Development of Forecast Models for COVID-19 Hospital Admissions using Mobile Network Data: A Privacy-Preserving Approach

Jalil Taghia (jalil.taghia@ericsson.com)
Ericsson AB, Ericsson Research

Valentin Kulyk
Ericsson AB, Ericsson Research

Selim Ickin
Ericsson AB, Ericsson Research

Mats Folkesson
Ericsson AB, Ericsson Research

Cecilia Nyström
Ericsson AB, Ericsson Business Area Managed Services

Kristofer Ågren
Telia Company AB

Thomas Brezicka
Sahlgrenska University Hospital, Department of Quality and Patient Safety

Tore Vingare
Sahlgrenska University Hospital, Department of Analysis and Project Management,

Julia Karlsson
Sahlgrenska University Hospital, Department of Analysis and Project Management,

Ingrid Fritzell
Sahlgrenska University Hospital, Department of Analysis and Project Management,

Ralph Harlid
Södra Älvsborg Hospital

Bo Palaszewski
Västra Götaland Regional Council

Magnus Kjellberg
Sahlgrenska University Hospital, AI Competence Center

Jörgen Gustafsson
Ericsson AB, Ericsson Research

Research Article

Keywords:

Posted Date: April 5th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1369613/v1

Development of Forecast Models for COVID-19 Hospital Admissions using Mobile Network Data: A Privacy-Preserving Approach

Jalil Taghia[1,*], Valentin Kulyk[1], Selim Ickin[1], Mats Folkesson[1], Cecilia Nyström[2], Kristofer Ågren[3], Thomas Brezicka[4], Tore Vingare[5], Julia Karlsson[5], Ingrid Fritzell[5], Ralph Harlid[6], Bo Palaszewski[7], Magnus Kjellberg[8], and Jörgen Gustafsson[1,*]

[1]Ericsson Research, Kista, 164 40 , Sweden
[2]Ericsson Business Area Managed Services, Kista, 164 40, Sweden
[3]Telia Company AB, 169 94 Solna, Sweden
[4]Sahlgrenska University Hospital, Department of Quality and Patient Safety, Gothenburg, 413 45, Sweden
[5]Sahlgrenska University Hospital, Department of Analysis and Project Management, Gothenburg 413 45, Sweden
[6]Södra Älvsborgs Sjukhus, Sjukhusledningen, Borås, 501 82, Sweden
[7]Department of Data Management and Analysis, Västra Götalandsregionen , Gothenburg, 405 44, Sweden
[8]Sahlgrenska University Hospital, AI Competence Center, Gothenburg, 413 45, Sweden
[*]corresponding author(s): Jalil Taghia and Jörgen Gustafsson (jalil.taghia@ericsson.com, jorgen.gustafsson@ericsson.com)

ABSTRACT

Reliable near-time forecast of COVID-19 hospital admissions can help enable effective resource management which is vital in reducing pressure from healthcare services. The use of mobile network data has come to attention in response to COVID-19 pandemic leveraged on their ability in capturing people social behaviour. Crucially, we show that there are latent features in irreversibly anonymised and aggregated mobile network data that carry useful information in relation to the spread of SARS-CoV-2 virus. We describe development of the forecast models using such features for near-time prediction of COVID-19 hospital admissions. In a case study, we verified the approach for two hospitals in Sweden, Sahlgrenska University Hospital and Södra Älvsborgs hospital, working closely with the experts engaged in the hospital resource planning. Importantly, the results of the forecast models were used in year 2021 by logisticians at the hospitals as one of the main inputs for their decisions regarding resource management.

Introduction

COVID-19 outbreaks have exhausted the healthcare around the world. Concentration of admitted patients during outbreaks and limited resources at hospitals put pressure on the healthcare systems. Knowing an estimated number of admitted patients in near-time horizons of two-to-three weeks can significantly facilitate management of the resources. Forecasts of the number of admitted patients can serve as an important input for prediction of the hospital resource allocation. However, developing forecast models for admitted COVID-19 patients has proven to be challenging[1–5] due to, among others, lack of historical data, involvement of many external factors, and most notably the evolving nature of COVID-19 outbreaks including the evolution of SARS-CoV-2[6], the dynamic nature of people behavioural response to external factors such as regulations set by the authorities[7], increasing number of people with antibodies, uncertainties in the measurements of antibodies, and evolution of antibody immunity to SARS-CoV-2[8].

Merely considering historical data on the number of admitted patients for prediction of the future number of admitted patients is not sufficient and can result in forecast models that lack novelty factor - in the sense that they fail to capture novel trends for which there are no precedence in the past. Capturing the trend changes help proactive decision making which is vital during outbreaks. Inclusion of external factors in the forecast models might improve the model efficacy in capturing the novel trends. However, it is not straightforward as there are many external factors for which it is difficult to determine their importance[9–14]. Examples of such external factors are various temporal seasonalities (e.g., weekly and monthly seasonalities), public holidays, events, weather condition, regulations set by authorities, and changes in behaviour of people at different phases of the pandemic. Aware of this, here, we argue in favor of using mobile network data of user activities as one of the main inputs for construction of the forecast models for near-time prediction of the number of admitted COVID-19 patients. We motivate use

of mobile activity data for development of the forecast models by their inherent ability in capturing social behaviour of people with respect to their physical movements in the society.

Inclusion of the mobile activity data in addition to the historical data on the number of admitted COVID-19 patients enables us to construct forecast models that maintain their novelty factor, leveraged on the approximate time lag between the point in time when people first come into contact with virus and the time when they are hospitalized. In other words, number of admitted patients at a given time can be explained in part by the patterns of people social activities in past registered by mobile network data of user activities. Our underlying hypothesis is that the user activities are positively correlated with the number of admitted patients, as the higher activity means concentration of more individuals in a limited area and in turn higher risks of exposure to SARS-CoV-2 virus.

The use of mobile activity data has seen several applications during the pandemic, such as to inform reopening strategies[15,16], for informing evidence-based policy making by authorities in attempt to manage the spread of SARS-CoV-2[17–19], early detection of COVID-19 outbreaks[20,21], and for informing COVID-19 forecast models[22]. In this work, we use mobile network data of user activities as one of the main inputs for construction of the forecast models of COVID-19 admission data motivated by their representative capabilities in capturing social behaviour of people in relation to COVID-19.

While in certain cases mobile network data contain useful information for analysis of COVID-19, the use of such data must be considered alongside understanding of their privacy risks, limits on their representative capabilities, effect of external factors on their usefulness, and the biases that they may impose[16]. In use of mobile activity data in construction of our forecast models, we took necessary steps to address these concerns. To mitigate privacy risks, user activities were irreversibly anonymised and aggregated hourly using a privacy-preserving technique that maintains overall representative capabilities of such data[17]. External factors that are related to the development of antibody in population, such as statistics on antibody-test positive rates and vaccination rates, can directly limit the representative capabilities of mobile activity data. To maintain the usefulness of mobile activity data, we introduce procedures that help limit the adverse effect of antibody development on their representative capabilities. Finally, locations at which user activities are measured may come with biases which can misinform the forecasts. Taking a data-driven approach instead of a hypothesis-driven approach helped reduce the effect of such biases in development of the forecast models.

In a case study, we used irreversibly anonymised and aggregated geographical grid-level hourly mobile network data of user activities in Västra Götaland county in Sweden provided by Swedish operator Telia Sverige AB, and developed forecast models for prediction of the number of admitted COVID-19 patients at Sahlgrenska University Hospital (SU) and Södra Älvsborgs hospital (SÄS). The predictions from the forecast models were provided as weekly inputs to the hospitals as part of the planning for their resource management. A notable aspect of this study is the collaboration between industry verticals: Telecom vendor Ericsson AB, Swedish mobile operator Telia Sverige AB, and Region Västra Götaland including the two hospitals SU and SÄS. We describe the development of the forecast models and discuss how the insight from the models were used in planning and prediction of healthcare demands and resources.

Results

Development of the forecast model

Development of the forecast models for the near future prediction of the number of admitted COVID-19 patients using mobile network activity data is one of the main results of this study. Our forecast model pipeline is composed of three interconnected models, namely: the grid selection model, the spatiotemporal model, and the prediction model. Figure 1 shows the key components of the forecast model. A detailed description of the mathematical formulation and algorithmic implementations are provided in Methods.

There are three types of input data provided to the forecast model, namely: (i) historical data on the number of daily admitted COVID-19 patients aggregated daily per hospital, hereafter referred to as the COVID-19 admission data, (ii) external factors in the form of antibody and vaccination data, and (iii) privacy-preserving anonymised and hourly aggregated mobile activity data. These data are described in Methods. The forecast model is fully data-driven. When it is seen as a module, it takes the three types of inputs and produces prediction of the number of admitted patients for the duration of the forecast window.

We constructed two 21-day forecast models, one for SU and one for SÄS, providing predictions for the duration of 21 days. These two forecast models share the same underlying architecture, however, they are optimized separately. We proceed with introducing the main three components of the forecast model pipeline.

Grid selection model. Mobile network data contain timeseries of aggregated hourly activities per grid in order of thousands. The grids are spread out across 49 municipalities in the Västra Götaland region, as shown in Supplementary Figure 1. While the hourly mobile activity data from the grids carry useful information about user activities in the area, not all the grids are equally relevant to the behavioural aspects related to COVID-19. Thus, there was a need for selection of the most relevant grids.

We opted for a data-driven approach for the grid selection such that selected grids of interest dynamically change over time as do user behaviours throughout the pandemic. Our approach in construction of the grid selection model is described in Methods. Seen as a module, as shown in Figure 1.a, it takes as its inputs both historical data on grid-level hourly mobile activity data and COVID-19 admission data. It then selects the clusters of grids that are most related to the user activities in connection to COVID-19. Grid selection was performed on a weekly basis as planning at the hospitals were done weekly. Figure 2 and Figure 3 illustrate the selected clusters of grids at selected analysis dates used for construction of the forecast models for SU and SÄS, respectively. Using tags taken from OpenStreetMap[23], one can see which geographical objects the selected clusters of grids represent at a given analysis date.

Spatiotemporal model. Hourly mobile activity data from selected clusters of grids contain useful spatial information about user activities. Additionally, these data are temporal in nature whose dynamics are affected not only by short-to-long range seasonalities, such as hourly and weekly seasonalities, but also various external factors, such as possibly evolution of antibody development and regulations set by authorities. This implies to the need for capturing temporal dynamics in modelling of such data. Our hypothesis was that there are certain temporal patterns hidden in the mobile activity data that are particularly useful for the analysis of COVID-19. Thus, the forecast model was equipped with a spatiotemporal model. As shown in Figure 1.b, the spatiotemporal model takes as its inputs hourly mobile activity data from selected clusters of grids, historical COVID-19 admission data and antibody data. It then constructs a spatiotemporal memory containing useful information about the short-to-long term dynamics in data. Specifically, the spatiotemporal memory contains latent spatiotemporal patterns in mobile activity data that satisfy the following two conditions: (i) they are one of the major patterns in the data, (ii) and they are either statistically positively or negatively correlated with the number of admitted patients. The mathematical formulation and algorithmic description of the spatiotemporal model are described in Methods. Figure 4 and Figure 5 show the (Pearson) correlation between the positively correlated spatiotemporal patterns and the daily number of admitted patients at SU and SÄS, respectively.

Prediction model. The forecast model is equipped with a predictive model in the form of a regressor. Seen as a module, as shown in Figure 1.c, the predictive model takes as its inputs all available historical frames of the spatiotemporal memories, historical data on vaccination data, and historical COVID-19 admission data. It then produces predictions for the number of admitted patients for the duration of the forecast window, 21 days. Here, we use a multilayer perceptron (MLP) regressor as the choice of predictive model. MLP belongs to the class of fully connected neural networks. The use of MLP as a simple neural network architecture in favor of recurrent neural nets (RNNs), which are designed for modelling temporal data, is motivated by the fact that the input to the MLP already contains carefully engineered short-to-long term spatiotemporal features. Being a non-temporal model, MLP would preserve the structure while RNNs will not. The construction of the predictive model is described in Methods.

Considerations in development of the forecast model

Validation of the forecast models. Validation of the forecast models for COVID-19 were challenging due to the dynamic nature of pandemic. We took the following approach for the validation of the forecast models. For a given analysis date, we divided available historical data into a train set and a validation set. The forecast model parameters were tuned guided by the results on the validation set. We varied the size of the validation set, from one week to six weeks to find the best setting for the parameters of the forecast model. The setting of the parameters that performed well on average across all validation sets were used for the final analysis, referred to as the optimal parameter setting. Next, we trained the model on the entire historical data, using the optimal parameter setting, which provided final forecasts for the duration of the forecast window. Such validation procedure was performed on a weekly basis for both SU and SÄS forecast models.

Evaluation of the forecast models. Evaluation of the forecast models were done based on both visual inspection by healthcare subject matter experts and objective measures. The visual inspection of the forecasts were done to examine model performance in capturing important trends in data. We found that using primarily objective measures for evaluation of the forecast models while useful can be sometimes misleading. As an example, a forecast model can miss out on capturing important trend changes while yet achieving reasonable performance based on the objective measures. It was found that the visual inspection of the predictions for evaluation of the forecast models, by the healthcare experts, can provide complementary insights.

Addressing the degeneracy problem of the forecast model. Training the forecast models involved minimizing a loss function between true and predicted number of admitted patients. We found that the forecast models often fall into degenerate solutions. The problem of degeneracy of a forecast model arises when a forecast model learns to "repeat the past" and by doing so achieves misleadingly a low loss. This may be explained by noting that the COVID-19 admission data generally can be seen as a stationary signal containing long steady state regions followed by sudden rare increasing or decreasing trends. The

main issue with the forecasts from a degenerate model is that they lack novelty and fail to predict novel trend changes. To reduce the degeneracy problem, we introduced a regularization to the loss function of the forecast models. The regularization was designed to discourage forecasts that are similar to the past and encourage uncovering novel trends. The addition of the regularization was the key in reducing the degeneracy problem in our forecast models for SU and SÄS. Construction of the regularization is discussed in Methods.

Inclusion of the external factors in the forecast model. Our main hypothesis in using mobile network data was that the user activities are positively correlated with the spread of SARS-CoV-2 virus such that the higher user activities are, the higher risk of spread of virus between individuals will be. As the level of antibody increased in population, user activities captured by the mobile activity data became less correlated with the spread of the virus. Antibody test data and vaccination data were the two external factors considered in the forecast models. Between the two, the vaccination data were given higher importance by the forecast model. In the case of the antibody test data, they were included indirectly through the spatiotemporal model while for the case of the vaccination data, they were included directly through the predictive model, as shown in Figure 1. In Methods, we discuss the mathematical formulations in inclusion of these external factors in the forecast models.

Near-time forecasts of the number of admitted patients at SU and SÄS

Figure 6.a shows the predicted number of admitted patients at SU during course of pandemic from February 15, 2021, until June 23, 2021, provided by the 21-day forecast model. Forecasts were delivered as inputs to SU on a weekly basis through 17 analysis dates (deliverable dates). That means the forecast models were run at various analysis dates while providing predictions for the duration of the next 21 days. Figure 6.b and Figure 6.c show the error in prediction per analysis date in terms of mean-absolute-error (MAE) score and the percentage error (relative error) between true and predicted number of admitted patients, averaged across the duration of the forecast window. For the resource-planning purpose, the forecasts from the most recent models were used by logisticians at SU. The most recent model is referred to the model built on using the latest available data at the time. Figure 6.d shows forecasts from the latest models. The evolution of the forecast models is highlighted with markers indicating the major changes to the forecast model.

Similarly, Figure 7.a shows the predicted number of admitted patients at SÄS from April 19, 2021, until July 4, 2021, provided by the 21-day forecast model. Forecasts were delivered as inputs to SÄS on a weekly basis through 14 analysis dates. Figure 7.b and Figure 7.c objectively evaluate the error in prediction per analysis date in terms of MAE score and the percentage error between true and predicted number of admitted patients, averaged across the duration of the forecast window. Figure 7.d shows forecasts from the latest models provided by the 21-day forecast model which were used by logisticians for resource-planning purpose.

Figure 6.c and Figure 7.c show the averaged percentage error across the entire duration of the forecast window, 21 days. Alternatively, we can study the quality of forecasts by partitioning the forecast window into three separate weeks, namely, the first week, the second week and the third week. Figure 8 shows the averaged percentage error per each partition for SU and SÄS.

Discussion

The goal of our project was to help in hospitals resource management alleviating the pressure for the healthcare services to meet the requirements for patient care during the COVID-19 pandemic. We hypothesized that privacy-preserving mobile network data of user activities, that are irreversibly anonymised and aggregated, were reflective of the social activity of people in terms of their physical movement in the society. To validate our underlying hypothesis, we took a model-based approach as we believed the aspects of the mobile network data, that are of interest for the analysis of COVID-19, are latent in the data. Thus, the idea was to extract those latent spatiotemporal patterns in mobile activity data that are of utmost relevance to the analysis of COVID-19 admission data.

The first step in achieving this goal was to extract the spatial information by selecting the geographical grids of interest to COVID-19. We first considered a hypothesis-driven approach in selection of the grids based on their geographical locations. However, the hypothesis-driven approach does not take into the dynamic nature of people behavioural response to the evolving pandemic situation. As an example, restrictions set by the authorities affect people social behaviour and that in turn affects grid activities differently. Therefore, we opted for a data-driven approach in selection of the grids. Figure 2 and Figure 3 show that the selected clusters of grids - and what they represent - change dynamically throughout the pandemic. As an example, the effect of season change in people behaviour has been captured in the selected grid clusters so that in winter time grids of interest are mostly concentrated around city center areas while in summer times, in addition to the city center areas, there are clusters of grids representing areas outside of the city such as parks and cottage areas.

The spatially relevant grid clusters selected by our grid selection model were in the form of timeseries and it was important to capture as well those temporal dynamics in the data that are relevant for the analysis of COVID-19 admission data. We modeled

the timeseries using the spatiotemporal model by decomposing data into a number of spatiotemporal patterns representing various temporal dynamics in the timeseries data. We then showed that there are indeed latent spatiotemporal patterns in mobile activity data that are statistically correlated with the number of admitted patients at the hospitals. Figures 4 and Figure 5 show the correlation scores for the positively correlated spatiotemporal patterns throughout the pandemic for SU and SÄS, respectively. We observed that the correlation scores were considerably higher for SU than SÄS. This could be explained partly by the fact that SU is a larger hospital and its catchment area includes municipalities with higher population densities than the ones for SÄS. Hence, user activities captured in mobile activity data are better reflective of the people behaviour in connection to COVID-19.

In spatiotemporal modelling of the mobile activity data for the extraction of the correlated spatiotemporal patterns, we considered various lags between mobile activity data and the number of COVID-19 admitted patients. The lag duration was varied from 7 days to 49 days with a step size of 7 days. For different lags, we computed Pearson correlation between spatiotemporal patterns extracted from mobile activity data and the number of admitted patients. We found that higher correlation scores are achieved for longer lags between 28 days and 45 days while, in most cases throughout the pandemic, the highest correlation score was achieved for 35-day lag. This is shown for SU in Figure 4 and for SÄS in Figure 5.

Leveraged on the predictive capabilities of the correlated spatiotemporal patterns with COVID-19, we built the predictive model of the number of admitted patients which uses these patterns as one of the main input features in addition to the historical COVID-19 admission data. We found the spatiotemporal patterns having a complementary role which proved useful for construction of our 21-day forecast models. This is explained as follows. Historical COVID-19 admission data on the number of admitted patients are better predictive of the future number of admitted patients for shorter lags (lags smaller than 7 days) and they lose their predictive relevance as the lag increases beyond 14 days. However, mobile activity data were shown to be most relevant for longer lags (28 to 42 days) but to have relatively limited predictive relevance for shorter lags, smaller than 7 days. Taking into account these two input features concurrently helped the forecast models to harvest useful information for near-time (i.e., 21 days) prediction of the number of admitted COVID-19 patients.

In addition to the historical COVID-19 admission data and mobile activity data, the forecast models were enriched with additional inputs provided by the external factors related to the development of the antibody in population, namely antibody test and vaccination data, when they became available. Purposefully, we did not include effect of external factors that are implicitly captured in mobile activity data, such as weather forecast, season, public transportation. Contrary to such external factors, antibody development plays against our hypothesis in using mobile activity, that is the user activities are positively correlated with the spread of SARS-CoV-2 virus such that the higher user activities are, the higher risk of spread of the virus between individuals will be. As the level of antibody increased in population, user activities captured by the mobile network data became less correlated with the spread of the virus. This can be seen by monitoring the correlation scores throughout the pandemic. As shown in Figures 6 and 7, the correlation scores were higher in the beginning and reduced as the pandemic evolved. This could be explained not only by the change in social behaviour of people but also partly by the evolution of antibody development in the population. To compensate for the effect of the antibody development in reducing predictive capabilities of the mobile activity data in relation to COVID-19, the effect of this external factor was explicitly considered in the forecast models. This was done by inclusion of antibody test data in the spatiotemporal model and vaccination data in the predictive model of the forecast model.

It is important to note that the forecast model pipeline was developed during pandemic. As the pandemic evolved, we needed to make changes to the forecast model. This is referred to as the evolution of the forecast model pipeline. Major changes to the models are highlighted in Figures 6 and 7. The changes to the forecast models are mostly related to the introduction of external factors to the forecast models. Antibody test data were included in the forecast models for SU on 2021-02-25 and on 2021-04-24 for SÄS. However, due to concerns regarding the quality of data, they were excluded from 2021-05-17 onward for both SU and SÄS forecast models. The antibody test data were included in the forecast models of SU and SÄS when the test was free of charge for public. Vaccination data were included in the forecast models of SU and SÄS on 2021-04-26. Initially, we did not have access to the age groups. From 2021-05-10, the effect of age group of the vaccinated population was included in the models as such data became available to us. Prior to 2021-05-25, we used linear effect in inclusion of the vaccination data. Since 2021-05-25, we used nonlinear effect where the non-linearity was learned from Israel vaccination experience[24] as described in Methods. In terms of the methodology, the only major change to the forecast model pipeline was related to the grid selection model. Since 2021-04-19, we changed the method of grid selection from the distance correlation to the periodograms, as described in Methods. The transformation of the timeseries data to periodograms added reliability and freedom to choose the seasonality related frequencies to be present in the data reflecting the latest state of the pandemic.

Forecast models for SU and SÄS were run regularly on a weekly basis as deliverables to the hospitals. At each deliverable, the forecasts for the duration of 21 days were provided to the logisticians at the respective hospitals. Figures 6 and 7 summarize the results. Crucially, in 16 out of 17 deliverables to SU, the percentage error, averaged across the duration of the forecast window of 21 days, was below 30% as shown in Figure 6.c. In the case of SÄS, excluding the analysis dates where the total

number of admitted patients were fewer than 15 patients, in 8 out of 9 deliverables, the percentage error was less than 30%.

In development of the forecast models, we have made several assumptions with respect to the input data, namely, data from external factors (i.e., antibody test and vaccination data), COVID-19 admission data, and mobile network data. As discussed earlier, through evolution of the forecast models, some of these assumptions were addressed - as an example effect of antibody development in the population was included in the forecast models through inclusion of vaccination data when they became available. However, some other assumptions remained throughout, which we believe addressing them could have improved the quality of the predictions from the forecast models. The first assumption is with respect to the use of mobile network data. In this study, we used data provided by Swedish operator Telia Sverige AB. Telia has the largest market share based on the number of mobile subscriptions in Sweden with about 34.6 percentage. We have made an assumption that the data from Telia is representative of the population. However, this assumption may have introduced potential bias in our analysis specially considering the age group of the base subscribers. The second assumption was with respect to the data from PCR testing which were not used by the forecast models. We believe including such data in the models, as an additional input, could have helped the forecast models - particularly, if such data were available at an early stage and were made on a population basis on all individuals presented with symptoms of COVID-19.

At SU, forecasts of the number of admitted COVID-19 patients were primarily processed by logisticians, as one of the key inputs for prediction of hospital beds for patients with COVID-19. It was done heuristically by adding the average of days the COVID-19 patients are hospitalized to each hospitalization case provided by the forecast model. Logisticians then could calculate the number of beds needed to take care of these patients. In addition to the forecast of admitted COVID-19 patients, at SU other inputs were used for the hospital resource management. These inputs varied over time because of changes in behaviour of the population. The inputs used at SU for the longest of time was analyzing the increase and decrease of positive PCR-tests from primary care (Vårdcentral) in Gothenburg which is the area where the patients hospitalized at SU live. Other inputs used during the pandemic were for example the increase or decrease in number of travelers at Västtrafik, which is the public transport company in Gothenburg and calls to Vårdguiden 1177 with the symptoms that are correlated to COVID-19 (Vårdguiden 1177 is a Swedish service providing healthcare by telephone and the central national infrastructure for Swedish healthcare online). SU also followed the content of virus in sewage water as yet another factor in their considerations[25]. These different inputs were weighted together and then presented to the group in charge of the hospital resource management. This group subsequently made decisions whether to open or close wards and beds dedicated for patients with COVID-19.

At SÄS, the forecasts on the number of admitted patients were used together with other indicators such as number of positive PCR in the community and cluster outbreaks in part of the region in order to make an estimation whether the number of admitted patients would increase, decrease or remain stable for the next 14-21 days. This estimation was used to adjust the estimated number of beds available for COVID-19 patients.

Collaboration between operative and academic departments have proved to be a key factor of success in the presented study. At the hospital level, there was a profound knowledge of how the disease itself influenced the need of both intensive care resources as well as of ordinary care facilities. It was observed that a rather constant factor of the admitted patients who were hospitalized needed intensive care (approximately 15%) and a higher fraction of the beds were occupied during the high waves of the pandemic by the same patients (approximately 25%). This insight called for a need of being able to estimate the number of patients that were admitted from time to time in order to always being able to correctly allocate resources to all patients who were imperatively in need of hospital care. SU found it extremely important to be able to on a weekly basis continuously forecast the number of patients that were admitted.

Collaboration with the academics and industry was regarded as a necessity in providing new opportunities for developing models that later proved very useful. The collaboration was performed with an open mind for the skill in each area that the different actors could provide. This resulted in a dynamic evolution of the knowledge of how models that could be of use could be produced. There were no economic or other constraints, such as a pre-designed overall protocol of research in the collaboration which allowed for free thinking that we feel is of great importance for developing this kind of models. Due to the immense complexity of a pandemic, such constraints would rather hinder than facilitate research that had to be performed at a reasonable pace in order to be operationally useful as the pandemic developed in its own unpredictable way. From the hospital perspective we have learned a lot from both the way the collaboration was initiated and performed and what kind of data we think would be of great use in future pandemic situation, in order to forecast the need of hospital resources. Decision makers can draw important insights from this work when it comes to the need of at early stages formulate sustainable strategies when it comes to recommendations to inhabitants how to behave, how to proceed with shut downs of municipal services and providing healthcare region testing for infection at a high level.

This project has been a collaborative effort between the two hospitals (SU and SÄS) and the private companies (Ericsson and Telia). It was initiated as an effort to handle the difficult and alarming pandemic situation. The rapid project initiation and the positive project outcome show the importance of forming and maintaining active networks across industries, both private and public sectors. The project can be seen as an excellent example of how society can benefit from digitalization, e.g., mobile

phones, mobile networks and data-driven model development. An interesting aspect of the project was that the project outcomes could be used timely in operational plannings. This was at first through insights from visualization of the mobile network data. However, as the project progressed more advanced outcomes, the forecast models, were generated and used gradually in practice. Throughout the project, close communication between the parties was prioritized and maintained at various stages of the project including problem formulation, interpretation of the results and proper usage of the forecast models.

Finally, the results would not have been possible without the close three-party collaboration, and certainly not in such a timely fashion that it was used while the pandemic was still ongoing.

Methods

Data

Mobile activity data. Mobile activity data used in this study were provided by Swedish operator Telia Sverige AB. Hourly mobile activities are obtained from user equipments (UEs) and are aggregated at the grid level. A grid is defined as a geographical square-like area, as shown in Supplementary Figure 1. The total number of activities are obtained per grid on hourly basis. It counts as one, if a UE is connected to the base station in a grid for at least 20 minutes. The same UE will not be counted again while it remains within the same grid[17]. The raw data are made privacy-preserving through a procedure consisting of anonymization, aggregation, and extrapolation[17].

Vaccination data. Aggregated vaccination data was supplied by Region Västra Götaland. In all analysis, we considered only the effect of dose 1.

Antibody data. From October 2020 the habitats in Region Västra Götaland could get an antibody test to verify whether they have antibodies for COVID-19, and the test was free of charge. In April 2021, the free-of-charge antibody test was cancelled, and the habitats had to pay for this test. The associated cost with the antibody test could have imposed biases in our forecast models as the results of the tests may not have been representative for the whole population. Therefore, the decision was made to not use antibody test data for the subsequent analysis.

COVID-19 admission data on the number of admitted patients. At SU, for a patient to be counted as an admitted COVID-19 patient, there should be a positive PCR-test at the earliest of 14 days before the hospitalization. Therefor the number of admitted COVID-19 patients consists of not only patients confirmed with COVID-19 at the hospital when they seek care but also if patients who have been tested positive for COVID-19 at for example a primary care center before seeking care at SU.

At SÄS, data on the daily number of COVID-19 admitted patient were provided by SÄS patient registering system. Patients that came to the emergency ward and were categorized as "pandemic cause" were tested and if found positive they were registered as admitted for COVID-19. The data are not public.

Preprocessing of mobile activity data

The mobile activity data were collected from geographical grids with dynamic sizes[17]. It was decided to normalize the data by the respective grid area. This changed the data units from raw mobile activity counts to mobile activity per square meter.

The outlier removal was a part of the cleaning procedure and was performed on the raw mobile activity data. It was based on two statistical concepts: (i) the median absolute deviations (MAD)[26], and (ii) kurtosis score computed from the historical data. A data point was marked as an outlier when its MAD distance and kurtosis score were greater than the respective thresholds.

Due to existence of trend changes (or concept drifts) in mobile activity data, we used double median absolute deviation (left-and-right MADs). The left-MAD was used to calculate the distance from the median of all points less than or equal to the median while the right-MAD was used to calculate the distance for points that were greater than the median. Thus, the MAD threshold was calculated as:

$$\Delta = \frac{\Lambda(X)}{\Gamma(X)}, \tag{1}$$

where $\Lambda(X)$ is absolute deviation of the timeseries X and $\Gamma(X)$ is the double MAD of X.

For a given data point, a kurtosis score was calculated using the historical data including the current data point. If the kurtosis score was larger than the threshold, the data point was flagged as the outlier. The threshold for the kurtosis was set experimentally to 3 which corresponds to the kurtosis value for a univariate normal distribution.

No handling of the missing data were required during preprocessing step of the mobile activity data due to the fact that our grid selection model could handle missing data.

As shown in Figure. 1.a, the grid selection model takes as its inputs (i) hourly mobile activity data from all grids, and (ii) daily COVID-19 statistics on the number of admitted patients. It then provides the hourly mobile activity data from grid clusters

that are best reflective of user activities in connection to COVID-19 admission data. The main components of the grid selection model are shown in Figure 9.a. Here, we describe these components.

At the first step, we convert both mobile activity timeseries and COVID-19 admission data into periodograms using Lomb-Scargle approach[27]. The mobile activity data contained missing values due to, among others, global mobile network outage, problems with the data collection, and mobile network maintenance. The duration of missing values varied from one to several days. The use of Lomb-Scargle approach was motivated by the fact that it can effectively handle the missing data. For our analysis, we decided to use the log-periodogram representations which are computed by taking natural logarithm of the periodograms. The use of logarithmic representations helps reduce the high-frequency noise in periodograms due to the measurement noise. To further reduce high-frequency noise in log-periodograms, a Butterworth low pass filter was applied to the representations.

Next, the log-periodogram representations were clustered using Ward hierarchical clustering approach[28]. To find the optimal number of clusters, we used the information theoretic measure of interaction information[29], which is a multivariate generalization of the mutual information. For this purpose, the interaction information was calculated between the cluster centers and COVID-19 admission data used as the prediction targets. The negative interaction was chosen as the criterion for selection of the number of clusters. The selection of the negative interaction over the positive interaction was experimentally verified. The number of clusters which resulted in the highest negative interaction information score was selected as the optimal number of clusters. Supplementary Figure 2 exemplifies the grid clusters resulting from the clustering step. Not all these clusters are reflective of people behaviour in connection to COVID-19. The final step was to select clusters of interest defined as grid clusters that are best related to COVID-19 admission data. We selected the smallest set of clusters with the highest negative interaction information.

Evolution of the grid selection model. The method of grid selection changed over the course of pandemic, as highlighted in Figures 6 and 7. The changes to the grid selection model are related to use of the timeseries conversion of the grids to the periodograms from 2021-04-19. Initially, the Pearson correlation distance was used to cluster the timeseries. However, the distance metric based on the Pearson correlation showed to be unreliable in combination with constantly growing size of the timeseries due to addition of new data on a weekly basis. The transformation of the timeseries data to periodograms added reliability and freedom to choose the seasonality related frequencies to be present in the data to be clustered. Additionally, that helped to test the timeseries clustering with different duration and allowed to find the best time duration reflecting the latest state of the pandemic. Thus, we selected several data chunks with different time duration, from 4 to 24 weeks into the past. These data chunks were clustered separately. The result was several clustering configurations. The clustering with the highest negative interaction information including the COVID-19 admission data was used for the subsequent selection of the cluster combination.

Spatiotemporal model

As shown in Figure. 1.b, the spatiotemporal model takes as inputs (i) the hourly mobile activity data from selected grid clusters, (ii) antibody test data, and (iii) daily COVID-19 statistics on the number of admitted patients. It then constructs a memory matrix of the spatiotemporal patterns in mobile activity data that are best related to the number of admitted COVID-19 patients, where the degree of relevance is measured in terms of the Pearson correlation score between the two. The main components of the spatiotemporal model are shown in Figure 9.b. Here, we describe these components and visualize them through an example - the example taken here is for the analysis date on 2021-06-07 for SU, and similar procedures were taken for other analysis dates for SÄS and SU.

At the first step, we obtain cluster representations of the grid clusters, where each cluster representation is the cluster's global center obtained by taking average of the grid timeseries in the cluster. The cluster representations are hourly mobile activity data averaged across grids, referred to as the hourly cluster-grid mobile activity representations. As an example, the grid clusters and the hourly cluster-grid mobile activity representations are shown in Supplementary Figure 3.

Next, these representations are modelled using Bayesian switching state space (BSDS) model[30]. BSDS is a temporal model which models data through a number of states. The states are modeled by Gaussian distributions where each state is fully described by the mean and covariance matrix of its corresponding Gaussian distribution. States are connected to each other through the first-order Markov chain, modeled via hidden Markov model (HMM). Throughout the analysis, maximum number of states was set to 10 where we relied on the Bayesian model selection in BSDS for deciding on the optimal number of states. Specifically, in Bayesian model selection of BSDS, those states that have insignificant contributions in describing data are assigned weights approaching zero and are pruned out by the model automatically.

Measures extracted from BSDS that we use here include occupancy rate, temporal evolution of states, and state posterior probabilities. For our working example, Supplementary Figure 4.a shows the state distributions where each (remaining) state is presented by a Gaussian distribution. Supplementary Figure 4.b shows the state posterior probabilities from which the temporal evolution of states are obtained, shown in Figure 4.c. The temporal evolution of states indicates which state is active at a given

time and at a given day. Supplementary Figure 4.d shows the occupancy rate of each state, computed from temporal evolution of states. The occupancy rate of states indicates the activity of the states and are between zero and one such that the most active state has the highest occupancy rate. We set a threshold equal to the median of the occupancy rates. The states with occupancy rates greater than the threshold are maintained and are referred to as the active set of states.

For the states in active set of states, we compute the state projections. State projections are computed by multiplying the state posterior probabilities (Supplementary Figure 4.b) to the daily cluster-grid mobile activity representations (Supplementary Figure. 3.c). The state projections are referred to as the spatiotemporal patterns. Note that spatiotemporal patterns are only computed for the states in the active set of states. Given the daily spatiotemporal patterns, we compute the correlation between these patterns and the COVID-19 admission data on the daily number of admitted patients. Those spatiotemporal patterns that are statistically positively or negatively correlated with the daily number of admitted patients are maintained and are referred to as the correlated spatiotemporal patterns. Note, that we discard the spatiotemporal patterns that are not statistically correlated. The correlated spatiotemporal patterns are those latent patterns in mobile activity data that are most related to the COVID-19 admission data. The correlated spatiotemporal patterns are shown in Supplementary Figure. 5.a and their correlation scores are shown in Supplementary Figure. 5.b.

Finally, we build a spatiotemporal memory from the extracted correlated spatiotemporal patterns by stacking the correlated spatiotemporal patterns into a matrix. The size of the memory matrix varies based on the number of correlated spatiotemporal patterns extracted from data.

Predictive model

As shown in Figure. 1.c, the predictive model takes as its inputs historical frames of the spatiotemporal memory matrix, historical data on the vaccination data, and COVID-19 admission data. It then produces forecasts for the duration of the forecast window. We use a multilayer perceptron (MLP) regressor as the choice of predictive model. MLP belongs to the class of fully connected neural network. Table 1 shows the design choices in construction of the MLP neural network. The same neural network architecture was used for SU and SÄS. Note that, the input later dimensionality depends on the size of the input spatiotemporal memory matrix which is determined automatically by the spatiotemporal model depending on the data under consideration for the given analysis date.

Reducing degeneracy problem of the forecast models. To reduce the degeneracy problem, we modified the loss function by adding a regularization term to the loss function. The regularization term discourages forecasts that are similar to the past and hence encouraging uncovering of novel trends. Let L denote the loss between the true number of admitted patients Y_t and the predicted number of admitted patients $\hat{Y}_t$ at the day t, defined as:

$$L = \sum_{t=1}^{T} \ell\left(Y_t, \hat{Y}_t\right), \tag{2}$$

where T is the number of days in the training set, and ℓ denotes the mean-square-error between true and predicted number of admitted patients. We obtain a regularized version of the loss defined as:

$$L_{\text{regularized}} = \ell(Y_{0:W}, \hat{Y}_{0:W}) + \sum_{t=1}^{T} (1-\beta)\ell\left(Y_{iW:(i+1)W}, \hat{Y}_{iW:(i+1)W}\right) - \beta\ell\left(\hat{Y}_{(i-1)W:iW}, \hat{Y}_{iW:(i+1)W}\right). \tag{3}$$

In above equation, W is the length of the forecast window; $Y_{iW:(i+1)W}$ denotes the W elements of the true number of admitted patients at the frame i and $\hat{Y}_{iW:(i+1)W}$ is the corresponding predicted number of admitted patients; β is a regularization term where $0 < \beta < 0.5$. The second term in the summation is a regularization term which penalizes the loss if the forecasts at frame i is similar to the ones at the previous frame, $i-1$. The optimal value of β shall be chosen thorough cross validation. In all analysis, β was set to 0.25. The ML neural network is trained using the regularized loss in Equation. 3

Inclusion of the antibody data in forecast model. The effect of antibody test data was included in the forecast model through the spatiotemporal model. For that, the user mobile activities are scaled up through an antibody curve computed from the cumulative statistics on the antibody positive rate, as

$$Z_t \leftarrow (1+\gamma_t) Z_t, \tag{4}$$

where Z_t denote the average user activities at a given day t, and γ_t is the cumulative number of antibody positive rate at the day t. The antibody curve is shown in Figure 10.a.

Inclusion of the vaccination data in forecast model. The effect of vaccination data was included in the forecast model through the predictive model. Before introduction of the vaccination data to the forecast model pipeline, the predictive model at the learning phase was trained to learn the following problem

$$\hat{Y}_t = f(X_t), \quad \forall t, \tag{5}$$

where f is the predictive model which takes as the inputs X_t and produces the forecasts $\hat{Y}_t$. After introduction of the vaccination data, the predictive model was trained to solve instead the following problem,

$$\hat{Y}_t = (1 - g(\alpha_t))f(X_t), \quad \forall t \tag{6}$$

where g is the vaccination efficacy model, and α_t is the vaccination rate at the day t computed from the cumulative number of vaccinated individuals normalized by the population. Effect of vaccination data was included in the forecast models from 2021-04-26. Up until 2021-05-25, we considered linear effect, that means $g(\alpha_t) \approx \alpha_t$. However, since 2021-05-25 onward, we considered a nonlinear effect where the vaccination efficacy model g was learned from the vaccination efficacy curve computed from vaccination campaign in Israel[24]. Figure 10.b shows the vaccination efficacy curve from which vaccination efficacy model g was learned by fitting a polynomial model, as discussed in Methods.

Data availability

Data permission For both anonymous and aggregated mobile network data and vaccination data, we obtained permissions to use. For using anonymous and aggregated mobile network data, we have signed legal agreements with Telia Sverige AB. The vaccination data were provided directly from Swedish healthcare, Region Västra Götaland. Both Telia Sverige AB and Region Västra Götaland were part of the project.

Data availability statement Vaccination data can be shared upon request. Anonymous and aggregated mobile network data are confidential and cannot be shared publicly. Anonymous and aggregated mobility data from Telia Sverige AB is commercially available as part of the service "Telia Crowd Insights." For more information, please see https://business.teliacompany.com/crowd-insights

References

1. Bertozzi, A., Franco, E., Mohler, G. O., Short, M. B. & Sledge, D. The challenges of modeling and forecasting the spread of COVID-19. *Proc. Natl. Acad. Sci. United States Am.* **117**, 16732 – 16738 (2020).
2. Baker, R. E., Park, S. W., Wagner, C. E. & Metcalf, C. J. E. The limits of SARS-CoV-2 predictability. *Nat. ecology & evolution* (2021).
3. Ioannidis, J. P. A., Cripps, S. & Tanner, M. A. Forecasting for COVID-19 has failed. *Int. J. Forecast.* (2020).
4. Moein, S. *et al.* Inefficiency of sir models in forecasting COVID-19 epidemic: a case study of Isfahan. *Sci. Reports* **11** (2021).
5. Dolton, P. J. The statistical challenges of modelling COVID-19. *Natl. Inst. Econ. Rev.* **257**, 46–82 (2021).
6. Singh, D. & Yi, S. V. On the origin and evolution of SARS-CoV-2. *Exp. & Mol. Medicine* **53**, 537 – 547 (2021).
7. Bavel, J. J. V. *et al.* Using social and behavioural science to support COVID-19 pandemic response. *Nat. Hum. Behav.* **4**, 460–471 (2020).
8. Gaebler, C. *et al.* Evolution of antibody immunity to sars-cov-2. *Nature* **591** (2021).
9. Lim, Y. K., Kweon, O. J., Kim, H. R., Kim, T.-H. & Lee, M.-K. The impact of environmental variables on the spread of COVID-19 in the Republic of Korea. *Sci. Reports* **11** (2021).
10. Poirier, C. *et al.* The role of environmental factors on transmission rates of the COVID-19 outbreak: An initial assessment in two spatial scales. *Sci. Reports* **11** (2020).
11. Azuma, K. *et al.* Environmental factors involved in SARS-CoV-2 transmission: effect and role of indoor environmental quality in the strategy for covid-19 infection control. *Environ. Heal. Prev. Medicine* **25** (2020).
12. Bherwani, H., Gupta, A., Anjum, S. G., Anshul, A. & Kumar, R. Exploring dependence of COVID-19 on environmental factors and spread prediction in India. *npj Clim. Atmospheric Sci.* **3**, 1–13 (2020).
13. Mccoy, D., Mgbara, W., Horvitz, N., Getz, W. M. & Hubbard, A. E. Ensemble machine learning of factors influencing COVID-19 across US counties. *Sci. Reports* **11** (2021).

14. Roy, S. & Ghosh, P. Factors affecting COVID-19infected and death rates inform lockdown-related policymaking. *PLoS ONE* **15** (2020).

15. Chang, S. H. *et al.* Mobility network models of COVID-19 explain inequities and inform reopening. *Nature* **589**, 82–87 (2020).

16. C, I. *et al.* Public mobility data enables COVID-19 forecasting and management at local and global scales. *Sci. Reports* **11** (2021).

17. Ågren, K., Bjelkmar, P. & Allison, E. The use of anonymized and aggregated telecom mobility data by a public health agency during the COVID-19 pandemic: Learnings from both the operator and agency perspective. *Data & Policy* **3** (2021).

18. Grantz, K. H. *et al.* The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology. *Nat. Commun.* **11** (2020).

19. Persson, J., Parie, J. F. & Feuerriegel, S. Monitoring the COVID-19 epidemic with nationwide telecommunication data. *Proc. Natl. Acad. Sci. United States Am.* **118** (2021).

20. Guan, G. *et al.* Early detection of COVID-19 outbreaks using human mobility data. *PLoS ONE* **16** (2021).

21. Wu, S., Fan, X., Chen, L., Cheng, M. & Wang, C. Predicting the spread of COVID-19 in china with human mobility data. *Proc. 29th Int. Conf. on Adv. Geogr. Inf. Syst.* (2021).

22. García-Cremades, S. *et al.* Improving prediction of COVID-19 evolution by fusing epidemiological and mobility data. *Sci. Reports* **11** (2021).

23. OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org (2017).

24. Rossman, H. *et al.* COVID-19 dynamics after a national immunization program in Israel. *Nat. medicine* (2021).

25. Saguti, F. *et al.* Surveillance of wastewater revealed peaks of SARS-CoV-2 preceding those of hospitalized patients with COVID-19. *Water Res.* **189**, 116620–116620 (2020).

26. Leys, C., Ley, C., Klein, O., Bernard, P. & Licata, L. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *J. Exp. Soc. Psychol.* **49**, 764–766 (2013).

27. Vanderplas, J. Understanding the Lomb–Scargle periodogram. *Astrophys. J. Suppl. Ser.* **236**, 16 (2018).

28. Jr., J. H. W. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **58**, 236–244 (1963).

29. Jakulin, A. & Bratko, I. Quantifying and visualizing attribute interactions. *ArXiv* **cs.AI/0308002** (2003).

30. Taghia, J. *et al.* Uncovering hidden brain state dynamics that regulate performance and decision-making during cognition. *Nat. Commun.* **9** (2018).

31. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. & LeCun, Y. (eds.) *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (2015).

Author contributions statement

The forecast model pipeline was designed and developed by J.T., V.K., S.I. M.F., C.N. and J.G. at Ericsson. Aggregated and anonymised mobile network activity data were provided by Telia. K.Å. at Telia provided insights about how data were collected and processed. Vaccination data were provided by Västra Götalandsregionen (VGR), and B.P. and M.K. provided insights in the usage of these data. COVID-19 admission data were directly provided by Sahlgrenska University Hospital (SU) and Southern Älvsborg Hospital (SÄS). Analysis were carried out at Ericsson by J.T., V.K., S.I. M.F., C.N. and J.G. Forecasts were analyzed by T.B., T.V., J.K. and I.F. at SU, and by R.H. at SÄS. All co-authors contributed in writing and revising the manuscript.

Competing interests

The authors declare no competing interests.

Figures & Tables

Table 1. MLP neural network used as the predictive model of the forecasts models.

Net	Number of Units	Activation Function	Batch Normalization	Drop Out
Input Layer	D_{input}	ReLU	True	0.2
Hidden Layer 1	50	ReLU	True	0.2
Hidden Layer 2	50	ReLU	True	0.2
Hidden Layer 3	50	ReLU	True	0.2
Output Layer	1	Linear	False	False
Loss Function	Mean-Square Loss			
Optimizer	Adam[31], learning rate = 0.001			

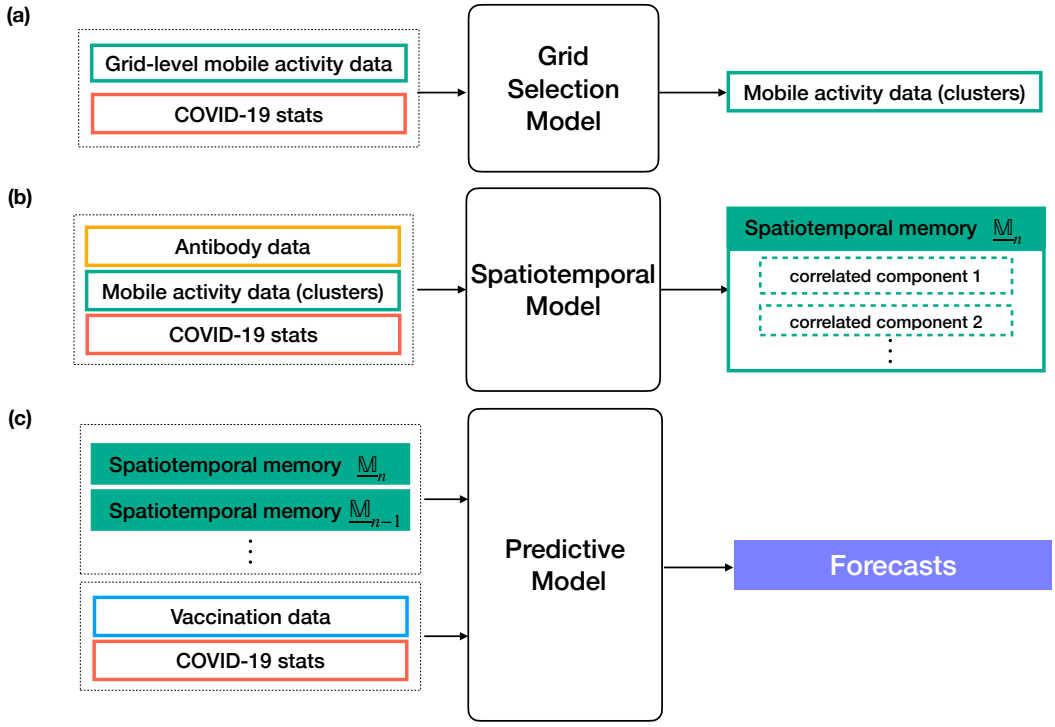

Figure 1. Modular description of the forecast model pipeline and its main components. (a) The grid selection model selects clusters of grids from mobile activity data that are best reflective of user activities in relation to COVID-19. (b) The spatiotemporal model uses hourly mobile activity data from selected clusters of grids and constructs a spatiotemporal memory of correlated components. These correlated components are the latent spatiotemporal patterns in mobile activity data that are either statistically positively or negatively correlated with daily number of admitted COVID-19 patients. (c) The predictive model uses historical frames of spatiotemporal memory matrices, together with historical COVID-19 admission data and vaccination data, in order to produce the predicted number of admitted patients for the duration of the forecast window.

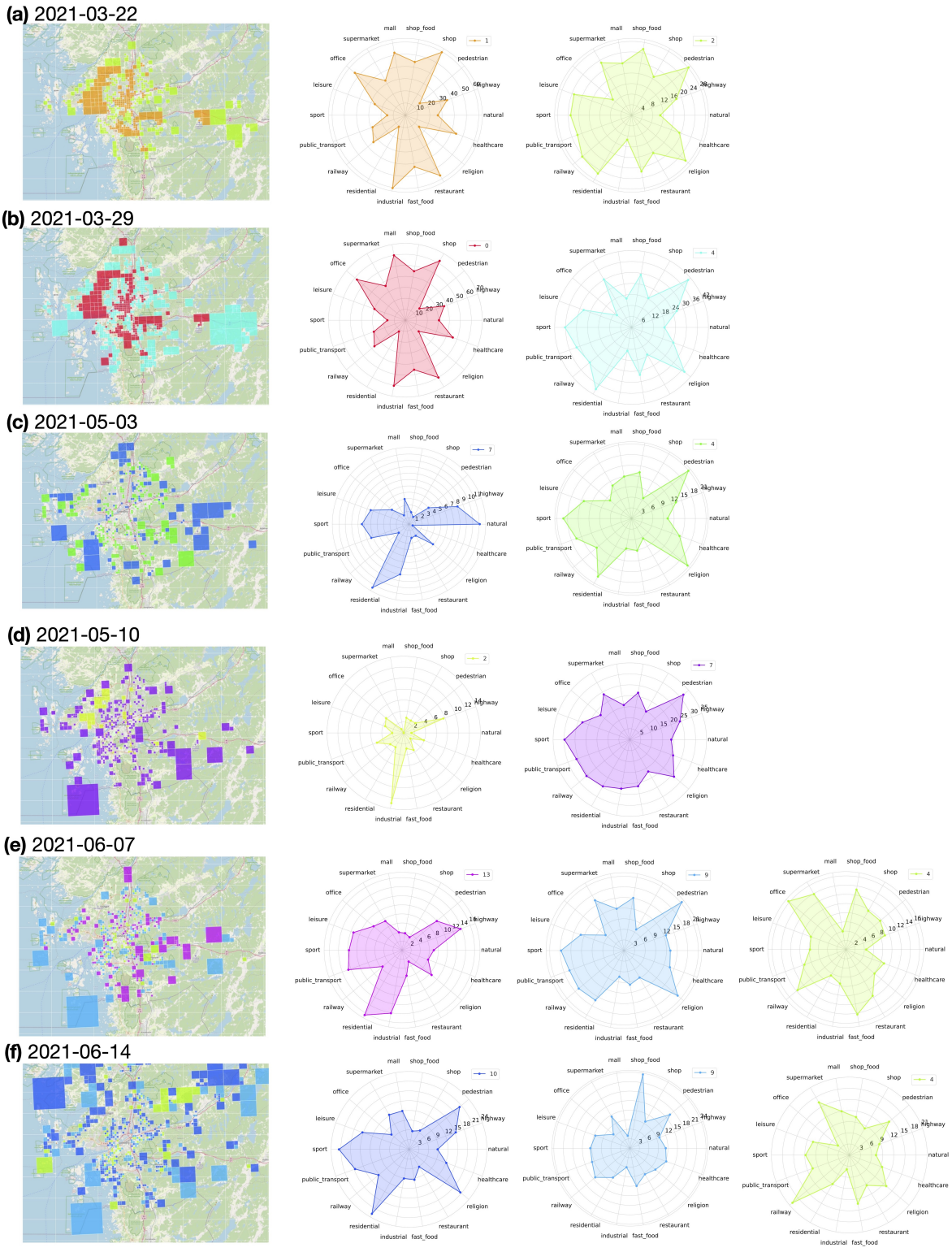

Figure 2. Selected clusters of grids by the grid selection model used in construction of the forecast models for Sahlgrenska University Hospital. (a-f) The figure shows the results for selected analysis dates. Note that there is no one-to-one correspondence between the same colors across two arbitrary analysis dates.

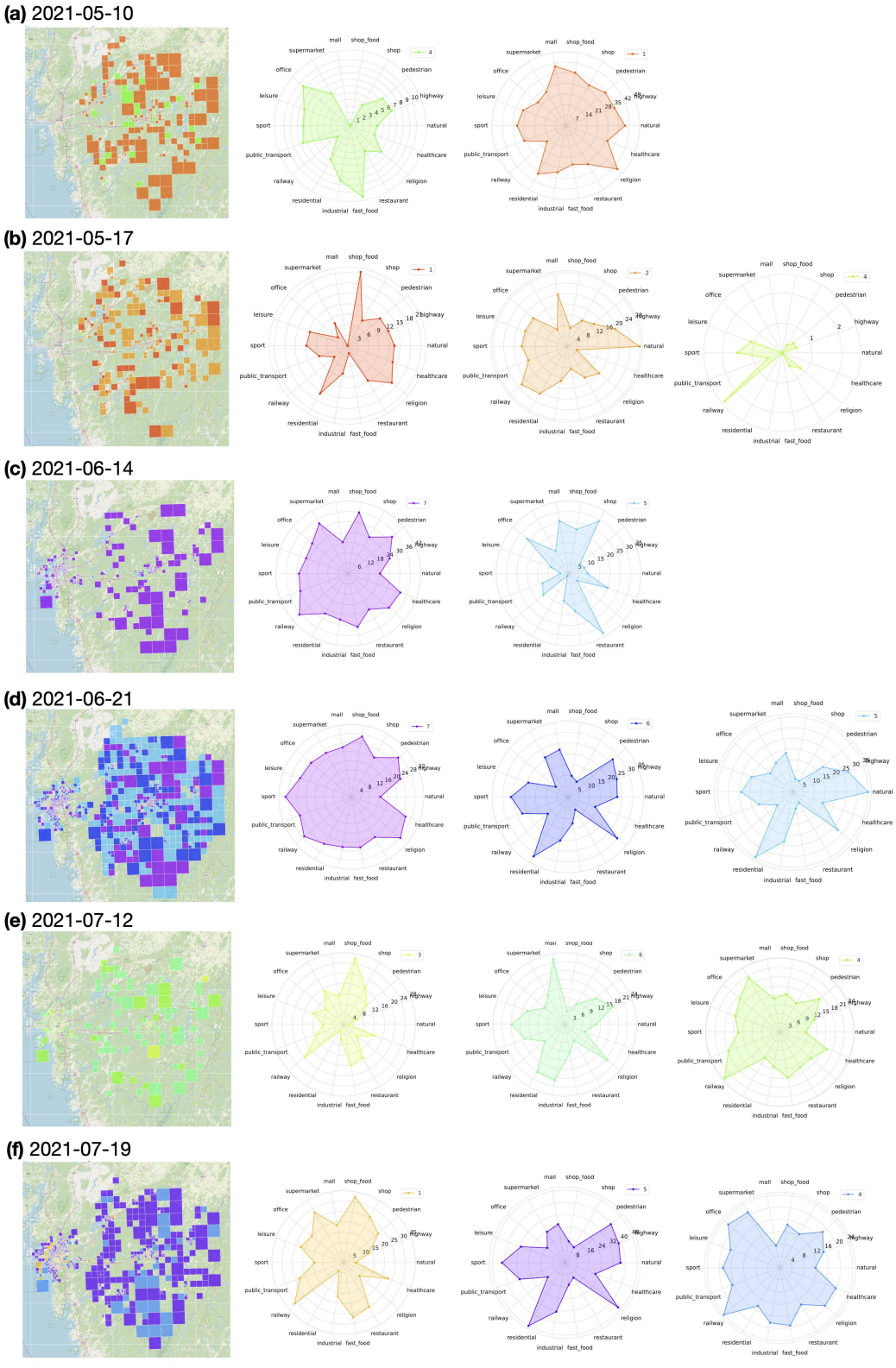

Figure 3. Selected clusters of grids by the grid selection model used in construction of the forecast models for Södra Älvsborgs hospital. (a-f) The figure shows the results for selected analysis dates. Note that there is no one-to-one correspondence between the same colors across two arbitrary analysis dates.

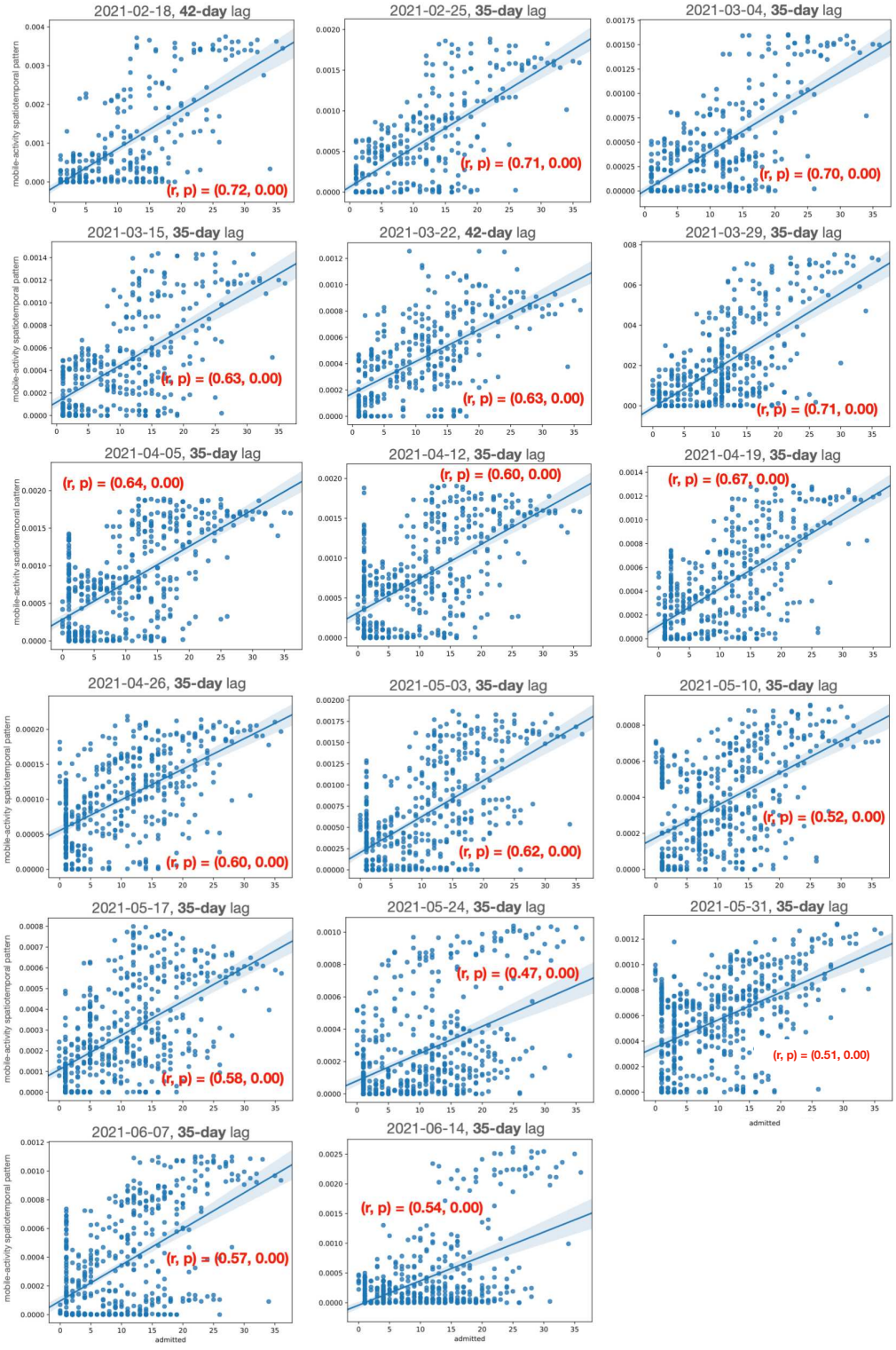

Figure 4. Pearson correlation between the positively correlated spatiotemporal component and the daily number of admitted patients at SU. The positively correlated spatiotemporal components are extracted from mobile activity data by the spatiotemporal model. The figure shows the correlation (r) score and the statistical significance (p) score across various analysis dates. The figures also show the optimal time lag at which the Pearson correlation between the spatiotemporal components and daily number of admitted COVID-19 patients is at its highest value. The optimal time lag is 35 days for all analysis dates except for 2021-02-18 and 2021-03-22 which is 42 days.

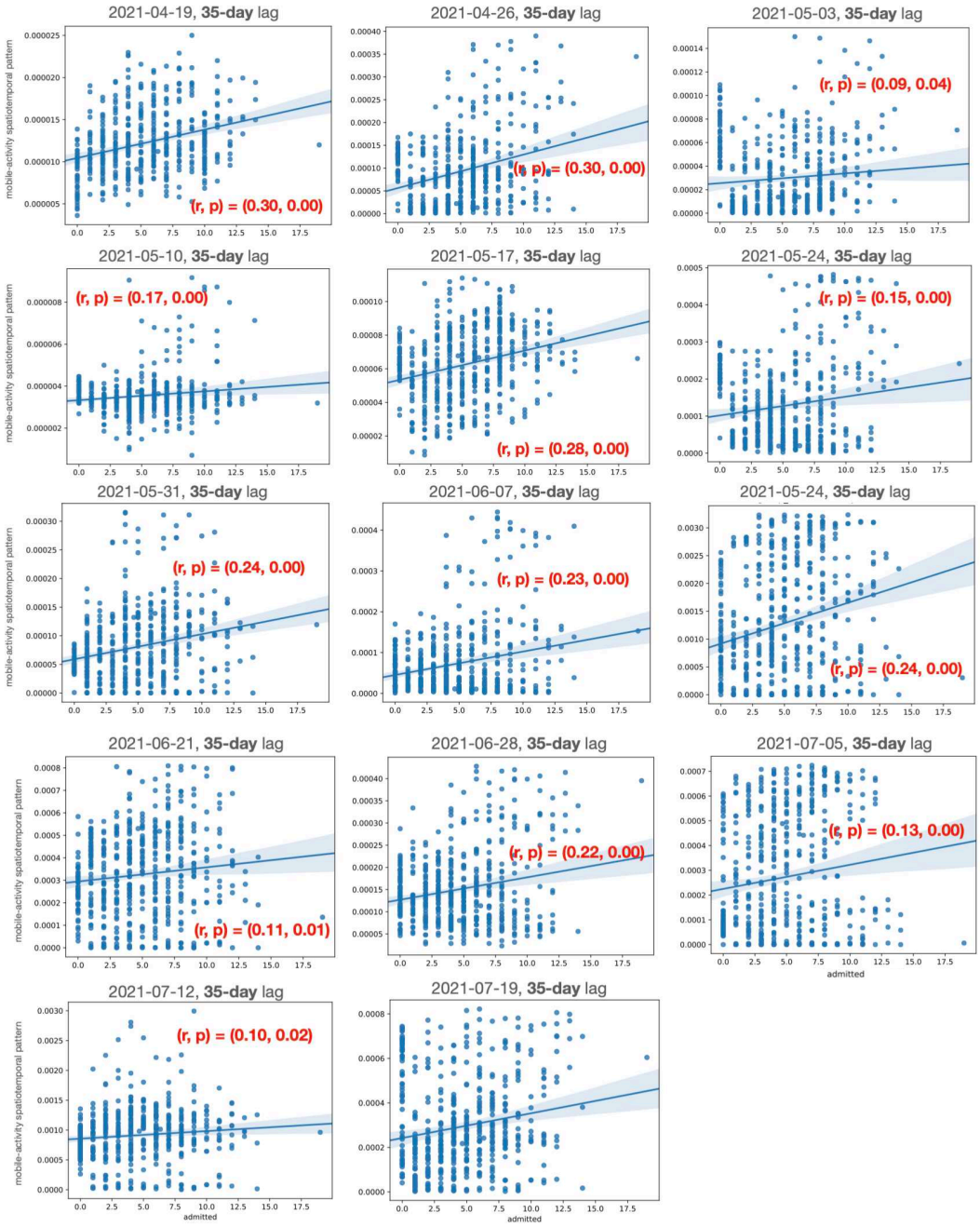

Figure 5. Pearson correlation between the positively correlated spatiotemporal component and the daily number of admitted patients at SÄS. The positively correlated spatiotemporal components are extracted from mobile activity data by the spatiotemporal model. The figures show the correlation (r) score and the statistical significance (p) score across various analysis dates. The figures also show the optimal time lag at which the Pearson correlation between the spatiotemporal components and daily number of admitted COVID-19 patients is at its highest value. The optimal time lag is 35 days for all analysis dates.

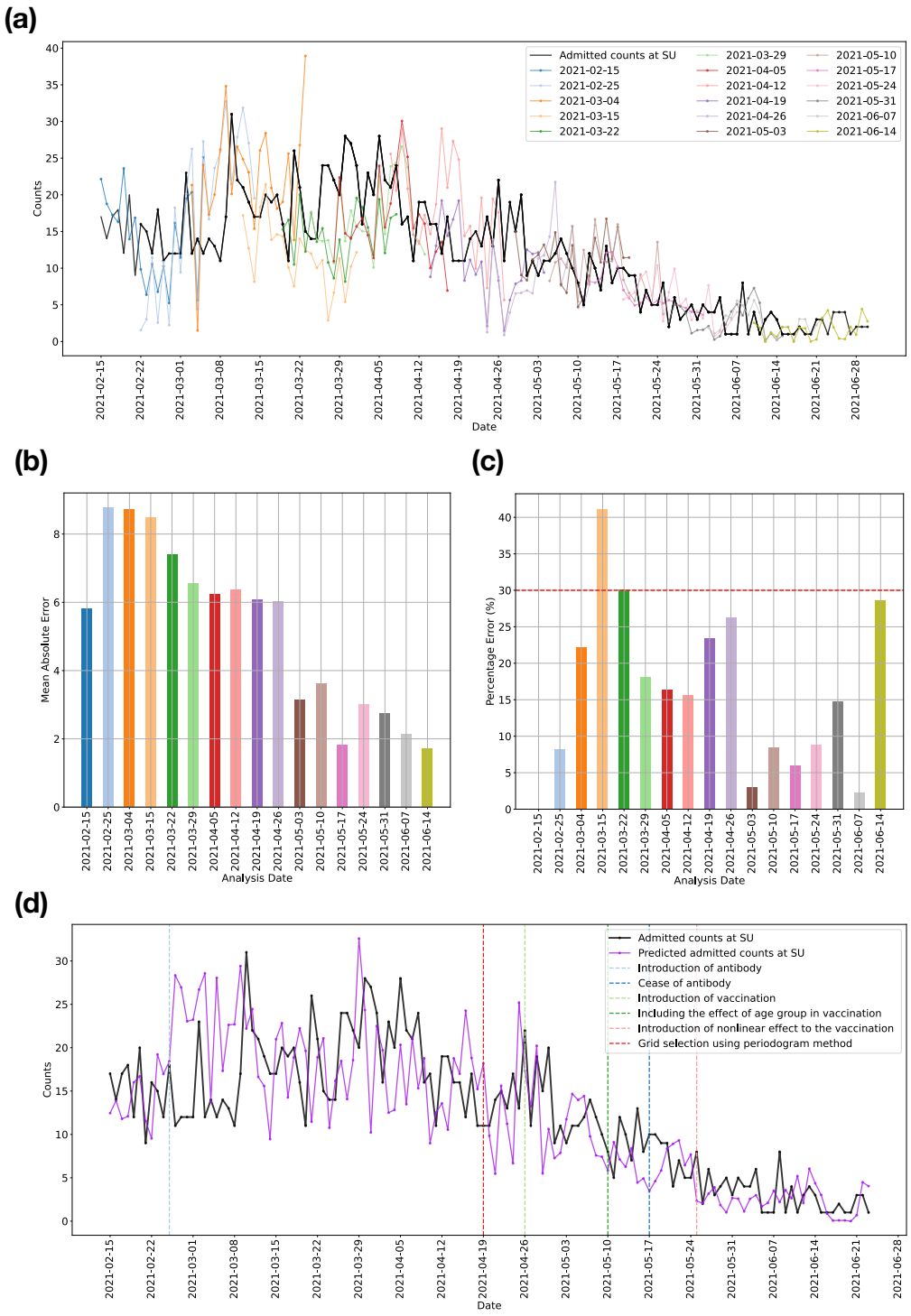

Figure 6. Prediction of the number of admitted COVID-19 patients at Sahlgrenska University Hospital (SU). Forecast models were run regularly for 17 analysis dates (deliverable dates) starting from February 15, 2021 until June 14, 2021. (a) Forecasts from the 21-day forecast model at each analysis date. (b) Error in prediction in terms of the mean-absolute-error between true and predicted number of admitted patients, averaged across the duration of the forecast window. (c) Error in prediction in terms of the percentage error score per analysis date, averaged across the duration of the forecast window. Across all analysis dates, the total number of admitted patients were more than 15 patients. In 16 out of 17 analysis dates, the error is less than 30%. (d) Prediction of the number of admitted patients considering forecasts from latest models, built using the latest available data. The evolution of the forecast models is highlighted with markers indicating the major changes to the forecast model.

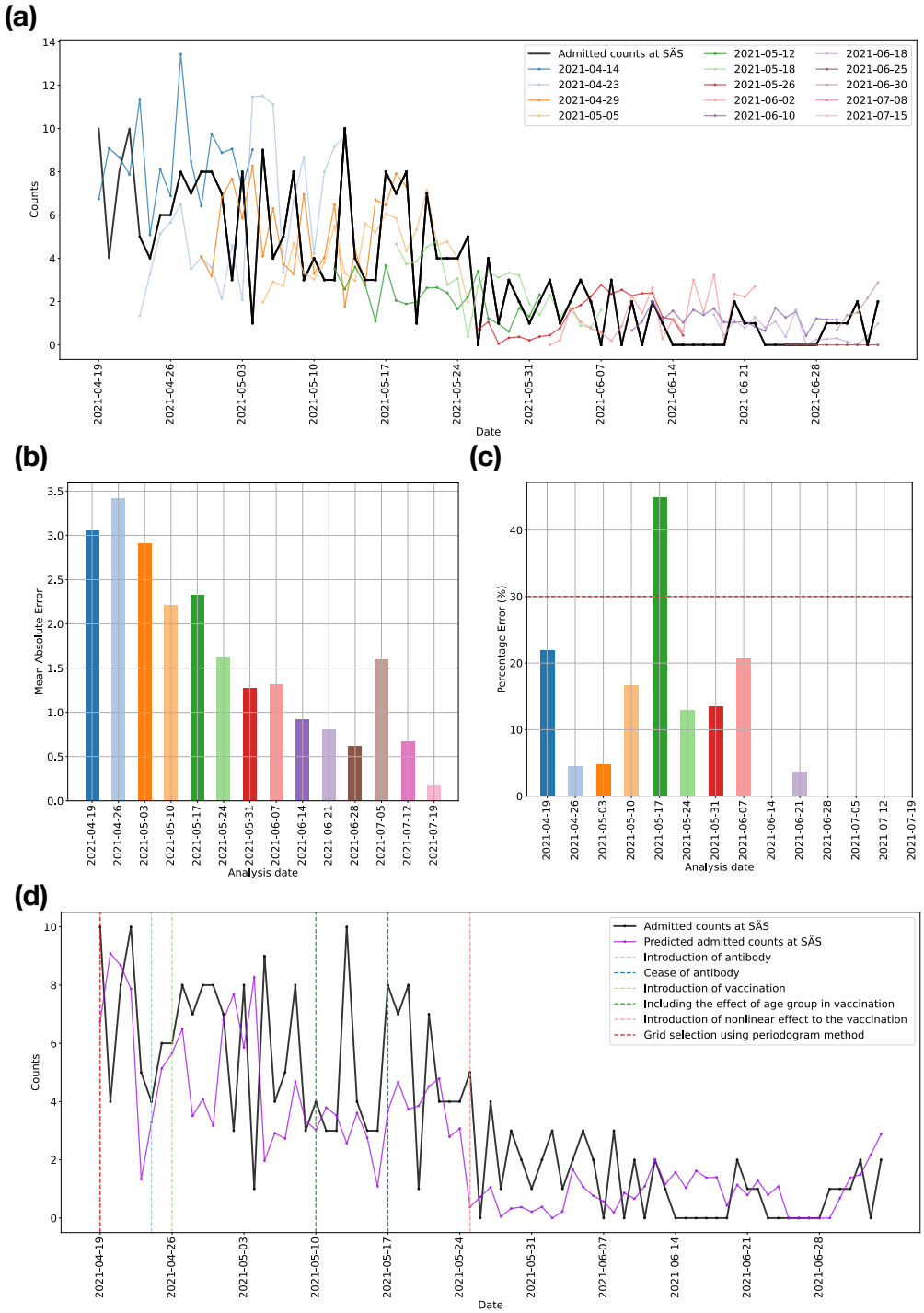

Figure 7. Prediction of the number of admitted COVID-19 patients at Södra Älvsborgs hospital (SÄS). Forecast models were run regularly for 14 analysis dates (deliverable dates) starting from April 19, 2021 until July 19, 2021. (a) Forecasts from the 21-day forecast model at each analysis date. (b) Error in prediction in terms of the mean-absolute-error between true and predicted number of admitted patients, averaged across the duration of the forecast window. (c) Error in prediction in terms of the percentage error score per analysis date, averaged across the duration of the forecast window. Percentage errors are not shown for those analysis dates for which the total number of admitted patients were fewer than 15 patients. In 8 out of 9 analysis dates, for which the total number of admitted patients were more than 15 patients, the error is less than 30%. (d) Prediction of the number of admitted patients considering forecasts from latest models, built using the latest available data. The evolution of the forecast models is highlighted with markers indicating the major changes to the forecast model.

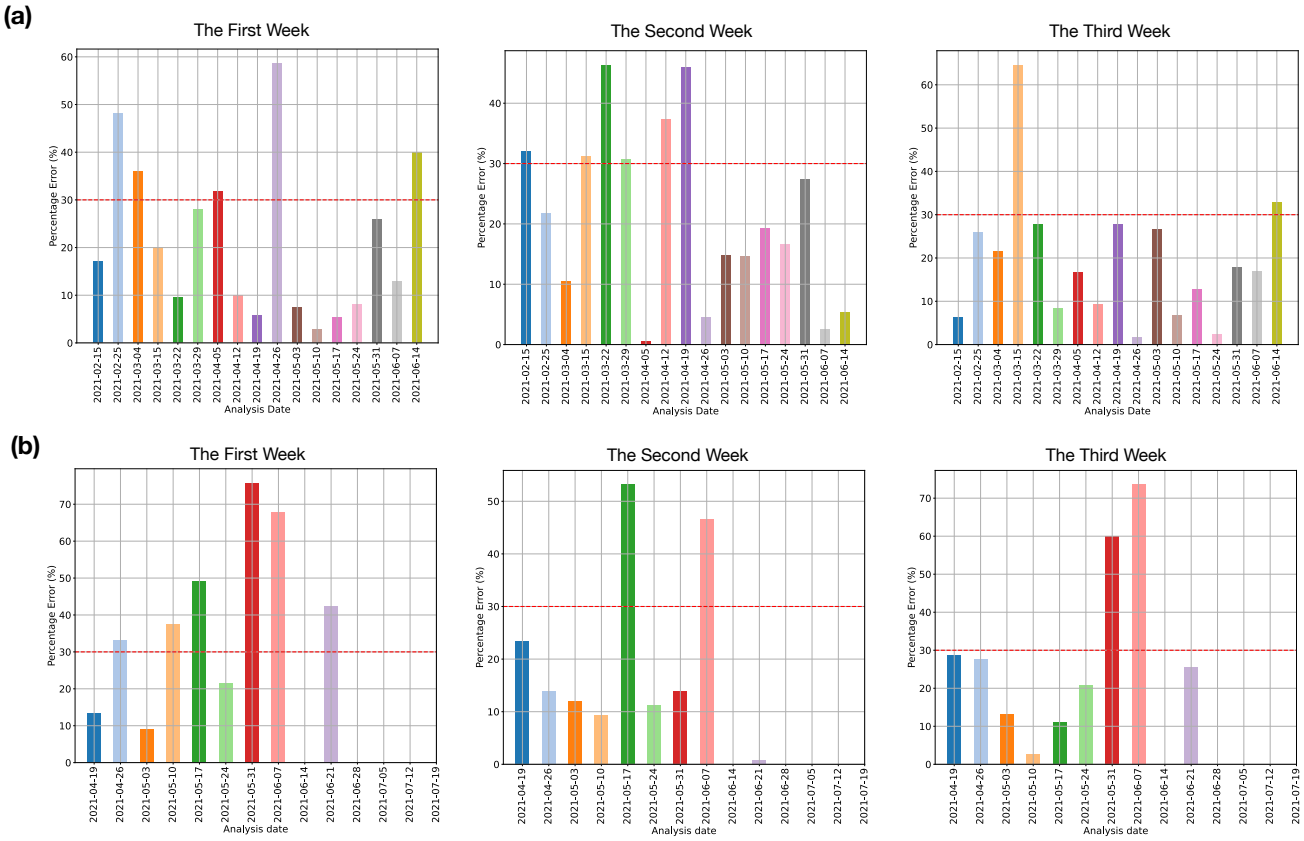

Figure 8. Error in prediction in terms of the percentage error score per analysis date. The duration of the forecast window of 21 days is partitioned into three parts, namely, the first week, the second week, and the third week. (a) Percentage error per partition and per analysis date, averaged across 7 partition days, for SU and (b) for SÄS. Percentage errors are not shown for those analysis dates for which the total number of admitted patients were fewer than 15 patients.

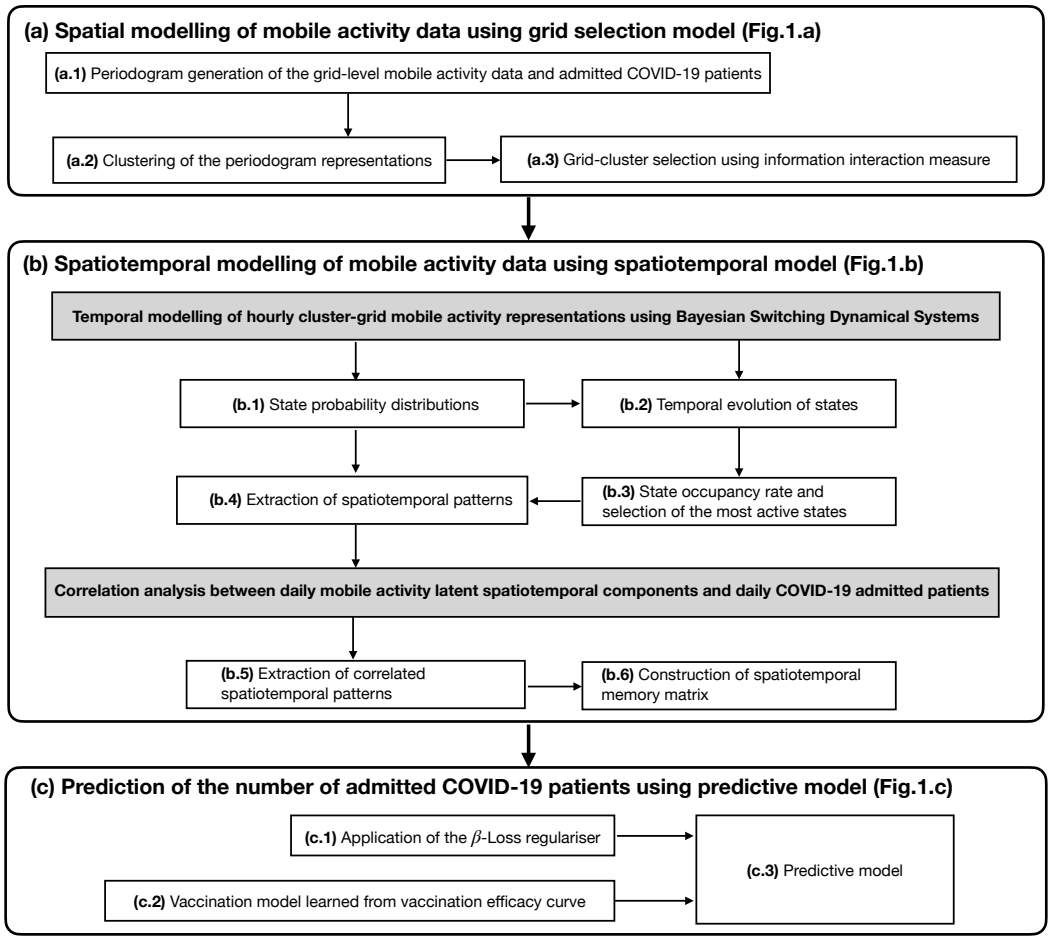

Figure 9. Overview of the main components in the forecast model pipeline and their functionalities.

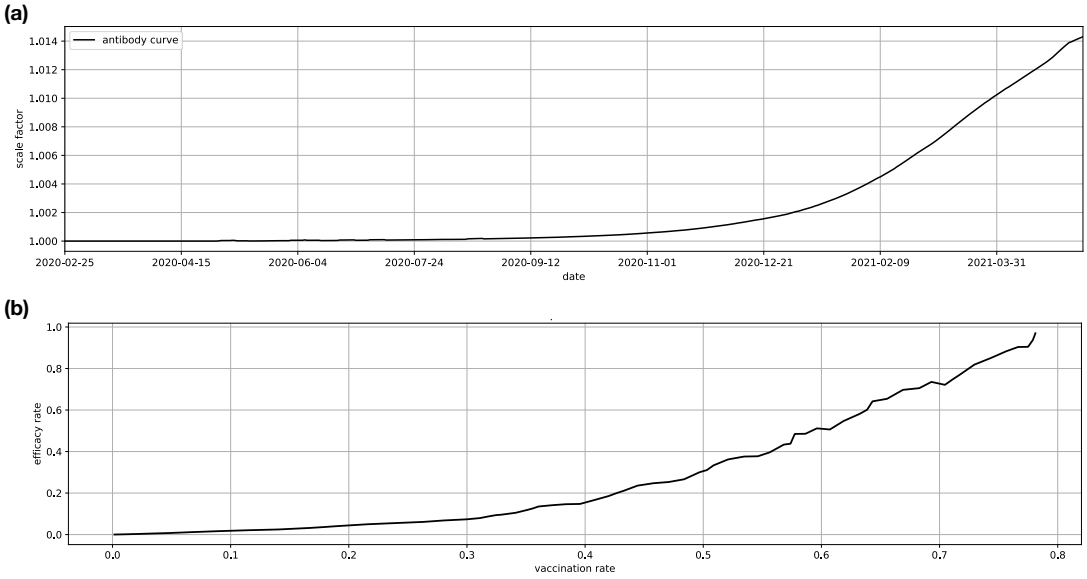

Figure 10. External factors. (a) Antibody curve which is computed from cumulative statistics on the antibody positive rates. (b) Vaccination efficacy curve obtained from Israel vaccination campaign data available at the time[24].

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementarymaterials.pdf