 9 till 25 dagar framåt [7]. I termen är $d$ en parameter mellan 0 och 1 som ska skattas. Anledningen till just detta intervall är att $-k(1 + d \cdot m(t - 21))$ inte ska kunna bli positivt.

Datan från Västtrafik var veckovis men inskrivningsdatan var dagsvis. Därför användes linjär interpolation för att göra mobilitetsdatan från Västtrafik dagsvis. Mobilitetsdatan från Västtrafik sträckte sig inte tillräckligt långt bak i tiden så det antogs att resandet vecka 6, 7 och 8 var samma som vecka 9.

Denna modell löstes på samma sätt som den logistiska modellen med **curve_fit()** och metoden från 2.3.1 men med andra parametrar, ett till dataset och en gradient, $\nabla_{\boldsymbol{\beta}} h$, beräknad från (10).

## 3.5 Logistiska modellen med mobilitetsdata och begränsat tak

Den fjärde statistiska modellen var (10) men där $L$ var begränsad till att ligga mellan 1450 och 1550. Detta gjordes på samma sätt som för den logistiska ekvationen med begränsat tak.

## 3.6 Mobilitetsbaserad SIR-modell

Den första mobilitetsbaserade SIR-modellen baserades på Google mobilitetsdata hämtad från *Our world in data*. Detta dataset kommer från Google's *Community Mobility Report* som publicerades på grund av covid-19-pandemin. På liknande sätt hämtades även mobilitetsdatan för kollektivtrafik

från Västtrafik. För att kunna använda mobilitetsdatan som en tidsberoende parameter interpolerades datapunkterna till tidsberoende funktioner. Dessa mobilitetsfunktioner användes sedan i SIR-modellen som tidsberoende parametrar, dock med koefficienter för att bilda en viktad linjärkombination:

$$\kappa = c_0 + c_1 * \textbf{Mobility data function (t)}$$

där $c_0$ representerar en basal infektivitetshastighet, $c_1$ viktar mobilitetsfunktionen och $\kappa$ är kontakter per tidsenhet. Dessa koefficienter skattades sedan med hjälp av SciPy's **curve_fit()**-funktion vilket returnerar de optimala värdena samt deras kovarians. Här skattades $c_0$ och $c_1$ så att felet mellan antalet infekterade individer $I$ från modellen och inskrivningsdata minimerades. Återhämtningstakten $\gamma$ sattes till 0.2 för dagliga modellen och till 1.4 veckoliga modellen [7].

Tidsstegen i modellen har fram tills nu varit dagsvis, men en veckobaserad modell implementerades också för att ge en mjukare trend jämfört med den mer oregelbundna dagsmodellen. Veckomodellen konstruerades genom att summera den dagliga inskrivningsdatan till veckolig och sedan parameteroptimera till detta dataset istället.

### 3.6.1 Konfidens- och prediktionsintervall för SIR-modeller

Eftersom SIR-modellen saknar en analytiskt lösning kommer konfidens- och prediktionsintervall beräknas med hjälp av en parametrisk bootstrap [7]. För beräkning av konfidensintervall definierades en matris av modellvärden, genom att generera nya parametervärden för modellen utifrån en normalfördelning i varje tidsintervall. För normalfördelningen användes de optimala värdera för parametrarna som medel samt deras kovarians som erhölls från **curve_fit()**. Konfidensintervallen beräknades med stödfunktionen **quantile()** från Numpy. Denna funktion tar emot en matris av modellvärden samt en konfidensintervallgräns och returnerar värdena för konfidensintervallet.

Prediktionsintervallen beräknades på liknande sätt som konfindensintervallen, med skillnaden att både parametrarnas varians och modellens mätfel måste inkorporeras. En fördelning av parametrar plockades ut som tidigare ur en normalfördelning. Medelvärdet antogs vara 0 och standardavvikelsen av felet mellan modellen och datan beräknades. För att skatta felet användes en t-fördelningsfunktion och värden genererades för varje tidssteg. Parameterfördelningen samt mätfels-fördelningen användes för att generera datapunkter. Med hjälp av **quantile()**-funktionen beräknades den undre och övre gränsen av prediktionsintervallet i varje tidssteg.

### 3.6.2 Global parameteroptimering

Utöver parameteroptimering med **curve_fit()** gjordes också en global sökning av parametervärden. Här genererades ett intervall av parametervärden, varefter varje kombination av dessa olika värden ställdes upp i en matris för att skapa ett parameterrum. Parameterrummet utvärderades sedan genom att skapa en konturplot av log(RMSE) mot parameterrummet, vilket gjordes enkelt med Plotly. Detta möjliggör att visuellt hitta lokala minimum samt bra startvärden för **curve_fit()** att optimera från. Detta gjordes för data från både Västtrafik och Google.

## 3.7 Jämförelse av modellerna

För att jämföra modellerna användes tre olika metoder. Alla gick ut på att anpassa modellerna till ett ökande antal datapunkter $n$, där en datapunkt motsvarar en dag. Sedan beräknades hur bra modellens prediktioner stämde överens med datan som följde efter träningsdatan. De tre metoderna är:

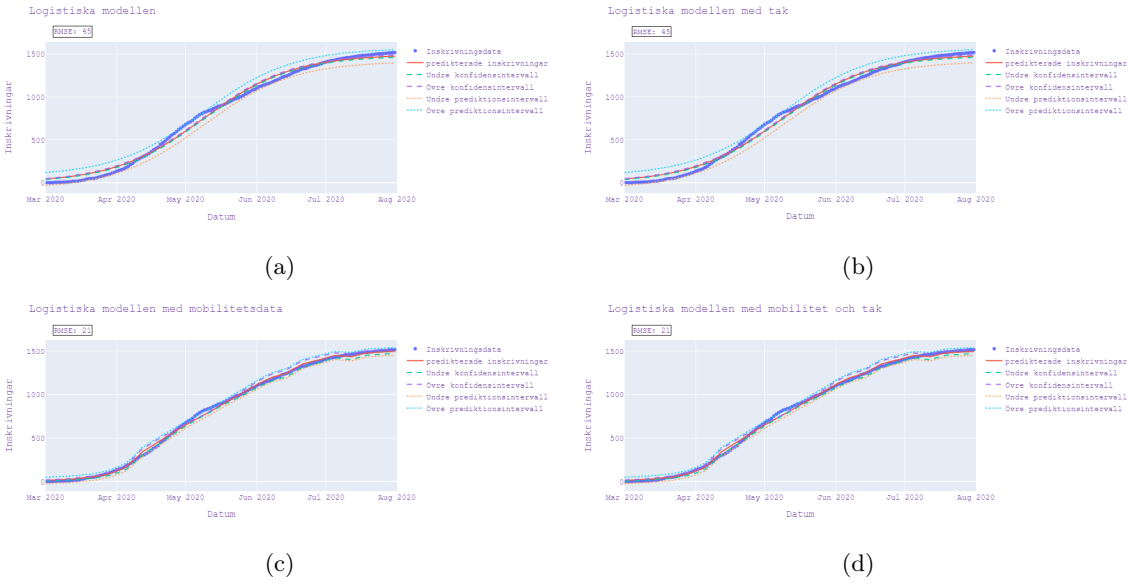
- MAPE, som beräknades för de 21 nästkommande dagarna efter $n$. Se Ekvation (7).
- $\text{IS}_\alpha$ scoring rule, som beräknade osäkerheten i parametrarna. Här beräknades medeltalet över de nästkommande 21 dagarna efter $n$. Se Ekvation (8).
- Jämförelse av det totala antalet inskrivna vid den sista tidpunkten, 2020-07-31.

# 4 Resultat

Följande avsnitt innehåller resultat för statistiska respektive SIR-modellerna, samt en jämförande del där olika utvärderingsmetoder används. Dessa ger en bättre förståelse av prediktionsförmåga, stabilitet och applicerbarhet av modellerna.

## 4.1 Statistiska modeller

I Figur 3 ser vi en logistisk modell anpassad till all inskrivningsdata, det vill säga 153 datapunkter. I Figur (3a) ser vi modellen där endast inskrivningsdatan använts för att anpassa den logistiska ekvationen. I Figur (3b) inkluderar modellen även ett tak på totalt antal inskrivningar. Villkoret säger att det totala antalet inskrivningar vid slutdatum ska vara mellan 1450 och 1550 inskrivningar. I Figur (3c) anpassas den logistiska ekvationen till inskrivningsdata samt mobilitetsdata från Västtrafik. I Figur (3d) används inskrivnings- och mobilitetsdata, och det sätts ett tak på totalt antal inskrivningar vid slutdatum mellan 1450 och 1550 inskrivningar [1].




(a) (b)

(c) (d)

Figur 3: (a) visar den statistiska modellen, (b) visar den statistiska modellen med tak, (c) visar den statistiska modellen med mobilitetsdata och (d) visar den statistiska modellen med mobilitetsdata samt tak plottad mot inskrivningsdata med konfidens- och prediktionsintervall.

## 4.2 SIR-modeller

I detta avsnitt kommer de samlade resultaten för SIR-modellerna, inklusive parametersökningen som gjordes för att hitta goda startvärden för parameteroptimisering av **curve_fit()**.

### 4.2.1 Global paramtersökning

Resulatet från parametersökningen kan ses i Figur 9a-10b i Appendix 9.1. Vi observerade att RMSE skiljer sig mycket berode på vart man placerar sig i parameterrummet, och därför minskades sökområdet för parameter $c_1$. Utifrån denna plot valdes en initial gissning för parameteranpassningen för Västtrafik mobilitetsdata till $c_1 = 1, c_2 = 0.07$.

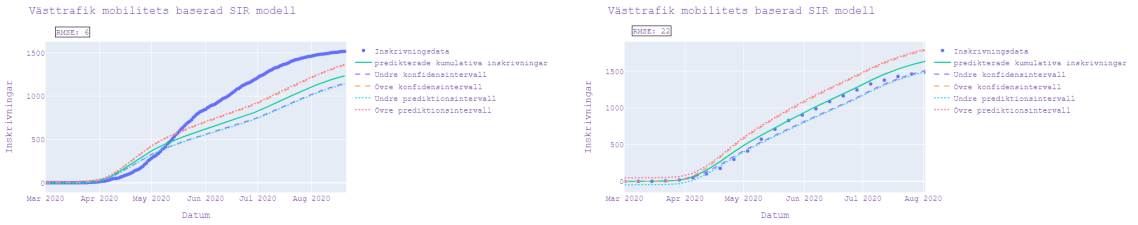
Den globala parametersökningen för Googles mobilitetsdata gav ett mindre tydligt resultat. Vi observerade att ett fel log(RMSE) under 4.4 hittas i en stor del av parameterrummet. En nedskalning av parameterrummet gav bättre resultat. Utifrån resultatet valdes en initial gissning för Googles

[1] Grafer med alla resultat för alla olika $n$ finner du här.

mobilitetsdata till $c_0 = 3, c_1 = 0.6$.

Med de implementerade initialgissningarna fortsatte Västtrafiks mobilitetsdata att ge goda resultat vid modellanpassning, medan Googles mobilitetsdata var för oregelbunden för parameteranpassning med **curve_fit()**. På grund av detta togs beslutet att förkasta modellen baserad på Google mobilitetsdata och endast fortsätta med den Västtrafik-baserade modellen. Preliminärt resultat för Västtrafik baserad modell kan ses i Figur 11 i Appendix 9.1.

De slutgiltiga resultaten för de dagsbaserade samt veckobaserade modellerna, som blivit anpassade till hela datasetet av inskrivningar, kan ses i Figur 4 vilket visar det kumulativa antalet inskrivningar.




(a) (b)

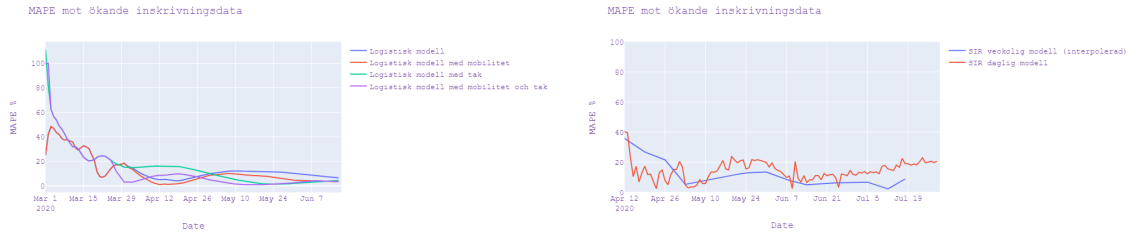
Figur 4: (a) visar den dagsbaserade kumulativa SIR-modellen. (b) visar veckobaserade kumulativa modellen plottad mot inskrivningsdata med konfidens- och prediktionsintervall.

## 4.3 Jämförelse av modeller

Nedan följer jämförelser av modellerna med tre olika metoder: MAPE, $\text{IS}_\alpha$ score samt totalt antal inskrivna vid ett slutdatum. Utifrån dessa kommer modellerna utvärderas och diskuteras.

### 4.3.1 MAPE

I Figur 5 ser vi resultatet av MAPE för både de logistiska modellerna och SIR-modellerna mot ökande mängd anpassningsdata.

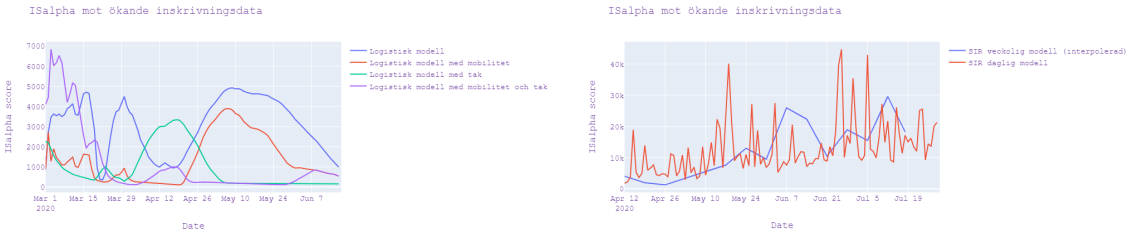



(a) (b)

Figur 5: MAPE beräknat som en funktion av ökande anpassningsdata för (a) de logistiska-modellerna, och (b) SIR-modellerna.

### 4.3.2 Interval score

I Figur 6 ser vi resultatet av $\text{IS}_\alpha$ score för både de logistiska modellerna och SIR-modellerna mot ökande mängd anpassningsdata.

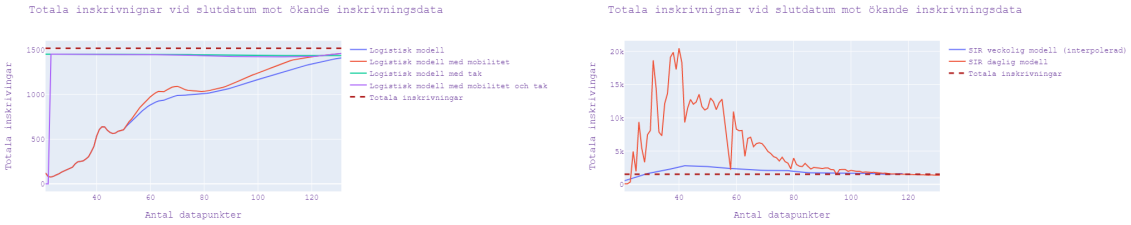



(a) (b)

Figur 6: $IS_\alpha$ score beräknat som en funktion av ökande anpassningsdata för (a) de logistiska-modellerna, och (b) SIR-modellerna.

### 4.3.3 Totala inskrivningar

I Figur 7 ser vi resultatet av totala antalet inskrivningar vid 31 juli 2020 för både de logistiska modellerna och SIR-modellerna mot ökande mängd anpassningsdata.




(a) (b)

Figur 7: Totalt antal inskrivningar beräknat som en funktion av ökande anpassningsdata för (a) de logistiska-modellerna, och (b) SIR-modellerna.

# 5 Diskussion

I detta avsnitt diskuteras resultatet i samma ordning som de är presenterade i föregående avsnitt. En diskussion hålls om varje jämförelsemetod för de logistiska modellerna och SIR-modellerna, där de gås igenom individuellt och sedan jämförs med varandra. Därefter diskuteras behovet av ytterligare kunskap och eventuella framtida problemställningar inom ämnet.

## 5.1 Analys MAPE

I Figur 5a ser vi att alla fyra logistiska modeller blir bättre på att prediktera de 21 nästkommande dagarna ju mer data de tränas på. Detta kan dels bero på att modellerna i allmänhet stämmer mer överens med datan då de tränas på mer data. Det kan dock också bero på att förändringen i antalet inskrivningar avtar då $t$ är stort, detta kan göra att prediktionerna blir säkrare.

I Figur 5a ser vi även att de modeller med begränsat tak är sämre på att prediktera de 21 nästkommande än de som inte har begränsat tak när de har data fram till april att anpassa sig efter. De är dock betydligt bättre på att prediktera när de har data fram till maj och juni att anpassa sig efter. Att de är bättre senare beror på att det begränsade taket gör störst nytta i slutet då modellerna enligt begränsningarna hamnar nära det faktiska värdet. Att de är sämre med mindre data, fram till april, beror på att inskrivningsdatan inte är helt utformad som en logistisk kurva, se Figur 3. Detta gör att de modeller med begränsat tak hela tiden begränsas till att ha fel form samtidigt som de utan begränsat tak kan anpassa sig mer till hur inskrivningsdatan ser ut fram till april.

För SIR-modellerna observerar vi i Figur 5b att det finns en generell trend av sjunkande MAPE

med ökande mängd anpassningsdata. Dock ökar MAPE mot slutet av anpassningsdatan för den dagliga modellen. Vi observerar generellt en relativt stor varians för MAPE, det vill säga MAPEs värde är känsligt för hur många och vilka datapunkter som utgör anpassningsdatan. Över lag observeras ett mindre fel i veckoliga modellen.

Om vi jämför MAPE-resultatet för logistiska och SIR-modellerna ser vi att SIR-modellen har ett lägre initialt MAPE som sedan stannar mellan 0-20%, medan de logistiska modellerna har ett högt initialt MAPE som till slut, med tillräckligt mycket anpassningsdata, når ett fel som är jämförbart med SIR-modellerna.

## 5.2 Analys $\text{IS}_\alpha$

I Figur 6a ser vi att de två modellerna med begränsat tak har liknande kurvor och att de modellerna utan begränsat tak också har liknande kurvor. Detta beror på att de modeller med begränsat tak har stora prediktionsintervall samtidigt som de utan begränsat tak har smala prediktionsintervall. Detta gör att det framförallt är bredden på prediktionsintervallen som bestämmer $\text{IS}_\alpha$ för de med begränsat tak. För de modeller utan begränsat tak är det istället strafftermerna som dominerar $\text{IS}_\alpha$. Särskilt stor skillnad mellan modellerna med och utan begränsat tak är det i slutet. Detta eftersom de med begränsat tak har smala prediktionsintervall som ändå innesluter inskrivningsdatan och därmed får små $\text{IS}_\alpha$ värden. De med obegränsat tak får något större prediktionsintervall men trots detta innesluter de inte inskrivningsdatan. Därav får de väldigt stora $\text{IS}_\alpha$-värden.

Om vi i Figur 6a jämför modellerna utan begränsat tak ser vi att modellen med mobilitetsdata hela tiden har lägre värden än den utan. Detta beror på att den har en parameter extra som kan göra att den bättre anpassas till inskrivningsdatan samtidigt som den generellt sett har större prediktionsintervall och därmed oftare innesluter inskrivningsdatan. Om vi istället jämför de med begränsat tak så ser vi att modellen utan mobilitetsdata har lägre $\text{IS}_\alpha$ värden förutom runt april. Detta beror på att båda modellerna i stort följer inskrivningsdatan och att prediktionsintervallen för båda innesluter inskrivningsdatan. Prediktionsintervallen för modellen utan mobilitetsdata är dock mindre än de för modellen med mobilitetsdata, därav lägre $\text{IS}_\alpha$-värden. Att $\text{IS}_\alpha$-kurvan får en topp runt april beror på att prediktionsintervallen minskar kring april och inskrivningsdatan hamnar då utanför prediktionsintervallen. Detta beror på att inflexionspunkten på den logistiska kurvan hamnar runt april-maj och då blir $(t - t_0)$-termen i $\nabla_{\boldsymbol{\beta}} h$ väldigt liten vilket gör att prediktionsintervallen minskar.

$\text{IS}_\alpha$ för SIR-modellerna ser dock ganska annorlunda ut, vi observerar direkt en stor skillnad i storleksordning av $\text{IS}_\alpha$ mellan logistiska modellerna och SIR-modellerna. Logistiska modellerna når ett maxvärde på cirka 7000 medan SIR-modellerna når maxvärden över 30 000. Vi observerar också att värdet ökar med mängden inskrivningsdata. De höga $\text{IS}_\alpha$-värdena för den dagsbaserade modellen kan vara ett resultat av att modellen var svår att anpassa till datan, se Figur 4a. Det vill säga breda prediktionsintervall som växer med mängden inskrivningsdatan. Den veckoliga modellen visar ett mycket stabilare $\text{IS}_\alpha$ dock också med en uppåttrend, vilket förklaras av att prediktionsintervallets bredd växer med mängden inskrivingsdata även i detta fall.

Jämför vi logistiska och SIR modellerna i Figur 6 kan vi konstatera att logistiska modellerna speciellt dem med tak genererar bäst resultat med andra ord högst säkerhet i prediktionen.

## 5.3 Analys totalt antal inskrivningar

I Figur 7 ser vi resultatet av beräknade kumulativa inläggningar på IVA mellan första mars och sista juli. Vi observerar i Figur 7a att de logistiska modellerna med tak direkt predikterar att slutgiltiga antalet inskrivningar ska vara cirka 1500, vilket är förväntat av dessa modeller på grund av att initialgissningen var 1500 inskrivningar. För modellerna utan tak observerar vi att prediktionerna för slutgiltigt antal inskrivningar blir grovt underskattade med liten mängd inskrivningsdata. Detta resultat är inte helt oväntat då modellen gör en lång framtidsprediktion med liten datamängd. Dessa resultat medför att en modell med tak är att föredra för långsiktiga prediktioner, förutsatt

att det finns goda kunskaper eller argument att grunda värdet av taket på. För kortsiktiga prediktioner erbjuder modeller med eller utan tak liknande resultat som vi såg i MAPE-resultatet.

I Figur 7b ser vi resultatet av beräknade kumulativa inskrivningar för båda SIR-modellerna. Vi observerar att kurvan stiger väldigt hastigt för den dagsbaserade modellen, medan mer rimliga värden presenteras om beräkningen görs veckovis. Anledningen till detta kan vara att modellen svarar för fort på inskrivningsdatan vid den dagliga beräkningen. Efter ett tag når modellen ett tak och därmed sjunker antalet inskrivningar fort igen. Vad gäller den veckovisa modellen hinner modellen anpassa sig till taket innan antalet inskrivningar ökar för mycket, därav ser kurvan mer stabil ut. Över lag erbjuder dagliga SIR-modellen ett dåligt resultat för långsiktiga framtida prediktioner med låg mängd inskrivningsdata medan veckoliga modellen erbjuder ett jämförtbart bra resultat även med låg mängd inskrivningsdata.

Totala antalet inskrivningar vid ett slutdatum visade sig vara problematiskt att prediktera för alla modeller då datamängden var liten, förutom logistiska modellerna med tak som diskuterats tidigare. De logistika modellerna har en tendens att underskatta de totala inskrivningarna vilket är förväntat med tanke på logistiska funktionens mittvärdesparameter, $t_0$, som uppskattas av **curve_fit()**. Detta leder till att modellen når sitt maximala värde för tidigt, vilket medför en underskattning av slutvärdet. Detta till skillnad från SIR-modellerna som överskattar slutgiltiga antalet inläggningar speciellt vid låga datamängder. Detta är också förväntat beteende då SIR-modellerna tidigt i pandemin beskriver exponentiell tillväxt, och om infektiviteten överskattas som vid låg datamängd kan responsen av modellen snabbt bli väldigt stor. Utifrån detta resultat kan vi konstatera att veckoliga SIR modellen producerar bäst långsiktig prediktion bortsätt från logistiska modellerna med tak.

## 5.4 Kvalitativ jämförelse av modeller

Både logistiska modellerna och SIR-modellerna producerade användbara resultat i en efterstudie som denna, men under en aktiv pandemi har båda metoderna olika styrkor och svagheter. För att kunna använda den logistiska metoden krävs att infektionen endast sker i *en* våg, eller att modellen endast beskriver en våg åt gången, vilken är en stor begränsning. De logistiska modeller där ett tak används, krävs också kunskap om vad det totala antalet infektioner kommer bli, vilket kan vara väldigt svårt att uppskatta tidigt i en pandemi. Fördelen med logistiska modellerna att de inte kräver mobilitetsdata, till skillnad från SIR-modellerna. Dessutom kräver SIR-modellerna kunskap om återhämtningstakten $\gamma$ samt längden på tidsfördröjningen mellan infektion och inskrivning på sjukhus. Båda dessa parameterar påverkar modellen starkt och är därför väldigt viktiga att få rätt, vilket också kan vara svårt i början på en pandemi. En klar styrka med SIR-modellen är dock att den kan beskriva flera vågor.

## 5.5 Ytterligare förkunskaper och framtida problemställningar

Som tidigare nämnts finns det en mängd olika variationer och valmöjligheter av vilka modeller som kan användas för modellering av smittspridning. Beroende på hur mycket förkunskap man har från smittspridningar med liknande karaktär kan man enklare anpassa det förebyggande arbetet till framtida pandemier. Problemet med covid-19 och dess mutationer är att mekanismerna för hur smittan fortlöper ser olika ut för varje variant. För att kunna prediktera och konstruera väl anpassade modeller för dessa krävs stora mängder data om de olika biologiska faktorer som spelar in på hur befolkningen smittar varandra genom kontakt samt hur kraftig reaktion smittan har.

En ambition med projektet var att även applicera bayesianska modeller på vårt aktuella område. Med en sådan modell kan man beräkna sannolikheten att smittspridningen kommer fortlöpa på ett specifikt sätt beroende på tidigare kunskap om smittspridningen och modellera därefter med ny insamlad empirisk data. Detta hade varit ett sätt att utveckla modelleringen vidare, men vi valde att begränsa oss till de urval av statistiska och mekanistiska modeller som diskuterats tidigare.

Framgången av artificiell intelligens är också ett fenomen som kan öppna möjligheter för en grad

förbättringar inom modellereingsområdet. Detta då man kan lära modeller att använda data som känns igen och med hjälp av detta anpassa ny data som uppkommer. Till exempel kan man tillämpa ”Neural Ordinary Differential Equations” på mekanistiska modeller och på så sätt få ut prediktioner från data utan att definiera ekvationssystemen själv på förhand [17]. Med denna utveckling kommer en mängd problem som behöver tas hänsyn till. Kan en AI vara tillräcklig pålitlig för att prediktera fortgången av en epidemi? Det arbete vi genomfört kan eventuellt användas som grund för att utveckla AI-baserade modeller och är därmed en viktig problemställning att uppmärksamma.

Något som inte togs hänsyn till i denna rapport är vaccinationens bromsning av smittspridningen av covid-19. Grunden till detta är att en mängd andra parametrar behöver hanteras och optimeras vilket leder till att mycket mer komplicerade modeller bör konstrueras. Dock finns stor potential för framtida problemställningar om just modellering med hänsyn till massvaccination för att utveckla dessa modeller mer.

# 6 Samhälleliga och etiska aspekter

Covid-19-pandemin har i över två års tid drabbat hela världens samhälleliga strukturer på en mängd olika plan. Allt från folkhälsan till statsekonomin världen över har påverkats, och flera år av återuppbyggnad kommer krävas för att återgå till hur det var innan pandemin.

Matematiska modeller, som de vi har behandlat i denna rapport, ska i huvudsak fungera som verktyg för samhällets olika organ, för att kunna se tillbaka på hur man tidigare hanterat pandemier och hur det påverkade smittans spridning och samhället i stort. Modellerna kan även användas för att förutspå framtida krissituationer. På så sätt kan de fungera som prediktioner för hur man bör agera för minsta påverkan på samhället i stort. Ett mer nischat område för våra modeller är planering inom sjukvården. Dels på grund av datan vi använt, som kommer från just inskrivningar på sjukhus, och dels på grund av det kortare tidsintervall som vi begränsat oss till. Modellerna förutspår smittans utveckling tre veckor framåt i tiden vilket skulle kunna användas för att planera sjukvården för kommande veckor.

Viktigt att tillägga är att modeller som dessa kan bli missvisande om de inte genomgår ständig granskning och hålls uppdaterade. Det är exempelvis inte hållbart att förlita sig på endast en modell. Detta beror på att sådana här modeller kan vara baserade på ett flertal olika parametrar som kan vara mer specifika för vissa områden. Till exempel konstruerades de modeller som vi hanterat på statistik från Västra Götalandsregionen vad gäller antal inskrivningar på sjukhus, samt befolkningens mobilitet under den tid som undersökts. Detta innebär att dessa modeller lämpar sig väl till regioner med liknande utbredning och storlek. Däremot uppstår problem då den appliceras på mindre befolkade områden eller för att ge ett helhetsperspektiv på covid-19-pandemins fortgång i hela Sverige.

En konsekvens av oförsiktighet eller förhastade slutsatser kan vara att felaktiga prognoser sprids i samhället. Detta leder till desinformation som är skadligt både för den offentliga och privata sektorn. Beroende på hur regeringar väljer att agera statspolitiskt utifrån de prediktioner som modellerna kan ge, påverkar detta hela samhället. Utifrån detta perspektivet kan man även diskutera hur korrupta regeringar kan använda felaktiga prognoser till sin egen fördel genom att styra befolkningens beteende. Ett viktigt arbete för att motverka detta är att se till att det finns ett tydligt och transparent samarbete mellan myndigheter och flera olika forskningsinstitutioner.

# 7 Slutsats

Sammanfattningsvis observerar vi från resultatavsnittet att MAPE över lag minskar med tiden och $IS_\alpha$ över lag ökar med tiden. Detta gäller för alla modeller förutom de logistiska modellerna med tak. Detta kan tolkas som att felet i modellerna minskar med mängden anpassningsdata medan osäkerheten av modellerna ökar. Logistiska metoden producerar bättre resultat så länge det är känt att infektionen endast kommer hålla i en våg och även bättre om taket kan beräknas. Vid prediktioner över en mer utdragen pandemi med flera vågor så lämpar sig en mekanistisk modell som SIR bättre då den kan modelera denna mer dynamiska ineffektivitet bättre.

# 8 Referenser


[1] D. Huremović, "Brief History of Pandemics (Pandemics Throughout History)", *Psychiatry of Pandemics*, s. 7–35, 2019. DOI: 10.1007/978-3-030-15346-5{_}2. URL: https://link.springer.com/chapter/10.1007/978-3-030-15346-5_2.

[2] "Visualizing the History of Pand Visualizing the History of Pandemics", URL: https://www.visualcapitalist.com/history-of-pandemics-deadliest.

[3] *WHO/Europe | Coronavirus disease (COVID-19) outbreak - About the virus*. URL: https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/novel-coronavirus-2019-ncov.

[4] *Smittspridning — Folkhälsomyndigheten*. URL: https://www.folkhalsomyndigheten.se/smittskydd-beredskap/utbrott/aktuella-utbrott/covid-19/om-sjukdomen-och-smittspridning/smittspridning/.

[5] A. Jöud, P. Gerlee, A. Spreco och T. Timpka, "Sammanställning och utvärdering av modeller för pandemiprediktion i Sverige under 2020", Chalmers tekniska högskola och Göteborgs universitet, tekn. rapport, 2021, s. 46. URL: https://libris.kb.se/bib/9pd99cc97s93zqr4.

[6] P. Gerlee och T. Lundh, *Vetenskapliga modeller : svarta lådor, röda atomer och vita lögner*. Studentlitteratur, 2012, ISBN: 9789144074207. URL: https://search.ebscohost.com/login.aspx?direct=true&db=cat07470a&AN=clc.bbbecf26.579d.49ce.b546.2bd319745313&site=eds-live&scope=site&authtype=guest&custid=s3911979&groupid=main&profile=eds.

[7] P. Gerlee, J. Karlsson, I. Fritzell m. fl., "Predicting regional COVID-19 hospital admissions in Sweden using mobility data", *Scientific Reports 2021 11:1*, årg. 11, nr 1, s. 1–8, dec. 2021, ISSN: 2045-2322. DOI: 10.1038/s41598-021-03499-y. URL: https://www.nature.com/articles/s41598-021-03499-y.

[8] A. Ruckstuhl, "Introduction to nonlinear regression", *IDP Institut für Datenanalyse und Prozessdesign. ZHAW Zürcher Hochschule für Angewandte Wissenschaften. stat. ethz. ch/stahel/courses/cheming/nlreg10E. pdf*, 2010.

[9] X. Yan och X. Su, *Linear Regression Analysis: Theory And Computing*. Singapore, SINGAPORE: World Scientific Publishing Company, 2009, ISBN: 9789812834119. URL: http://ebookcentral.proquest.com/lib/chalmers/detail.action?docID=477274.

[10] E. Stalidzans, M. Zanin, P. Tieri m. fl., "Mechanistic Modeling and Multiscale Applications for Precision Medicine: Theory and Practice", *https://home.liebertpub.com/nsm*, årg. 3, nr 1, s. 36–56, maj 2020. DOI: 10.1089/NSM.2020.0002. URL: https://www.liebertpub.com/doi/full/10.1089/nsm.2020.0002.

[11] M. Martcheva, *An Introduction to Mathematical Epidemiology*. Boston, MA: Springer US, 2015, vol. 61, ISBN: 978-1-4899-7611-6. DOI: 10.1007/978-1-4899-7612-3.

[12] *(PDF) Mathematical models for introduction, spread and early detection of infectious diseases in veterinary epidemiology*. URL: https://www.researchgate.net/publication/318394911_Mathematical_models_for_introduction_spread_and_early_detection_of_infectious_diseases_in_veterinary_epidemiology.

[13] H. Weiss, "The SIR model and the Foundations of Public Health", 2013.

[14] D. S. K. Karunasingha, "Root mean square error or mean absolute error? Use their ratio as well", *Information Sciences*, årg. 585, s. 609–629, mars 2022, ISSN: 0020-0255. DOI: 10.1016/J.INS.2021.11.036.

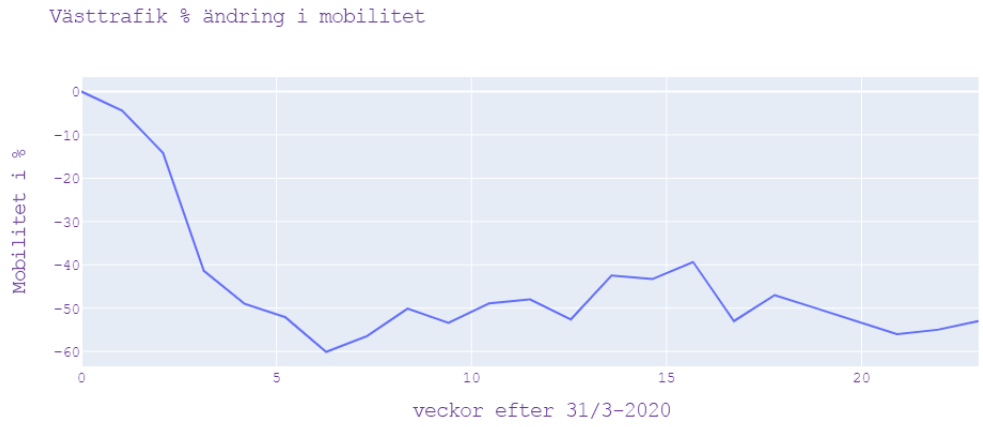
[15] A. de Myttenaere, B. Golden, B. Le Grand och F. Rossi, "Mean Absolute Percentage Error for regression models", *Neurocomputing*, årg. 192, s. 38–48, juni 2016, ISSN: 0925-2312. DOI: 10.1016/J.NEUCOM.2015.12.114.

[16] J. Bracherid, E. L. Ray, T. Gneitingid och N. G. Reichid, "Evaluating epidemic forecasts in an interval format", 2021. DOI: 10.1371/journal.pcbi.1008618.

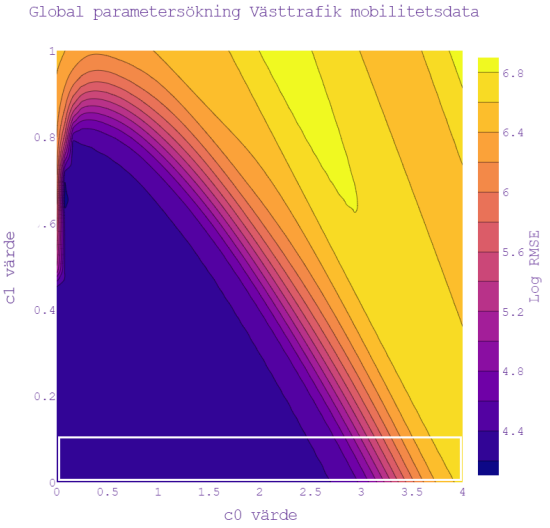
[17] R. T. Chen, Y. Rubanova, J. Bettencourt och D. Duvenaud, “Neural ordinary differential equations”, i *Advances in Neural Information Processing Systems*, vol. 2018-December, Neural information processing systems foundation, 2018, s. 6571–6583.

# 9 Appendix

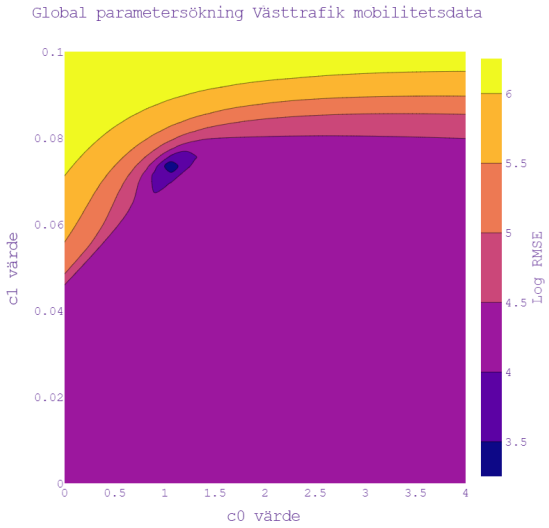
## 9.1 Figurer från SIR-modellen




Figur 8: Mobilitetsdata från Västtrafik i procentuell ändring från basvärdet.

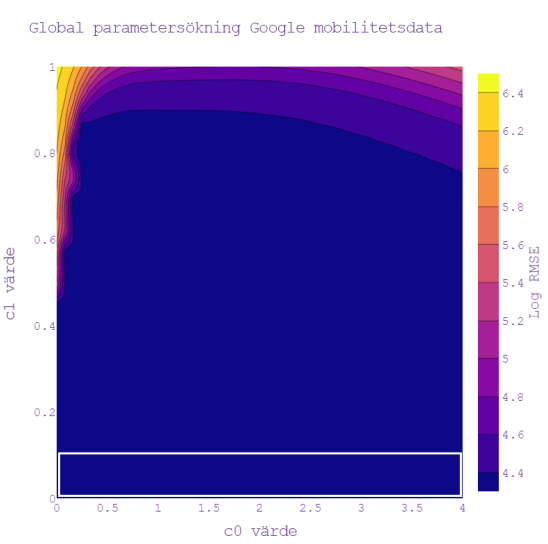



(a) Konturplot av log(RMSE) mot parameterrummet för Västtrafik mobilitetsdata.

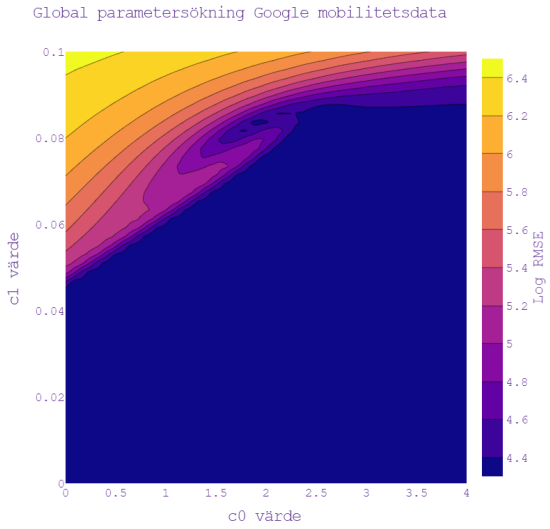



(b) Konturplot av log(RMSE) mot parameterrummet för mobilitetsdata från Västtrafik med förstoring av markerat område i (a) som sökområde för $c_1$.

Figur 9: Konturplot av log(RMSE) mot parameterrummet för mobilitetsdata från Västtrafik för olika stora sökområden.

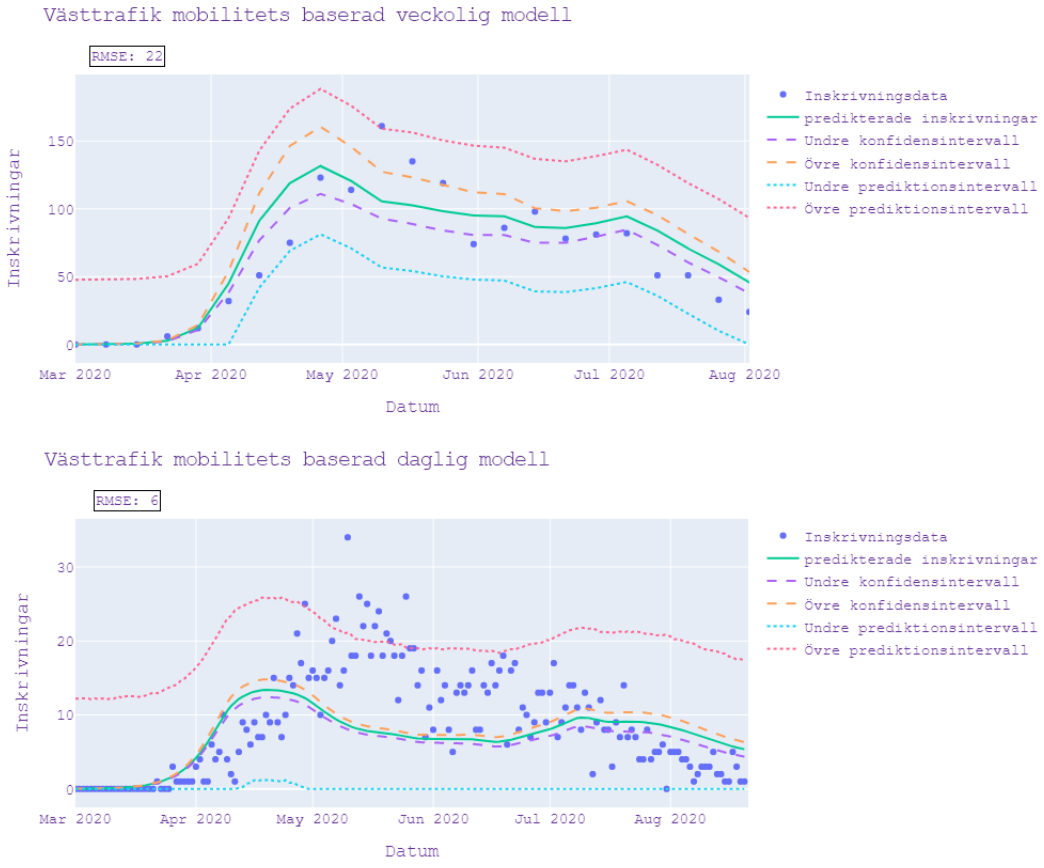



(a) Konturplot av log(RMSE) mot parameterrummet för mobilitetsdata från Google.




(b) Konturplot av log(RMSE) mot parameterrummet för mobilitetsdata från Google med förstoring av markerat område i (a) som sökområde för $c_1$.

Figur 10: Konturplot av log(RMSE) mot parameterrummet för mobilitetsdata från Google för olika stora sökområden.




Figur 11: Övre delen av figuren visar den veckobaserade modellen. Undre delen visar den dagsbaserade SIR-modellen plottad mot inskrivningsdata med konfidens- och prediktionsintervall.