

# Comparing forecasts performances on different synthetic outbreaks

BECHADE Grégoire

4 juillet 2024

Résumé

## 1 Introduction

The recent outbreak of COVID-19 has highlighted the importance of accurate forecasting models, for policymakers and population alike. The number of hospitalized has shown to be a key indicator in the last pandemic, as maintaining the number of hospitalized individuals below the capacity of the healthcare system was a major concern. Predicting the trajectory of a pandemic is a complex task, considering the number of parameters that have an influence on the outbreak. In this study, we explore the efficiency of different models in predicting the number of hospitalized individuals.

We implemented 13 individual models and an ensemble model to forecast the number of hospitalizations 7 and 14 days ahead, given the data up to the current day. These models are divided into two categories : those trained only on hospitalization data and those incorporating additional time-series such as mobility data and the number of infected agents. The individual models are : variations of the SIRH (Susceptible, Infected, Recovered, Hospitalized) model, ARIMA, VAR, moving average, exponential regressions, bayesian and linear regressors.

A complex agent-based model was used to simulate pandemics of a wide diversity. By varying key parameters and using different mobility patterns, we created a diverse set of 324 synthetic pandemic scenarios (consisting of daily reports of key quantities). This diversity allows for a robust evaluation of the models across various conditions.

The models were trained and tested on many points of these pandemics, allowing for a consistent analysis of their performance. Additionally, we classified the points in the pandemic trajectories based on the reproductive number to understand model performance at different phases of the pandemic.

This paper aims to determine the most consistent and reliable models for forecasting hospitalizations during pandemics, at different phases of the outbreak. By analyzing the performance of various models across a wide range of synthetic pandemic scenarios, we aim to provide insights that can guide the selection and development of predictive models for real-world applications. Our findings highlight the strengths and weaknesses of different modeling approaches and underscore the value of ensemble models in achieving robust and accurate predictions.

The simulations that are described in this paper and lead to all the figures presented below are available on this [GitHub Repository](#).

## Results

### 1.1 The data

The models were trained and tested on various points of 324 different pandemics generated with Covasim [7], a complex agent-based model. To generate those pandemics, four parameters that maxi-

mized the diversity of the set of pandemics generated were selected (see 3.0.3). Those parameters were varied on the three values 0.5, 1 and 2. These values correspond to scaling factors of the default values of the model. For each pandemic, the transmission probabilities were time dependent and followed one of the four different mobility patterns (see Fig.12). Thanks to those 4 parameters, each varied on three different values and to the four different mobilities, we obtained 324 simulations of pandemics. We have decided to focus on the number of hospitalized individuals as the target variable. Indeed, it is a key variable to monitor, as the hospital occupancy corresponds to the capacity of the state to treat the patients.

## The models

In this paper, we implemented 13 individual models and an ensemble model. Each model was trained on the data from the beginning of a pandemic and made predictions 7 or 14 days ahead (on user's choice). They output a point prediction and a set of confidence intervals on the prediction. Models of two types were implemented. The models of the first type were only trained on the time series of the number of hospitalized of the past few days. The models of the second type were trained on the time series of the number of hospitalized, but also on the mobility data and the number of infected.

### 1.1.1 SIRH

The SIRH (Susceptible, Infected, Recovered, Hospitalized) (Fig.13) is a variation of the classical SIR model, with another compartment : the Hospitalized compartment. The parameters of this model correspond to the rates of evolution from a compartment to another one.  $\beta$  corresponds to the transmission rate,  $\gamma_i$  and  $\gamma_h$  correspond to the recovery rate from both Infected and Hospitalized compartments, and  $h$  corresponds to the hospitalization rate. The value of  $S, I, R$  and  $H$  are linked through a system of differential equations (1). As the curve of the pandemic of a SIRH is deterministic, it is possible to generate it from the values of the parameters  $\beta, \gamma_i, \gamma_h$  and  $h$ , and to compare it to the value of the number of hospitalized observed from the beginning of the pandemic. The optimal parameters are found through minimization of the least squares error between the SIRH curve and the real curve. During the predicting phase, a 7 (or 14) days SIRH is run, with initial value estimated from the data and the fit of the training phase. The confidence interval is estimated through linearization of the regression (see 3.1.7)

Variations of the SIRH model were implemented, with  $\gamma_i$  and  $\gamma_h$  kept constant.

A SIRH model of the second type was also constructed. Following the same idea as the first SIRH model described above, this model has a time varying transmission rate, which is a linear combination of the mobility :  $\beta_t = a \times m_t + b$ , as it is done in [4]. The parameters to optimize are now  $a, b, \gamma_i, \gamma_h$  and  $h$  and both curve of the number of hospitalized and the number of infected are fitted to the data.

### 1.1.2 ARIMA

The ARIMA model is used for time series forecasting. It is the sum of an AR and a MA model (3.1.2) It is trained by maximizing the likelihood of the observed data.

A VAR model, which is a multi-dimensional AR model was also implemented. It is also optimized through maximum likelihood.

### 1.1.3 Moving Average

The moving average model is a baseline model that is used as a reference. It returns the value of the average of the past seven days and a confidence interval based on the variance of the values of the past seven days.

#### 1.1.4 Exponential regression

The exponential regression is a model that fits an exponential function ( $x \rightarrow ae^{bx} + c$ ) to the values of the number of hospitalized. The values of  $a, b$  and  $c$  are found by minimizing the least square difference between the prediction and the real values. The confidence interval is estimated through linearization of the regression (see 3.1.7).

An exponential regression of the second type was also implemented. The number of hospitalized is then an exponential regression of both mobility data, number of infected and number of hospitalized shifted.

#### 1.1.5 Regressors

A Bayesian and a linear regressor were implemented. They use the last 20 data points to predict the next one (see 3.1.5).

### Evaluation of the models

The models described above were tested on 14 points of each of the 324 pandemics generated, when the pandemic was significant (i.e when more than 0.01% of the population was hospitalized, in our case when the value of the number of hospitalized was above 100). For each point, a 7 and 14 days-ahead prediction was made. Each prediction was evaluated thanks to WIS and RMSE (see 3.1.7), and then ranked according to the performances of the other models. The points of evaluation were also classified in one of the following categories : 'big increase', 'increase', 'stable', 'decrease' and 'big decrease', based on the value of the reproductive number at this time of the pandemic (see 3.2 for more details). It is then possible to get global information on the rankings of the models. For instance, if the loss and the reach of prediction is fixed, we can look at the distribution of rankings of all the models for a type of point (to see the best model for a given type of point), as in Fig.1. This distribution of rankings of the model can be summed up in one single value : the expected value of the rank of a model (see Fig.2), which enables to get the idea of the best model for this type of point, this loss, and this prediction range on a more compact figure.

This new number loses information (for instance on bimodal rankings distribution) but enables to highlight the expected performance of a model. For each loss and range of prediction, the expected rank of the models for each type of point can be visualized on the same figure, which enables to have a global look of the performances of the models.(Fig.3)

The other heatmaps corresponding to RMSE for 14 days ahead prediction and WIS for 7 and 14 days ahead predictions can be found in the appendix (4) Some general interpretations can be made on these results. We noticed that regressors performances drop from 7-days ahead to 14-days ahead predictions. The family of the SIRH model does slightly better for long terms predictions than for short term ones. The ARIMA performance drops from 7-days ahead to 14-days ahead predictions. The exponential models are very bad and almost always the last ones.

### 1.2 The Ensemble model

It has been shown in the literature ([2], [12], [6]) that ensemble models, which are models that take into account the predictions of many individual models and output their mean, median, or any other function that combines the predictions, such as stacking (as it is done in [12]) or linear opinion pool (see [6]). We decided to implement our own ensemble models. They consist in a linear combination of the predictions of the 13 individual models (without exponential models) for 7 and 14 days-ahead predictions. The performances of these two models are shown in the two figures below (Fig.4 and Fig.5)

The distribution of the ranks of the ensemble model is almost always on the left side of the x-axis. It is a very consistent model, and almost always in the top models. The heatmaps of expected ranks on all type of points enables to see how consistent is the ensemble model (Fig.7 and Fig.6) compared to the other models.

Distribution of the ranks of each model for big increase points for RMSE and 7 days ahead prediction.

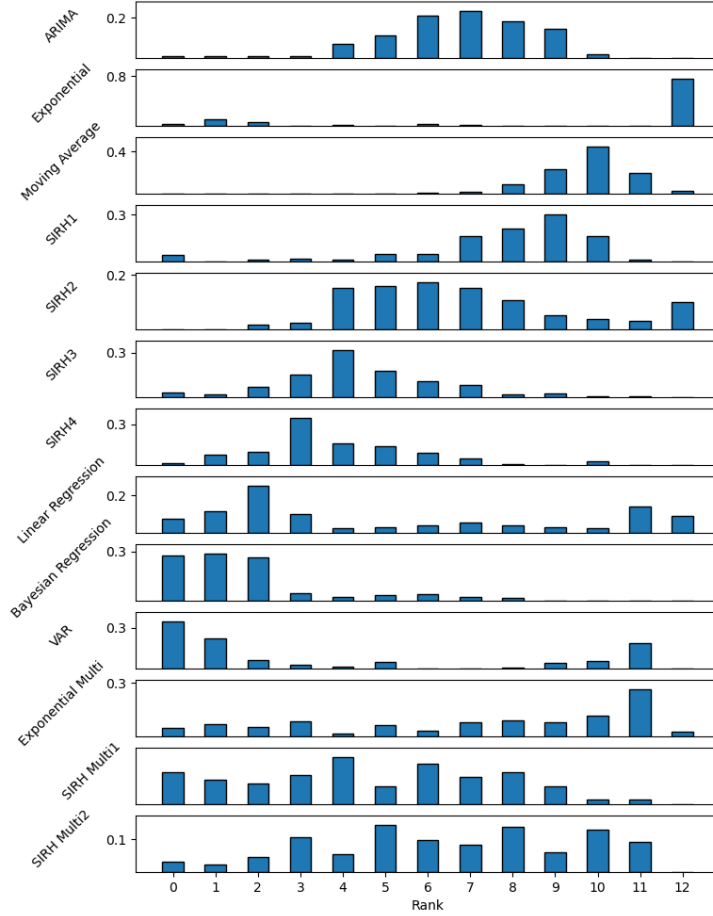


FIGURE 1 – Distribution of rankings of the models for big increase points for 7-days ahead predictions and RMSE loss

From these figures, we can see that the ensemble model is rarely the best, but never the last model, which is a result already observed in [2]. Indeed, for a loss, a type of point and a prediction range, one can look at the best models in terms of expected rank. For instance, on the Figure.2, one can see that the top 3 models are Bayesian Regression, VAR and SIRH4. On the test set, the Ensemble model is the only model that appears in all top-6 models for all types of points, all losses and all prediction ranges. Its consistency allows for accurate predictions and helps avoiding outlier predicted values that sometimes appear on the other good models (such as VAR and ARIMA).

### 1.3 Test on real data

We also tested the models on real data from the Covid-19 pandemic in Sweden. The data of the number of hospitalized individuals was taken from the website [Our World in data](#), the data of the number of infected was taken from the website [Worldometers](#), and the mobility used was the mobility from the Västtrafik reports ([4]). We acknowledge that the mobility reports from Västtrafik only concern the region of Västra Götaland, and that the mobility of the whole country would have been more

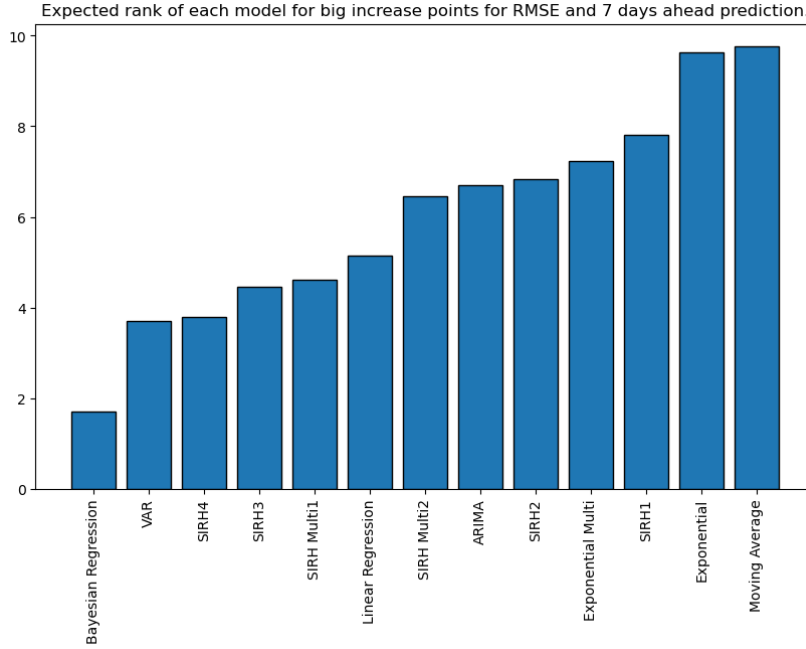


FIGURE 2 – Expected rank of the models for big increase points for 7-days ahead predictions and RMSE loss

relevant, but it is a good approximation, as the mitigation measures were at the state level in Sweden. The models were tested each 20 days from the 2nd of March to the 7th of December 2020. For each of these points, the models were trained on the data from the beginning of the pandemic to the day before the prediction, and outputted a 7-days ahead and a 14-days ahead prediction. As we wanted to compare the models with the ensemble model, we decided to assess their performances with the RMSE and not the WIS. For each point prediction, the models were ranked according to the RMSE between the prediction and the real value. The rankings are shown on fig 8. The predictions of the models are plotted with the real data on fig 9. We noticed similar results to what we observed on the synthetized set for Arima, VAR, Bayesian and Linear regressors. The performance of the ensemble model is disappointing during the period between the 20th of July and the 8th of October. This "flat" period was probably not present in the tes on which the ensemble model was optimized, as the real-world data is probably much more complex than the data generated with Covasim. Moreover, the Ensemble model puts lots of weight on the VAR model (see fig 14), which might be very sensitive to the errors in the data, as it takes into account the number of infected and the mobility. However, event if the ensemble model is not well-ranked, it lands not far from the real value, the other models being a little closer than him.

## 2 Discussion

### 2.1 Summary of Key Findings

This study enabled us to assess the consistency of the different models that were implemented. Indeed, comparing the models on a set of diverse pandemics highlighted the most consistent with the different pandemics. Moreover, the classification of the models with the type of points leads to the determination of the phase of the pandemics in which each model is the best, and on the other hand to assess which model is the best for a  $R_{eff}$  given. This result is promising, as the reproduction number

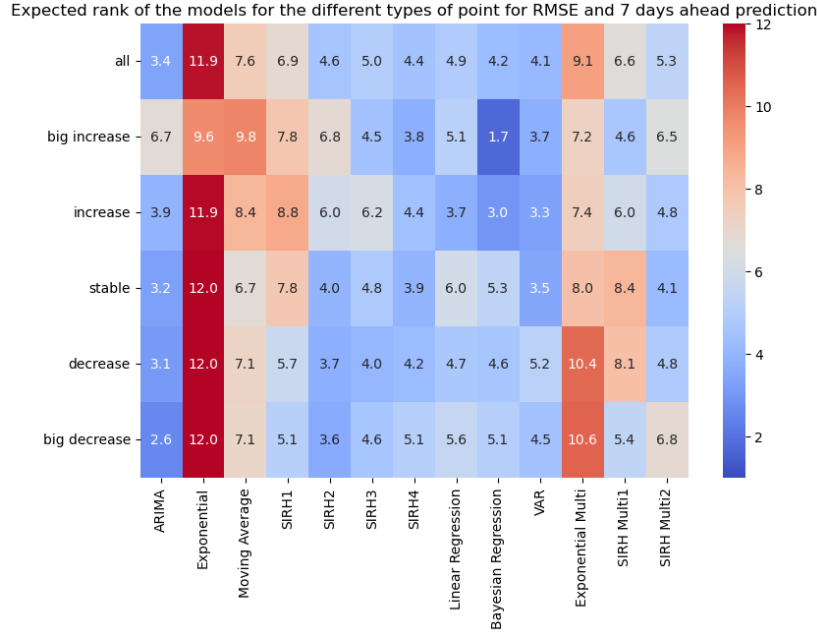


FIGURE 3 – Expected rank of the models for each type of point for 7-days ahead predictions and RMSE loss

is a quantity often estimated by the stakeholders during a pandemic. The interpretation based on the heatmaps Fig.15, Fig.3 and Fig.17 and Fig.16 shows the following results :

- Arima and VAR are very performant models in terms of expected ranking. They are the two best models for 7-days ahead predictions and in the top 4 models for the 14-days ahead predictions.
- Adding information with the mobility and the number of infected is not beneficial for the models. We observe almost always a poorer performance of the models of the second type compared to their equivalent model of the first type (VAR / Arima and SIRH Multi / SIRH ). Only the exponential multi outperforms the mere exponential model, but as they both perform pretty bad, this result is not interesting.
- The autoregressive models (Bayesian and Linear Regression, ARIMA and VAR) perform best for short term predictions rather than long term ones. On the other hand, the SIRH models, which have a physical meaning and try to fit the data to an understandable model perform better for long term predictions. This is also observed in the coefficients of the ensemble model (see Fig. 14). Indeed, the ensemble model puts more weight on the VAR and Arima models for short term predictions and more on the SIRH models for long terms predictions.
- The difference of ranking between the models is more pronounced when we rank them with RMSE rather than WIS. This might be due to the fact that the confidence intervals are not well calibrated and that the real value always lands out of it.
- The Moving Average model is not the worst model and is often ranked between 6 and 8. This might be due to the fact that as the curve of the number of hospitalized is continuous, the mean never diverges too much from the real value, while other models might misinterpret the trend of the curve and land far away from the real value.
- The Exponential models are almost always the last models, except for big increase points that are almost always at the very beginning of the pandemic. The fact that the beginning of a pandemic looks like an exponential curve is also shown with the SIRH model, when the number of infected is very small.

Distribution of the ranks of the Ensemble model for different type of points for 7 days ahead prediction.

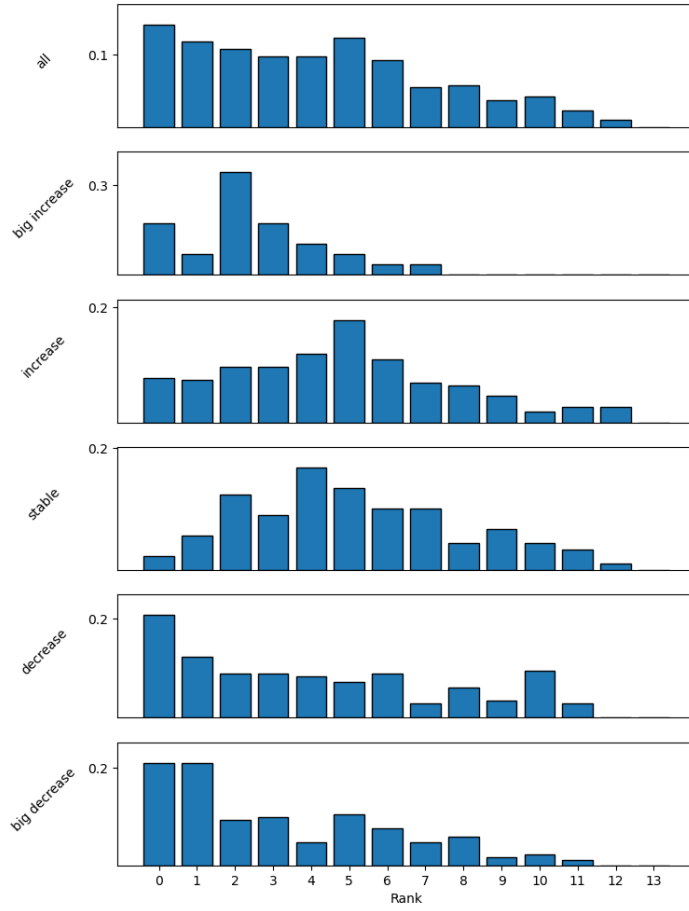


FIGURE 4 – Expected rank of the ensemble model for each type of point for 7-days ahead predictions and RMSE loss

Distribution of the ranks of the Ensemble model for different type of points for 14 days ahead prediction.

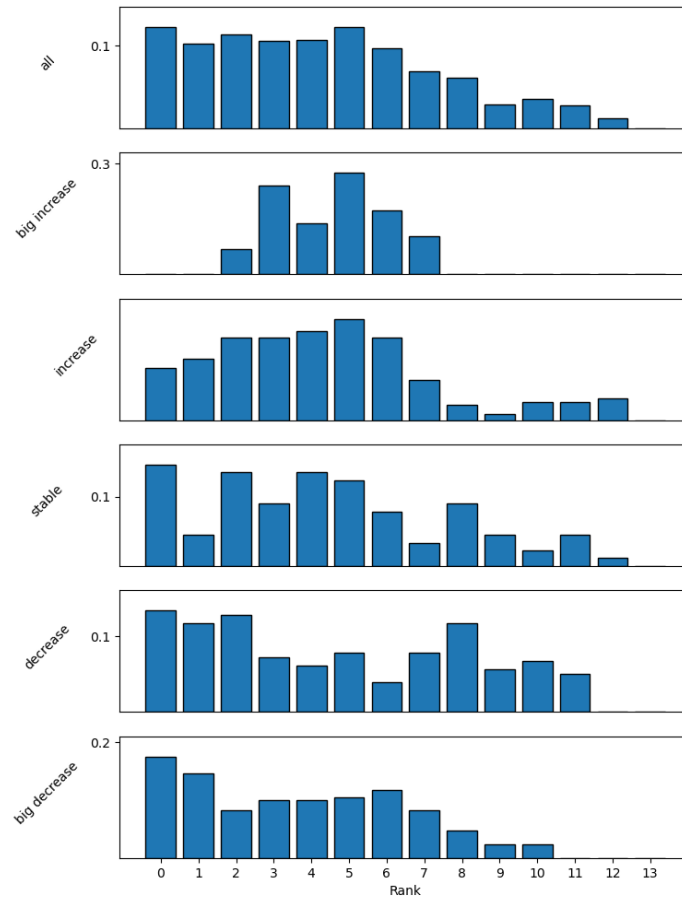


FIGURE 5 – Expected rank of the ensemble model for each type of point for 14-days ahead predictions and RMSE loss



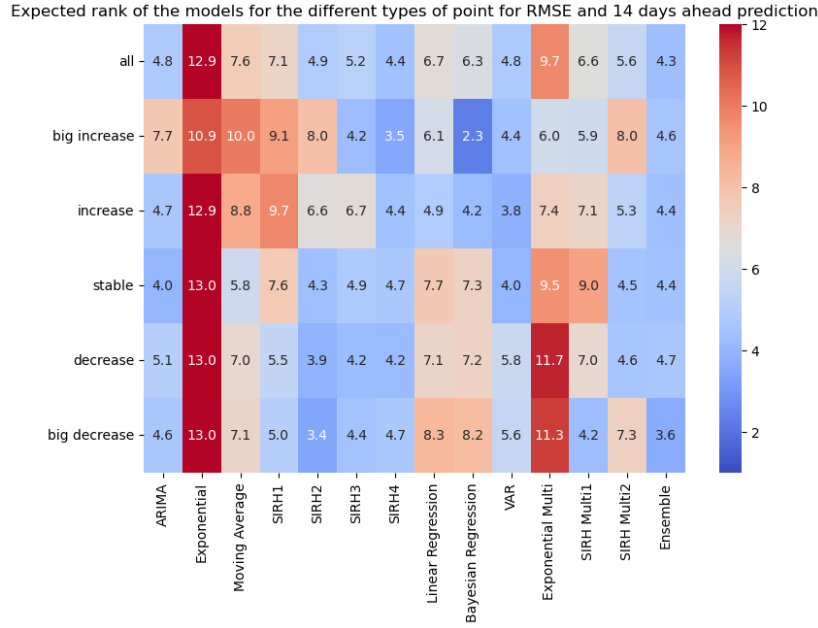


FIGURE 6 – Expected rank of all the models for each type of point for 14-days ahead predictions and RMSE loss on the test set

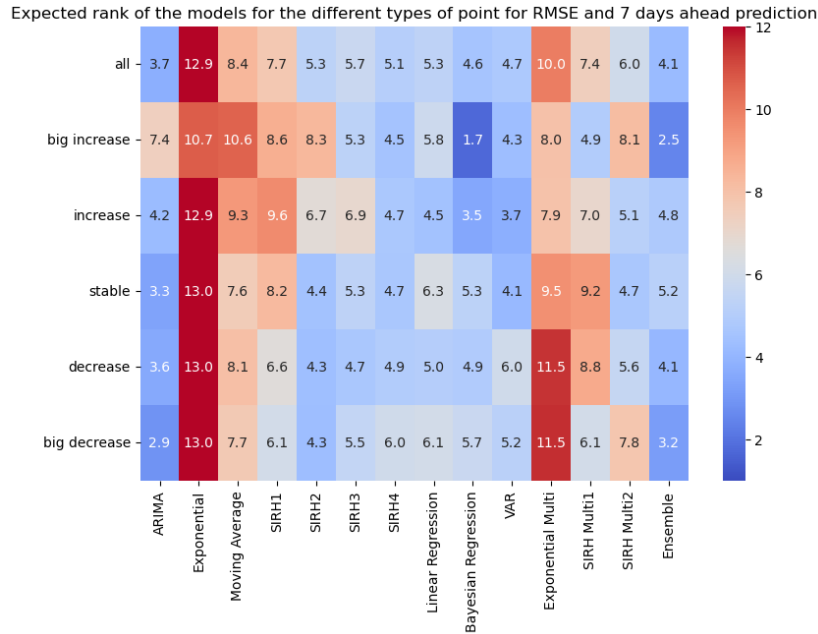
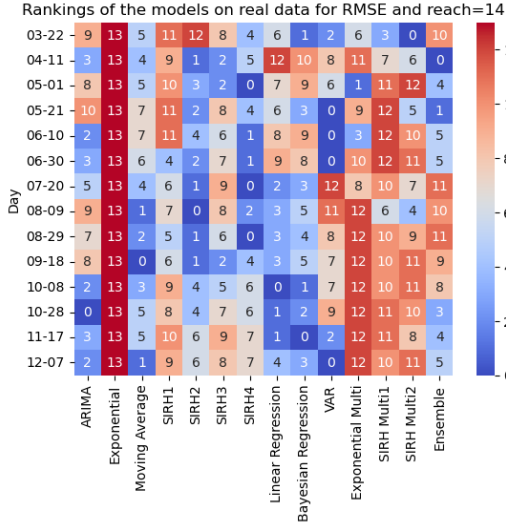
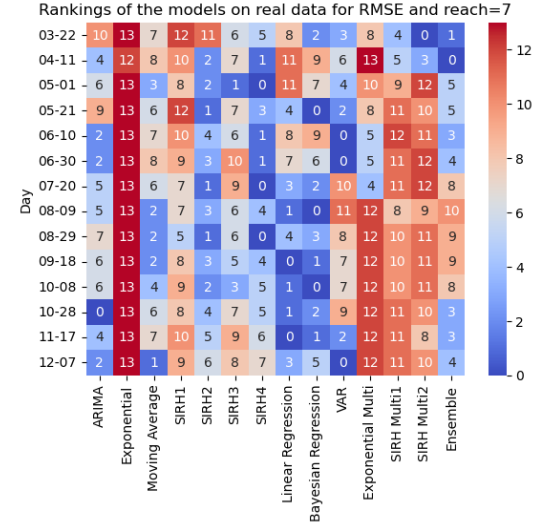


FIGURE 7 – Expected rank of all the models for each type of point for 7-days ahead predictions and RMSE loss on the test set

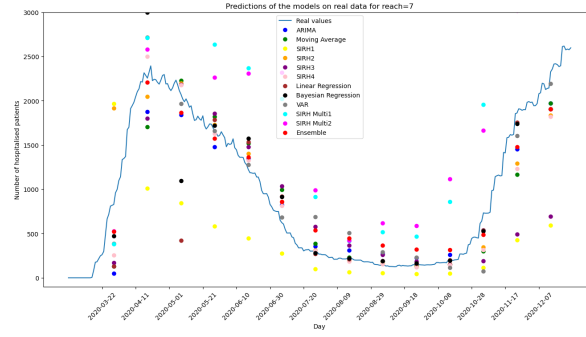


(a) Rankings for 7-days ahead predictions

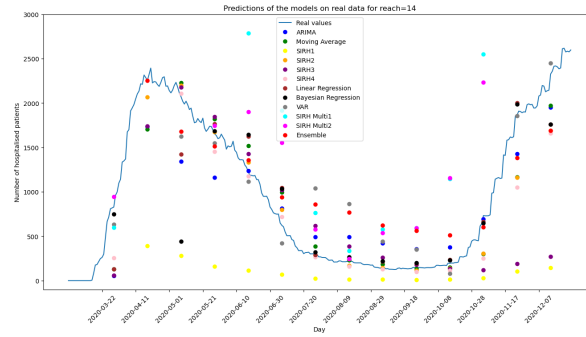


(b) Rankings for 14-days ahead predictions

FIGURE 8 – Rankings of the models during the Covid-19 pandemic in Sweden



(a) Predictions 7-days ahead



(b) Predictions 14-days ahead

FIGURE 9 – Predictions of the models during the Covid-19 pandemic in Sweden. The y-axis was cut at 3000.

— The Ensemble model performs quite well. We noticed that it is very rarely the best model for a type of point, while other models almost always have some point on which they produce the

best prediction. On the other hand, it is never the last, and has a very consistent ranking : it is the only model that appears in all top-6 models for all types of points, all losses and all prediction ranges.

## 2.2 Comparison with Previous Research

The very good performance of both VAR and ARIMA models is consistent with previous research. [8] and [14] both assess the performance of the ARIMA model in term of regional or national prediction accuracy. The interpretation of the coefficients of the ARIMA model enables to have information on the inherent parameters of the pandemic, or to discover links between the different regions of a country. The consistency of the ensemble model has been shown in many previous studies (see [2], [17] and [12]), who also observe that their ensemble model outperformed any individual model. They also observe the fact that the ensemble model is very consistent : rarely the best, but never the last, allowing to produce robust forecasts. In [11], the authors compared several model to understand the key features that enabled a model to produce accurate long term predictions. These features are : "*(1) capturing the physics of transmission (instead of using black-box models) ; (2) projecting human behavioral reactions to an evolving pandemic ; and (3) resetting state variables to account for randomness not captured in the model before starting projection.*" These findings are consistent with the results we have observed. Indeed, we have seen that the SEIR model (capturing the physics of transmission and resetting state ) performances increase from 7-days ahead to 14-days ahead predictions, while black-box models (such as Arima or VAR), are less performant for long terms prediction. This has to be qualified, as our "long term predictions" were a 14-days ahead prediction, while the study [11] gets interested to predictions up to 14-weeks reach.

The methodology of point classification has not be found in any other paper of the literature. Some studies assess the performance of the models during different times of a pandemic (see [6]), but this systematic point classification based on the reproduction number was not found.

## 2.3 Practical Implications

This study might be useful when facing a new pandemic. As the value of the  $R_{eff}$  is often estimated by the researchers during an outbreak, it could be used to decide which type of model to use during the current pandemic.

## 2.4 Limitations

This study did not take into account uncertainty on the data or imperfect data. Indeed, the Covasim library provided a daily data of high quality and without any noise. Real-life data is often full of mistakes, of missing values or of time-lag in the reports ([5]). It would be interesting to measure the robustness of the models with respect to these imperfections to assess if the models can still make good predictions. Moreover, the models of the second type used the number of infected to predict the number of hospitalized. This quantity is not directly accessible and is often estimated with tests, which could lead to huge variations in the predicted values. Moreover, the models were only tested on outbreaks generated with Covasim, which might lead to a bias in the results, despite the effort to create a diverse set of pandemics.

## Conclusion

This study should be seen as a try to reproduce previous results on the consistency of autoregressive and ensemble models, but also as a first attempt to assess the robustness of some models with the evaluation on a huge set of diverse pandemics. The methodology for point classification is also a new way to compare model performance, that might be useful during real outbreaks, as the reproduction number a popular quantity, which is often estimated. The results of the autoregressive and ensemble

models are promising, as many aspects of previous research have been highlighted once again. Future research on the consistency of models with error in the data might complete this one, by assessing that the observed robustness of the models is not only due to the quality of the data.

### 3 Methods

#### The data

As the goal of this study is to compare the accuracy of forecast in the context of many different pandemics, a large set of synthetic pandemics was generated, with a particular attention on the diversity of these pandemics.

##### 3.0.1 Covasim

To generate the pandemics, Covasim [7], a python library that can simulate the evolution of a pandemic, was used. Covasim is an agent-based model that can simulate many different pandemics and has a high diversity of outputs. This model takes as an input many parameters such as the population type, the population size, the age distribution etc, and outputs a complete description of the pandemic, with real-time values of each relevant piece of information, such as the number of severe, of asymptomatic etc, but also physical data such as the value of the reproduction number. Covasim enables to generate a huge diversity of pandemics, thanks to the plurality of parameters that can be given as input to the model, but also with interventions that can be planned by the users. These interventions can help to assess the impact of a vaccination campaign, with changes in the probability transmission. All the simulations in this paper are made on a fixed population of  $10^6$  agents.

##### 3.0.2 First pandemic

For the implementation and the first tests of the models, a single pandemic was generated.

TABLE 1 – Table of the parameters used to simulate the first pandemic.

| Parameter          | value          |
|--------------------|----------------|
| start day 1        | 2020-03-02     |
| end day            | 2021-01-01     |
| Population size    | 1000000        |
| Interventions      | interventions1 |
| population type    | hybrid         |
| $\beta$ initial    | 0.015          |
| location           | Sweden         |
| n infected initial | 100            |

The parameters used to generate this pandemic are described in the table 1 The parameters that are not specified are the default parameters of the Covasim library.

The interventions (interventions1) were based on mobility reports from Västtrafik (Fig.10), the public transport company of the city of Gothenburg. They were reported during the Covid-19 pandemic and have been retrieved from [4]. These interventions correspond to 53 relative weekly variations of the mobility, with a reference value of 1 for the first week of the report, which correspond to the 9-th week of 2020. As the interventions refer to relative changes in mobility, they were used as relatives changes in probabilities of transmission.

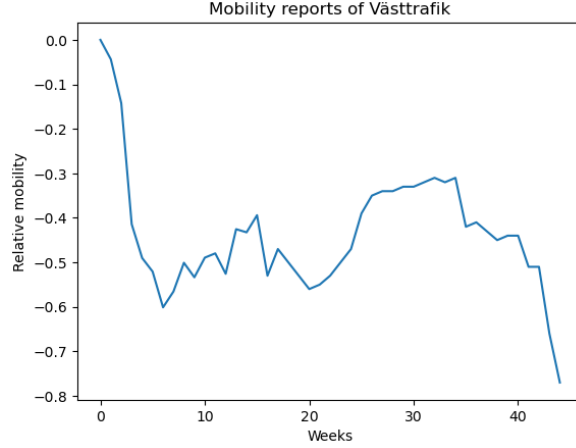


FIGURE 10 – Mobility reports from Västtrafik.

### 3.0.3 Generating diverse pandemics

In order to evaluate the performances of our models on a wide range of pandemics, a training set of pandemics was generated. A huge diversity of pandemics is needed to determine the most accurate forecasting model. It is therefore relevant to identify the key parameters that generate this diversity. As Covasim has a large set of inputs parameters, a first subset of key parameters was identified : the *spread* parameters and the *severity* parameters, which are parameters related to disease transmission and severity. The *severity* parameters are the 4 parameters that correspond to the probability for an individual to move from one disease compartment to another. The *spread* parameters are 9 parameters that represent the distribution of probability of the time spend by an agent in a compartment (such as infected, critical...) once he entered it. This distribution is a log-normal distribution, but the *spread* parameters correspond to the mean of this log-normal distribution. All the parameters have a default value of 1, which correspond to keeping the reference value. This set of 13 parameters is then noted as  $S$ . In order to test the forecasting models on a large, but still manageable set of different pandemics we decided to vary four parameters and that these should each take three different values in  $[0.5, 1, 2]$ , leading to a set of 81 pandemics. To select the 4 parameters that generated the most diversity out of the 13 possible, different diversity metrics were constructed.

Let  $Y_1, Y_2 \in \mathbb{R}^n$  be two time series of  $n$  days representing the number of hospitalized in two pandemics.

Let :

$$\mathcal{L}_1(Y_1, Y_2) = \|Y_1 - Y_2\|_{L_1}$$

$$\mathcal{L}_2(Y_1, Y_2) = \left\| \left( \frac{\max(Y_1)}{\max(Y_2)}; \frac{\max(Y_1')}{\max(Y_2')}; \frac{\max(Y_1'')}{\max(Y_2'')} \right), \|\tilde{Y}_1 - \tilde{Y}_2\|_{L_1}, \|\tilde{Y}_1' - \tilde{Y}_2'\|_{L_1}, \|\tilde{Y}_1'' - \tilde{Y}_2''\|_{L_1} \right\|_{L_2}$$

with  $Y'$  and  $Y''$  the first and second numerical derivatives of  $Y$

$$\mathcal{L}_3 = \mathcal{W}(\tilde{Y}_1 - \tilde{Y}_2), \text{ with } \mathcal{W} \text{ the Wasserstein distance.}$$

$$\mathcal{L}_4(Y_1, Y_2) = \left\| \left( \frac{\max(Y_1)}{\max(Y_2)}; \frac{\max(Y_1')}{\max(Y_2')}; \frac{\max(Y_1'')}{\max(Y_2'')} \right), \mathcal{W}(\tilde{Y}_1 - \tilde{Y}_2), \mathcal{W}(\tilde{Y}_1' - \tilde{Y}_2'), \mathcal{W}(\tilde{Y}_1'' - \tilde{Y}_2'') \right\|_{L_2}$$

with  $Y'$  and  $Y''$  the first and second numerical derivatives of  $Y$

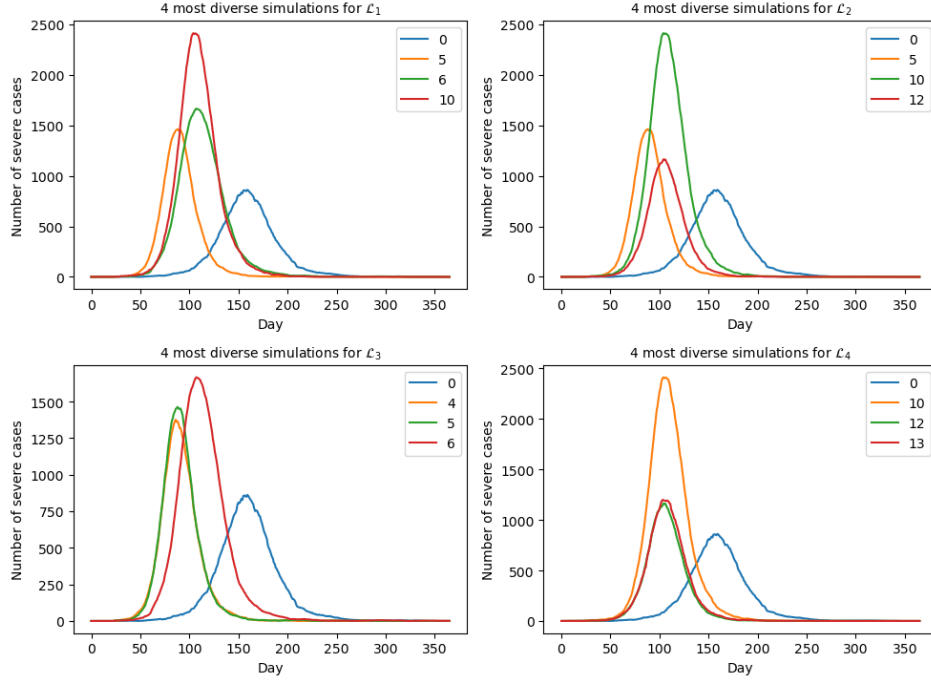


FIGURE 11 – 4 most diverse pandemics according to each norm.

It can be noted that  $\mathcal{L}_2$  is similar to the Sobolev norm  $\|\tilde{Y}_1 - \tilde{Y}_2\|_{W^{2,1}}$  with squared terms and with additional terms taking into account the amplitude.

To determine which measure to use, we generated 14 pandemics : one reference pandemic with all parameter values equal to the default values and the remaining pandemics having one parameter value doubled.

For each norm  $\mathcal{L}_k$ , we determined  $s$ , the subset of 4 pandemics that maximized the following quantity :

$$\mathcal{L}_k(S) = \sum_{i,j \in s, i \neq j} \mathcal{L}_k(Y_i, Y_j)$$

The 4 most diverse pandemics (in the subset of 14 pandemics generated) according to each norm are shown in the Fig.11. We decided according to this figure, that  $\mathcal{L}_2$  norm was the most relevant to determine the diversity of the pandemics. But, keeping the parameters from this evaluation ([0, 5, 10, 12]) would not be accurate, as the parameters were changed independently, and the diversity did not take into account the correlation between some of them.

Finding the parameters that maximise the  $\mathcal{L}_2$  diversity is equivalent to solving the following problem :

$s_{opt} = \operatorname{argmax}_{s \subset S, |s|=4} \mathcal{L}(s)$ , with  $\mathcal{L}(s) = \sum_{p_1, p_2 \in \mathcal{P}_g(s)} \mathcal{L}_2(p_1, p_2)$ , and  $\mathcal{P}_g(s)$  the set of the 81 pandemics generated with the 4 parameters of  $s$ .

However, generating a pandemic with **Covasim** is time consuming, and it is not possible to compute the diversity of each set of 4 parameters  $s$  included in  $S$  (the set of all 13 parameters).

An MCMC algorithm [3] was therefore implemented, to perform a clever grid search on the different subsets  $s \subset S$  of parameters. The MCMC algorithm is a method which is used to sample from a distribution that cannot be directly sampled. The main idea is to construct a Markov Chain whose stationary distribution is the objective distribution.

Let  $\mathcal{S} = \{s \subset S; |s| = 4\}$  be the support of the target distribution, which is, in our case, the set of all the 715 combinations of the 4 parameters among the 13 different possible, and let  $\pi$  be the target distribution on  $\mathcal{S}$ .  $\forall s \in \mathcal{S}, \pi(s) = \frac{\mathcal{L}(s)}{\sum_{s \in \mathcal{S}} \mathcal{L}(s)}$ .  $\pi$  is not directly computable as it is too time consuming to compute the denominator.

For each  $s = [a, b, c, d] \in \mathcal{S}$ , let  $ne(s)$  be the set of the neighbours of  $s$ , i.e the set of all the elements of  $\mathcal{S}$  who have only one parameter different from  $s$ . For instance,  $[0, 3, 9, 12] \in ne([0, 3, 10, 12])$ , but  $[0, 3, 9, 12] \notin ne([0, 3, 8, 10])$ .

Let  $U_n$  be a sequence of independent uniform random variables on  $[0, 1]$  and  $\forall s \in \mathcal{S}$ , let  $U_n^{ne(s)}$  be a sequence of independent uniform random variables on  $ne(s)$ . Let  $s_0 \in \mathcal{S}$  and let  $S_n$  be the random sequence defined as follow :

$$\begin{cases} S_0 = s_0 \\ \forall n \in \mathbb{N}, \alpha_n = \frac{\mathcal{L}(U_n^{ne(S_n)})}{\mathcal{L}(S_n)} \\ \forall n \in \mathbb{N}, S_{n+1} = U_n^{ne(S_n)} \mathbb{1}_{\{U_n < \alpha_n\}} + S_n \mathbb{1}_{\{U_n > \alpha_n\}} \end{cases}$$

This formula means that at each iteration, a neighbour of  $S_n$  is uniformly selected among all the neighbours of  $S_n$  (it is  $U_n^{ne(S_n)}$ ). The Markov Chain moves to this neighbour if the value of  $\mathcal{L}(U_n^{ne(S_n)})$  is higher than the value of the function  $\mathcal{L}(S_n)$  at the current state. If the new value of  $\mathcal{L}$  is smaller, the Markov Chain moves with a probability that is equal to the ratio of the two values. This algorithm enables to perform a clever grid search, as the Markov Chain tends to increase the diversity but is authorized to sometimes diminishes it to avoid to stay stucked in a local maximum, as a mere gradient descent would do. As  $S_{n+1}$  is a function of  $S_n$  and of other independent random variables, the sequence  $S_n$  is a homogenous Markov Chain.

The transition matrix of this Markov Chain is the following :

$$K(s, s') : \begin{cases} 0 \text{ if } s' \notin ne(s) \text{ and } s' \neq s \\ \frac{1}{Card(ne(s))} = \frac{1}{36} \text{ if } s' \in ne(s) \text{ and } \frac{\mathcal{L}(s')}{\mathcal{L}(s)} > 1 \text{ and } s' \neq s \\ \frac{1}{36} \times \frac{\mathcal{L}(s')}{\mathcal{L}(s)} \text{ if } s' \in ne(s) \text{ and } \frac{\mathcal{L}(s')}{\mathcal{L}(s)} \leq 1 \text{ and } s' \neq s \\ 1 - \sum_{s' \in \mathcal{S}, s' \neq s} K(s, s') \text{ if } s' = s \end{cases}$$

To prove that the stationary distribution of this Markov chain is  $\pi$ , we will show that it is an irreducible and aperiodic Markov chain and that  $\pi$  is reversible for  $K$ .

Let  $(s, s') \in \mathcal{S}^2$ . Let us suppose that  $s' \neq s$ , that  $s' \in ne(s)$ , and that  $\mathcal{L}(s) < \mathcal{L}(s')$  (the other case is symmetric).

$$\begin{aligned}
\pi(s)K(s, s') &= \frac{\mathcal{L}(s)}{\sum_{s \in \mathcal{S}} \mathcal{L}(s)} \times \frac{1}{36} \quad \text{as } \mathcal{L}(s) < \mathcal{L}(s') \\
&= \frac{\mathcal{L}(s)}{\sum_{s \in \mathcal{S}} \mathcal{L}(s)} \times \frac{1}{36} \times \frac{\mathcal{L}(s')}{\mathcal{L}(s')} \\
&= \frac{\mathcal{L}(s')}{\sum_{s \in \mathcal{S}} \mathcal{L}(s)} \times \frac{1}{36} \times \frac{\mathcal{L}(s)}{\mathcal{L}(s')} \\
&= \pi(s')K(s', s)
\end{aligned}$$

Indeed, each subset  $s$  has 36 neighbourh, as there are 13 parameters and one can replace each parameter of  $s$  by any of the 9 others.

If  $s' \notin ne(s)$ , then  $s \notin ne(s')$  and  $K(s, s') = 0$ . We directly have  $\pi(s)K(s, s') = 0 = \pi(s')K(s', s)$ .

Thus,  $\pi$  is **reversible** for  $K$ .

Let  $(s, s') \in \mathcal{S}^2$ . Let us note  $(a, b, c, d)$  and  $(a', b', c', d')$  the elements of  $s$  and  $s'$ . We denote :

$$\begin{aligned}
s_1 &= [a', b, c, d] \\
s_2 &= [a', b', c, d] \\
s_3 &= [a', b', c', d]
\end{aligned}$$

$$\begin{aligned}
\mathbb{P}(S_{n+4} = s' | S_n = s) &\geq \mathbb{P}(S_{n+4} = s' | S_{n+3} = s_3 \cap S_{n+2} = s_2 \cap S_{n+1} = s_1 | S_n = s) \\
&\geq \mathbb{P}(S_{n+4} = s' | S_{n+3} = s_3 \cap S_{n+2} = s_2 \cap S_{n+1} = s_1 \cap S_n = s) \\
&\quad \times \mathbb{P}(S_{n+3} = s_3 \cap S_{n+2} = s_2 \cap S_{n+1} = s_1 | S_n = s) \quad (\text{Baye's Formula}) \\
&\geq \mathbb{P}(S_{n+4} = s' | S_{n+3} = s_3) \\
&\quad \times \mathbb{P}(S_{n+3} = s_3 \cap S_{n+2} = s_2 \cap S_{n+1} = s_1 | S_n = s) \quad (\text{by Markov's property}) \\
&\vdots \\
&\geq \mathbb{P}(S_{n+4} = s' | S_{n+3} = s_3) \times \mathbb{P}(S_{n+3} = s_3 | S_{n+2} = s_2) \times \mathbb{P}(S_{n+2} = s_2 | S_{n+1} = s_1) \\
&\quad \times \mathbb{P}(S_{n+1} = s_1 | S_n = s) \quad (\text{by Markov's property}) \\
&\geq \left(\frac{1}{36}\right)^4 \times \min(1, \frac{\mathcal{L}(s')}{\mathcal{L}(s)}) \times \min(1, \frac{\mathcal{L}(s_3)}{\mathcal{L}(s_2)}) \times \min(1, \frac{\mathcal{L}(s_2)}{\mathcal{L}(s_1)}) \times \min(1, \frac{\mathcal{L}(s_1)}{\mathcal{L}(s)}) \\
&> 0
\end{aligned}$$

Thus,  $S_n$  is **irreducible**.

A Markov chain of transition matrix  $P$  on the support  $\mathcal{S}$  is said to be aperiodic if :  
 $\forall s \in \mathcal{S}, \forall s' \in \mathcal{S}, \exists N \in \mathbb{N}, \text{ s.t } \forall n > N, P(s, s')^n > 0$  [1]

First, note that  $\forall s \in \mathcal{S}$ ,  $s$  is a local minimum (i.e if  $\forall s' \in ne(s), \mathcal{L}(s') > \mathcal{L}(s)$ ) if and only if  $K(s, s) = 0$

Thus, if  $s$  is not a local minimum, then  $K(s, s) > 0$ . Moreover,  $\forall s \in \mathcal{S}, \forall s' \in ne(s)$ , if  $s \neq s'$ , then  $K(s, s') \neq 0$

Let  $(s, s') \in \mathcal{S}^2$ .

- If  $s'$  is not a local minimum,  
 $\forall n > 4, \mathbb{P}(S_n = s' | S_0 = s) \geq \mathbb{P}(S_4 = s' | S_0 = s) \times K(s', s')^{n-4} > 0$
- If  $s'$  is a local minimum,  $\forall s^* \in ne(s')$ ,  $s^*$  is not a local minimum and  $K(s^*, s^*) \neq 0$ .  
 $\forall n > 5, \mathbb{P}(S_n = s' | S_0 = s) \geq \mathbb{P}(S_3 = s^* | S_0 = s) \times K(s^*, s^*)^{n-4} \times K(s^*, s') > 0$



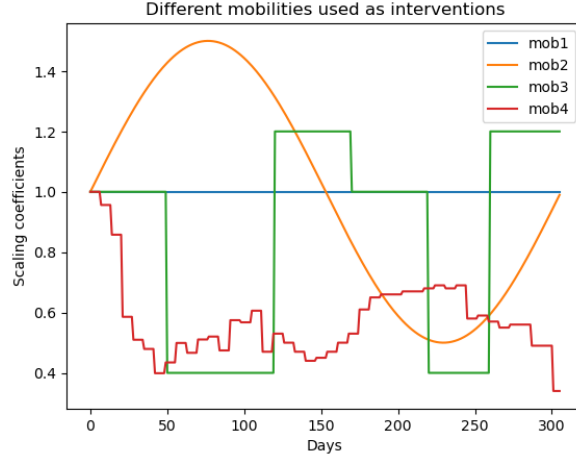


FIGURE 12 – Mobility reports.

Thus  $S_n$  is an **aperiodic** Markov Chain.

Finally, according to the **Theorem 5.5** from [1], as  $S_n$  is irreducible and aperiodic, as  $\pi$  is the stationary distribution, and as  $\mathcal{S}$  is countable,  $S_n$  converges in distribution to  $\pi$ .

The most probable set that will be sampled by  $S_n$  is the one that maximises the diversity. We implemented this MCMC algorithm to maximise  $\mathcal{L}_2$  on  $\mathcal{S}$ . The first value of the sequence was a clever starting point : the four parameters that maximized  $\mathcal{L}_2$  on the 14 pandemics generated earlier (Fig.11). After 200 iterations, the set of parameters that maximised the diversity was found to be [2, 4, 9, 10], which correspond to the parameters [sym2sev, asym2rec, rel\_symp\_prob, rel\_severe\_prob]. They correspond respectively to the mean of the log-normal distribution representing the time spent in the compartment 'symptomatic' before moving into the compartment 'severe', the mean of the log-normal distribution representing the time spent in the compartment 'asymptomatic' before moving into the compartment 'recovered', to the scale factor for proportion of symptomatic cases and to the scale factor for proportion of symptomatic cases that become severe. (see [7]). The  $\mathcal{L}_2$ -diversity increased from 62 353 to 93 553.

To create the most diverse set possible, we also created 4 different time-dependent mobilities (Fig.12), corresponding to constant mobility, annual variations, lockdown scenario and the mobility data from Västtrafik from [4]. These time-varying mobilities enabled us to model more complex behaviours of the pandemics. We finally modelled 324 pandemics. Indeed each of the 4 parameters was set to three different values : [0.5, 1, 2] and the 4 time-dependant mobilities were used.

### 3.1 The models

In this study, we define a model  $h_\theta$  as a function  $h$  defined on  $\mathbb{N}$  (corresponding to the days since the start of the pandemic), with parameters  $\theta$  and trained on the data  $\mathcal{D}$ . In the training phase,  $\hat{\theta}$ , an estimator of  $\theta$  is computed from  $\mathcal{D}$ , and used for the prediction. We considered two types of models : the first type correspond to models which are only trained on the time series we want it to predict (the number of hospitalized in our case), and the second type of models are trained on the time series we want to predict, but also on other time series that can be relevant to predict the number of hospitalized (the mobility and the number of infected). All of these models were implemented in Python, and are available on the github repository provided with this article (1). During the training or predicting

phase, the computation sometimes fails (for instance, when the matrix is non-invertible for the linear regression model). Instead of outputting a missing value, we let the model output the value of the moving average model (see 3.1.3), which can be interpreted as a naive output when the computation fails.

#### Task of a model :

Each model  $h$  is given :

1. A training set  $\mathcal{D}$
2. A reach of prediction  $r$
3. A confidence threshold  $\alpha$

And outputs :

1. A prediction  $\hat{Y}_r$
2. A  $(1 - \alpha)$  confidence interval on the prediction,  $I_{\alpha,r}$

The model was trained on the data  $\mathcal{D}$  to compute  $\hat{\theta}$  the parameter estimator and the output  $\hat{Y}_r = h_{\hat{\theta}}(r)$ .

#### 3.1.1 The SIRH model

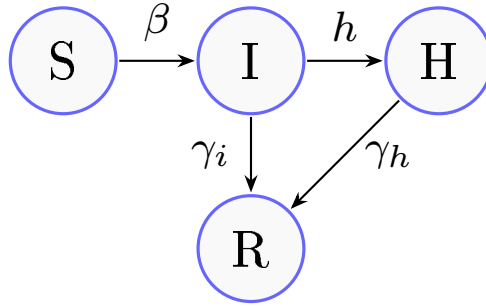


FIGURE 13 – Scheme of the SIRH model

The SIRH model (Fig.13) is an extension of the classic compartmental SIR (Susceptible-Infectious-Recovered) model used to describe the spread of infectious diseases. In the SIRH model, a fourth compartment, "H" for "Hospitalized," is added. Each compartment corresponds to the number of individuals in the health state corresponding to the compartment. The evolution of the number of individuals in each compartment is described by a system of ordinary differential equations :

$$\begin{cases} \frac{dS}{dt} = -\beta \frac{SI}{N} \\ \frac{dI}{dt} = \beta \frac{SI}{N} - \gamma_i I - hI \\ \frac{dR}{dt} = \gamma_i I + \gamma_h h \\ \frac{dH}{dt} = hI - \gamma_h H \end{cases} \quad (1)$$

At  $t = 0$ , the values of  $(S_0, I_0, R_0, H_0)$  are fixed to  $(10^6 - 1, 1, 0, 0)$ . As the system of equation cannot be directly solved, we use a Euler method to solve it :

We chose to fix  $dt = 0.001$  in the Euler's method.

To train this model, we minimize the least squares error between the curve of the number of hospitalized observed to the curve of the number of hospitalized of the training data with respect to  $\theta = (\beta, \gamma_i, \gamma_h, h)$ . We implemented some variations of the model in which either  $\gamma_i$ ,  $\gamma_h$  or both were fixed to the value 0.2 (see the Table.2).

|       | $\gamma_i$ | $\gamma_h$ |
|-------|------------|------------|
| SIRH1 | 0.2        | 0.2        |
| SIRH2 | 0.2        | free       |
| SIRH3 | free       | 0.2        |
| SIRH4 | free       | free       |

TABLE 2 – Difference between the SIRH models

In the prediction phase, an  $r$  day SIRH simulation is launched, with the parameter  $\hat{\theta}$  computed during the training phase. The initial values for  $S$  and  $I$  correspond to the last value of the fit of the training phase. The initial value for  $H$  corresponds to the last value of  $\mathcal{D}$ , the training data. The initial value of  $R$  is fixed by the previous values as the equation  $S_t + I_t + R_t + H_t = N$  is always true. The confidence interval of the prediction is computed thanks to a linearization and the use of the delta-method (see 3.1.7)

An SIRH model of the second type was implemented. It has the same structure but uses the mobility data and the number of infected in order to provide more accurate forecasts. The idea is the same, but there are two differences :

- $\beta$  varies with the time as a linear combination of the mobility :  $\beta_t = a + b \times m_t$  which is inspired from [4].
- The data is fitted to both the number of hospitalized and the number of infected.

To write it more formally, let  $H_\theta(t)$  and  $I_\theta(t)$  be the number of hospitalized and infected at time  $t$  in the SIRH model with parameters  $\theta$ . Let  $Y_{H,t}$  and  $Y_{I,t}$  be the number of hospitalized and infected at time  $t$  in the data. We have :

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^5}{\operatorname{argmin}} \sum_{t=1}^n \left( \frac{H_\theta(t) - Y_{H,t}}{\max(Y_{H,t})} \right)^2 + \left( \frac{I_\theta(t) - Y_{I,t}}{\max(Y_{I,t})} \right)^2$$

With  $\theta = (a, b, \gamma_i, \gamma_h, h)$  and  $m_t$  the mobility at time  $t$ . The normalization factors enables to put equal weight on the number of infected and hospitalized. Once again, a variation of the SIRH model was implemented, in which the value of  $\gamma_h$  and  $\gamma_i$  is fixed to 0.2. SIRH-multi-1 refer to the model in which  $\gamma_h$  and  $\gamma_i$  are free and SIRH multi 2 refer to the model in which  $\gamma_h$  and  $\gamma_i$  are fixed to 0.2.

### 3.1.2 ARIMA and VAR models

The ARIMA and VAR models are used for time-series forecasting and have outperformed many models in pandemic prediction (see [8] and [14]) The  $ARIMA(p, d, q)$  model is the sum of an  $AR(p)$  and a  $MA(q)$  model applied on the time series differentiated  $d$  times. It follows the equation :

$$Y_t^d = \alpha + \sum_{i=1}^p \beta_{t-i} Y_{t-i}^d + \sum_{j=1}^q \phi_{t-j} \epsilon_{t-j}$$

where  $Y_t^d$  is the time series at time  $t$ ,  $d$  is the order of the differentiation,  $\alpha$  is a constant,  $p$  is the order of the autoregressive part,  $q$  is the order of the moving average part and  $\epsilon_{t-j}$  is the difference between the prediction of the model and the real value at time  $t - j$ .

The coefficient are estimated through maximum likelihood estimation. This method is implemented in the `statsmodels` library, which directly provides prediction and confidence intervals. We performed a grid search on a single pandemic to identify the combination of parameters that would optimize the prediction accuracy. We found an optimal value for  $p = 3, d = 0, q = 3$ .

The VAR model is a multi-dimensional  $AR$  model, in which different variables are predicted. This model exploits the correlation between variables. Let  $Y_{1,t}, \dots, Y_{k,t}$  be the times series (in our case,  $k = 3$

and they correspond to the number of hospitalized, the number of infected and the mobility data).

$$\begin{aligned} VAR(p) : \begin{pmatrix} Y_{1,t} \\ Y_{2,t} \\ \vdots \\ Y_{k,t} \end{pmatrix} &= \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_k \end{pmatrix} + \begin{pmatrix} \phi_{11,1} & \phi_{12,1} & \cdots & \phi_{1k,1} \\ \phi_{21,1} & \phi_{22,1} & \cdots & \phi_{2k,1} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{k1,1} & \phi_{k2,1} & \cdots & \phi_{kk,1} \end{pmatrix} \begin{pmatrix} Y_{1,t-1} \\ Y_{2,t-1} \\ \vdots \\ Y_{k,t-1} \end{pmatrix} \\ &+ \cdots + \begin{pmatrix} \phi_{11,p} & \phi_{12,p} & \cdots & \phi_{1k,p} \\ \phi_{21,p} & \phi_{22,p} & \cdots & \phi_{2k,p} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{k1,p} & \phi_{k2,p} & \cdots & \phi_{kk,p} \end{pmatrix} \begin{pmatrix} Y_{1,t-p} \\ Y_{2,t-p} \\ \vdots \\ Y_{k,t-p} \end{pmatrix} + \begin{pmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \\ \vdots \\ \epsilon_{k,t} \end{pmatrix} \end{aligned}$$

Again, the  $\phi_{i,j,k}$  and  $c_i$  are estimated through maximum likelihood estimation with the `statsmodel` library. The confidence intervals are also directly provided by the library.

### 3.1.3 The moving average model

A mere moving average model was also implemented. It returns a constant prediction that correspond to the mean of the 7 past days, whatever the reach  $r$ . The confidence intervals are computed by assuming that the predictions follow a normal distribution of variance equal to the variance of the 7 last data-points. This model is used as a baseline. A model that does not manage to outperform the moving average model would not be useful.

### 3.1.4 Exponential regression

An exponential regression model was implemented. It corresponds to fitting the data of the number of hospitalized ( $Y_t$ ) to the function  $E_{a,b,c}(t) = a \times e^{bt} + c$ . The value of  $\theta = (a, b, c)$  is computed through a least square minimization method. The confidence interval on the prediction is estimated with the same method as SIRH (see 3.1.7).

An exponential regression of the second type was also implemented : The data of the number of hospitalized is fitted to the function  $E_{a,b,c,d,e}(t) = a \times e^{bm_t - i + ct + d \times inf_t - j} + e$ .  $m_t$  is the mobility data at time  $t$  and  $inf_t$  is the number of infected at time  $t$ . The value of  $\theta = (a, b, c, d, e)$  is computed through a least square method. The optimal value of the time lag  $i$  and  $j$  is optimized during the training phase through a grid search among all the values between 0 and 14. The confidence interval on the prediction is estimated with the same method as SIRH (see 3.1.7).

### 3.1.5 Machine learning models

In order to implement machine learning regressor, we converted the time-series  $Y_{t,t \in \{1, \dots, n\}}$  in a training set  $(X_i, Y_i)$  such that :  $\forall i \in \{1, \dots, n\}, X_i = (Y_{i-1}, Y_{i-2}, \dots, Y_{i-20})$ .

We then trained and optimized both regressors : the linear and the bayesian regressors, which were the only ones that provided reasonable outputs on the `scikit-learn` models including : linear regression, bayesian regression, Gradient boosting regressor, Random Forest regressor and support vector regression. The Bayesian ridge model was implemented with the default prior of the `scikit-learn` library (a gamma distribution with shape parameter equal to  $10^{-6}$  and scale parameter equal to  $10^{-6}$ ). The confidence interval for the linear regression prediction was computed as follow :

Let us suppose that the data follows a linear regression model :  $Y = X\beta + \epsilon$ , with  $Y \in \mathbb{R}^n$ ,  $X \in \mathbb{R}^{n \times d}$ ,  $\beta \in \mathbb{R}^d$  and  $\epsilon \sim \mathcal{N}(0, \sigma^2)$ . The least square estimator of  $\beta$  is  $\hat{\beta} = (X^T X)^{-1} X^T Y$ . If we have new data  $\tilde{X} \in \mathbb{R}^{1 \times d}$  that we want to predict, the prediction is given by :

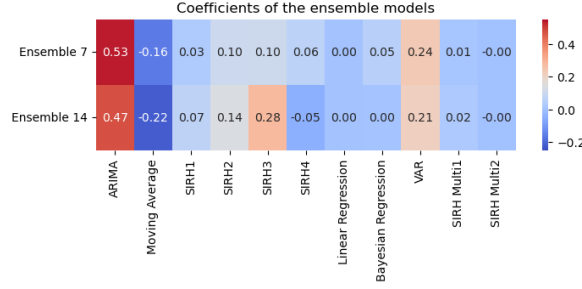


FIGURE 14 – Weights of the ensemble models

$$\begin{aligned}
\tilde{Y} &= \tilde{X}\hat{\beta} \\
&= \tilde{X}(X^T X)^{-1} X^T Y \\
&= \tilde{X}(X^T X)^{-1} X^T (X\beta + \epsilon) \\
&= \tilde{X}\beta + \tilde{X}(X^T X)^{-1} X^T \epsilon.
\end{aligned}$$

It implies that  $\tilde{Y}$  follows a normal distribution of expected value  $\tilde{X}\beta$  and variance  $\tilde{X}(X^T X)^{-1} \tilde{X}^T \sigma^2$ . The confidence interval on bayesian regression was directly computed with the variance of the parameters given by the scikit-learn library and the delta method.

### 3.1.6 Ensemble model

It has been shown (see [2] and [12]), that ensemble models, which combine the outputs of many models, can outperform individual models. We implemented an ensemble model which is a linear combination of the outputs of the 13 models described above, with the exponential models removed. Indeed, the exponential models almost always output very high values, and it had as a consequence that the optimization step of the linear regression problem failed. The weights of this model were found by minimizing the least-squared error between the prediction of the ensemble model and the real value of the number of hospitalized on a train set of approximately 80% of the pandemic generated. They are represented in the Fig.14. As the ensemble model only outputs a single value without confidence intervals, it is only evaluated with the RMSE metric (see 3.1.7).

### 3.1.7 Computing confidence intervals on the prediction

**Assumption :**

We suppose that the data of the pandemic observed follows the model  $h$ , of parameter  $\theta^* \in \mathbb{R}^d$ . Let  $Y_i$ ,  $i = 1, \dots, n$  be the number of hospitalized at each day. We suppose that :  $Y_i = h_{\theta^*}(i) + \epsilon_i$ , with  $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ , iid, and independent from all the other variables. The objective is to estimate  $\theta^*$ . We use  $\hat{\theta}$ , the least square estimator of  $\theta^*$  as an estimator of  $\theta^*$  :

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{i=1}^n (Y_i - h_{\theta}(i))^2$$

This problem is a non-linear regression problem on the regressors  $i$ . We also suppose that  $h_{\theta}$  is differentiable. Finally, we suppose that

$$\exists A_0 \in \mathbb{R}^{1 \times d}, \frac{1}{n} \sum_{i=1}^n \nabla_{\theta} h_{\theta^*}(i) \nabla_{\theta} h_{\theta^*}(i)^T \xrightarrow{n \rightarrow +\infty} A_0^T A_0. \quad (2)$$

This assumption is a condition for the sequence of the gradient of  $h_\theta$  not to move too much as  $n$  goes to infinity. This is a strong assumption, but required to have some essential convergences for the proof. Let :

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

$$h_\theta = \begin{pmatrix} h_\theta(1) \\ \vdots \\ h_\theta(n) \end{pmatrix}$$

We have :

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \|Y - h_\theta\|^2$$

Now, if  $\theta$  is close enough to  $\theta^*$ , we can write (from [13]) :

$$\forall i \in \{1, \dots, n\} : h_\theta(i) = h_{\theta^*}(i) + (\theta - \theta^*)^T \nabla_\theta h_{\theta^*}(i)$$

which leads to :

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \|Y - h_{\theta^*} - (\theta - \theta^*)^T \nabla_\theta h_{\theta^*}\|^2$$

Let us define :

$$\tilde{Y} = Y - h_{\theta^*}$$

$$\beta = \theta - \theta^*$$

$$\hat{\beta} = \theta - \hat{\theta}$$

and let us define the matrix  $A \in \mathbb{R}^{n \times d}$  such that  $\forall i \in \{1, \dots, n\}, \forall j \in \{1, \dots, d\}, A_{i,j} = \frac{dh_{\theta^*}}{d\theta_j}(i)$ .

The previous problem can be re-written as :

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \|\tilde{Y} - A\beta\|^2$$

This is now a linear regression problem on the regressors  $\nabla_\theta h_{\theta^*}(i)$ .

Let us solve this problem in the general case.

Let  $(A_i, \tilde{Y}_i)$  be the observations. In our case,  $A_i = (\nabla_\theta h_{\theta^*}(i))^T$ , the rows of  $A$ , and is a fixed quantity. Let us assume that  $\tilde{Y}_i = A_i \beta^* + \epsilon'_i$ , with  $\epsilon'_i \sim \mathcal{N}(0, \sigma'^2)$ .

The solution of this problem is explicitly (from [10]) :

$$\hat{\beta} = (A^T A)^{-1} A^T \tilde{Y}$$

This least-square estimator is unbiased :

$$\mathbb{E}[\hat{\beta}] = \beta^*$$

$$\hat{\beta} = \left( \sum_{i=1}^n A_i^T A_i \right)^{-1} \times \left( \sum_{i=1}^n A_i^T \tilde{Y}_i \right)$$

$$\hat{\beta} = \frac{n}{n} \left( \sum_{i=1}^n A_i^T A_i \right)^{-1} \times \left( \sum_{i=1}^n A_i^T \tilde{Y}_i \right)$$

$$\hat{\beta} = \left( \frac{1}{n} \sum_{i=1}^n A_i^T A_i \right)^{-1} \times \left( \frac{1}{n} \sum_{i=1}^n A_i^T \tilde{Y}_i \right)$$

Let us denote :

$$\hat{D} = \frac{1}{n} \sum_{i=1}^n A_i^T A_i, \quad \text{and} \quad \hat{\delta} = \left( \frac{1}{n} \sum_{i=1}^n A_i^T \tilde{Y}_i \right)$$

We have :

$$\hat{\beta} = \hat{D}^{-1} \hat{\delta}$$

$$\hat{D} \xrightarrow{a.s} D, \text{ with } D = \mathbb{E}[A_i^T A_i] = A_0^T A_0 \text{ (cf Hyp 2)}$$

$\hat{\delta}$  is the empirical mean of random variables independent but non identically distributed. We therefore cannot use the Kolmogorov's Strong Law of Large Numbers. There exists a law of large number for independent and non identically distributed random variables, which asserts that (see [15]) :

**Theorem 1.** *Let  $(X_k)$  be a sequence of independent random variables such that :*

1. *For all  $k \geq 1$ ,  $\mathbb{E}(X_k^2) < +\infty$ ;*
2. *There exists an increasing sequence  $(a_k)$  of strictly positive real numbers tending to infinity, such that*

$$\sum_{k=1}^{+\infty} \frac{\text{Var}(X_k)}{a_k^2} < +\infty.$$

*Then,  $\frac{S_n - \mathbb{E}S_n}{a_n}$  converges almost surely to 0. If additionally  $a_n^{-1} \mathbb{E}S_n \rightarrow m$ , then  $\frac{S_n}{a_n}$  converges almost surely to  $m$ .*

Let  $i \in \{1, \dots, n\}$ .

$$X_i = A_i^T \tilde{Y}_i = A_i^T A_i \beta^* + A_i^T \epsilon'_i.$$

The random values  $X_i$  are independent random variables.

$$\forall i \in \{1, \dots, n\}, \mathbb{E}[X_i^2] = \mathbb{E}[X_i]^2 + A_0^T A_0 \sigma'^2 < +\infty.$$

$$\sum_{i=1}^{\infty} \frac{\text{Var}(X_i)}{i^2} = \sum_{i=1}^{\infty} \frac{A_0^T A_0 \sigma'^2}{i^2} = A_0^T A_0 \sigma'^2 \sum_{i=1}^{\infty} \frac{1}{i^2} < +\infty.$$

$$\text{Moreover, } \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] = \frac{1}{n} \sum_{i=1}^n A_i^T A_i \beta^* = \frac{\beta^*}{n} \sum_{i=1}^n A_i^T A_i \xrightarrow{n \rightarrow \infty} \beta^* A_0^T A_0 \text{ (see 2).}$$

Therefore, with Theorem 1, we have :  $\frac{1}{n} \sum_{i=1}^n A_i^T \tilde{Y}_i \xrightarrow[a.s.]{} \beta^* A_0^T A_0$

$$\hat{\delta} \xrightarrow[a.s.]{} \delta = A_0^T A_0 \beta^*$$

$\hat{\beta} = \hat{D}^{-1} \hat{\delta} \xrightarrow[a.s.]{} D^{-1} \delta$ , as the following function  $\phi$  is continuous :

$$\phi : \begin{cases} \mathcal{GL}_n(\mathbb{R}) & \rightarrow \mathcal{GL}_n(\mathbb{R}) \\ M & \mapsto M^{-1} \end{cases}$$

Now, let us show that  $\hat{\beta}$  is asymptotically normal :

$$\begin{aligned} \sqrt{n}(\hat{\beta} - \beta^*) &= \sqrt{n}(\hat{D}^{-1} \hat{\delta} - \beta^*) \\ &= \sqrt{n}(\hat{D}^{-1} \hat{\delta} - \hat{D}^{-1} \hat{D} \beta^*) \\ &= \sqrt{n} \hat{D}^{-1} (\hat{\delta} - \hat{D} \beta^*) \\ &= \sqrt{n} \hat{D}^{-1} \left( \frac{1}{n} \sum_{i=1}^n A_i^T \tilde{Y}_i - \frac{1}{n} \sum_{i=1}^n A_i^T A_i \beta^* \right) \\ &= \frac{\sqrt{n}}{n} \hat{D}^{-1} \left( \sum_{i=1}^n A_i^T (\tilde{Y}_i - A_i \beta^*) \right) \\ &= \frac{1}{\sqrt{n}} \hat{D}^{-1} \left( \sum_{i=1}^n A_i^T \epsilon_i \right) \end{aligned}$$

This line is made of two terms. Let's show that each one of them converges in law.

$$\begin{aligned} \frac{1}{\sqrt{n}} \left( \sum_{i=1}^n A_i^T \epsilon_i \right) &= \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n A_i^T \epsilon_i \right) \\ &= \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n A_i^T \epsilon_i - 0 \right) \\ &\xrightarrow{\mathcal{L}} \mathcal{N}(0, \text{Var}(A_i^T \epsilon_i)) \end{aligned}$$

Yet, as  $\epsilon_i$  and  $A_i$  are independent, and  $\mathbb{E}[A_i^T \epsilon_i'] = 0$ ,  $\text{Var}(A_i^T \epsilon_i) = \mathbb{E}[A_i^T A_i \epsilon_i^2] = \mathbb{E}[A_i^T A_i] \sigma'^2 = A_0^T A_0 \sigma'^2$ .

Finally,  $\frac{1}{\sqrt{n}} (\sum_{i=1}^n A_i^T \epsilon_i') \xrightarrow{\mathcal{L}} \mathcal{N}(0, D \sigma'^2)$ .

On the other hand,  $\hat{D}^{-1} \xrightarrow{\mathcal{L}} D^{-1}$ , which is constant.

Finally, with Slutsky ([16]), we obtain that :

$$\begin{aligned} \sqrt{n}(\hat{\beta} - \beta^*) &\xrightarrow{\mathcal{L}} D^{-1} \mathcal{N}(0, D \sigma'^2) \\ &\xrightarrow{\mathcal{L}} \mathcal{N}(0, D^{-1} (D \sigma'^2) (D^{-1})^T) \\ &\xrightarrow{\mathcal{L}} \mathcal{N}(0, D^{-1} D \sigma'^2 (D^{-1})^T) \\ &\xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma'^2 D^{-1}) \\ &\xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma'^2 (A_0^T A_0)^{-1}) \end{aligned}$$



Let's get back to the first problem :  
As  $\beta^* = 0$  and  $\hat{\beta} = \hat{\theta} - \theta^*$ , we have :

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2(A_0^T A_0)^{-1})$$

and,

$$\hat{\theta} \sim \mathcal{N}(\theta^*, \frac{\sigma^2}{n}(A_0^T A_0)^{-1})$$

As a first conclusion, we have that  $\hat{\theta}$  is asymptotically normal.  
Let  $\Sigma$  be the covariance matrix estimated from the computation of  $\hat{\theta}$ .  
As  $\hat{\theta}$  is asymptotically normal, we can apply the delta-method :

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma)$$

$$\sqrt{n}(h_{\hat{\theta}} - h_{\theta^*}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \nabla_{\theta} h_{\theta^*}^T \Sigma \nabla_{\theta} h_{\theta^*})$$

And finally :

$$h_{\hat{\theta}} \rightarrow \mathcal{N}(h_{\theta^*}, \frac{1}{n} \nabla_{\theta} h_{\theta^*}^T \Sigma \nabla_{\theta} h_{\theta^*})$$

By estimating  $\frac{1}{n} \Sigma$  from `curve_fit`, we can compute the confidence interval of the prediction with the quantiles of the normal distribution. The gradient of  $h_{\theta}$  is approximated through numerical approximation :

$$\nabla_{\theta} h_{\theta}[i] \simeq \frac{h_{\theta + d\theta_i} - h_{\theta}}{dt}, \text{ with } dt = 0.0001 \text{ and}$$

$$d\theta_i = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_{i-1} \\ \theta_i + dt \\ \theta_{i+1} \\ \vdots \\ \theta_n \end{pmatrix}$$

## Assessing the performance of the models

### Metrics

Two metrics were used to assess the performance of the models. The first metric is the Weighted Interval Score (WIS), which is a metric commonly used in forecast evaluation (see [2] or [9]).

Let  $\alpha$  be in  $]0, 1[$ . Let  $\hat{y}$  be the prediction of the model and  $y$  the real value. Let  $[l, u]$  be the  $(1 - \alpha)$  confidence interval of the prediction. We define the Interval Score ( $IS$ ) as :

$$IS_{\alpha}([l, u], y) = \frac{2}{\alpha} \times (\mathbb{1}_{\{y < l\}}(l - y) + \mathbb{1}_{\{y > u\}}(y - u) + (u - l)).$$

This metric is made of three terms : a term of overprediction that punishes a model predicting a confidence interval which is above the real value, a term of underprediction that punishes a model whose

confidence interval is under the real value, and a term of range, that punishes too wide confidence intervals.

Let  $(\alpha_k)_{k \in \{1, \dots, K\}} \in ]0, 1[^K$  The WIS is defined as follow.

$$WIS([l, u], y) = \sum_{k=0}^K w_k IS_{\alpha_k}([l, u], y), \text{ with } (w_k)_{k \in \{1, \dots, K\}} \in \mathbb{R}_+^K \text{ weights chosen by the user.}$$

According to previous literature ([2]), we decided to set  $(\alpha_k)_{k \in \{1, \dots, K\}} = [0.02, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]$  and  $\forall k \in \{1, \dots, K\}, w_k = \frac{\alpha_k}{2}$ . One can notice that the WIS does not take into account the point prediction, but focuses on confidence interval accuracy.

The second metric chosen is the Root Mean Square Error (*RMSE*). With the same notations as above, we define the *RMSE* as follow :

$$RMSE(\hat{y}, y) = \sqrt{(y - \hat{y})^2}$$

This metric focuses on the point prediction, and does not take into account the confidence intervals.

The models were tested on all the 324 pandemics, on 14 data points different (at days 20, 40, 60, ..., 280). For each individual point, the models were trained on the previous days of the pandemic. A 7 and 14 days ahead prediction was asked, and  $[0.02, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]$  confidence-intervals were computed. The WIS and the RMSE of these predictions were then computed. For the analysis of the results, we decided to remove the points for which the number of hospitalized was below 100, (i.e less than  $10^{-4}$  hospitalized per 100 000 citizens) both point classification and model predictions were irrelevant at this period of the pandemic. Moreover, it is pointless to assess the performances of the models during the period in which there is no pandemic, as the interesting models are the ones that can predict the number of hospitalized during the waves. Not removing those points would lead to biased results, as they represent 44% of the dataset, and we would have concluded that the best model is the one that performs well when there is no pandemic.

### 3.2 Point classification

In order to compare the performance of the models at different point of the pandemic, we classified each day of the pandemic in six categories : *stable*, *increase*, *big increase*, *decrease*, *big decrease*. The  $R_{eff}$  number was used for point classification. This number is easily available from the Covasim simulation.  $R_{eff}$  represents the expected number of people that an infected person will infect. For a pandemic  $X_{i, i \in [1, \dots, n]}$  of reproduction number  $R_{i, i \in [1, \dots, n]}$ , and a day  $d \in [1, \dots, n]$ , the classification is made according Table.3.

| Condition             | Classification |
|-----------------------|----------------|
| $X[d] < 100$          | no pandemic    |
| $R[d] < 0.5$          | big decrease   |
| $0.5 \leq R[d] < 0.8$ | decrease       |
| $0.8 \leq R[d] < 1.2$ | stable         |
| $1.2 \leq R[d] < 3$   | increase       |
| $R[d] \geq 3$         | big increase   |

TABLE 3 – Classification algorithm

The values of the thresholds were chosen by hand. As the model are tested on the days [20, 40, 60, ..., 280], we labelled all of these points with the function described above. It resulted that among all those 4536 points, 698 were classified as 'big decrease', 579 as 'decrease', 388 as 'stable', 752 as 'increase', 132 as 'big increase', and 1987 as 'no pandemic'.

## 4 Appendix

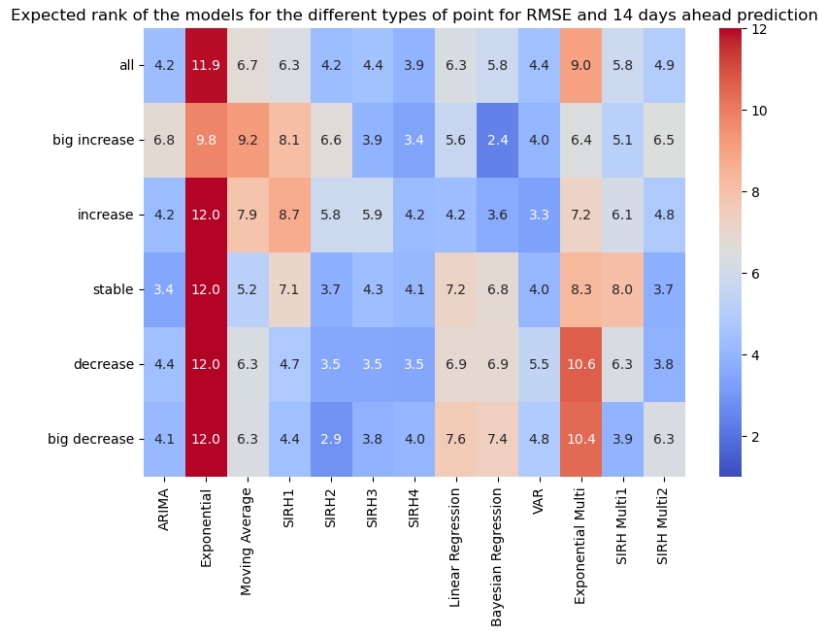


FIGURE 15 – Expected rank of the models for each type of point for 14-days ahead predictions and RMSE loss

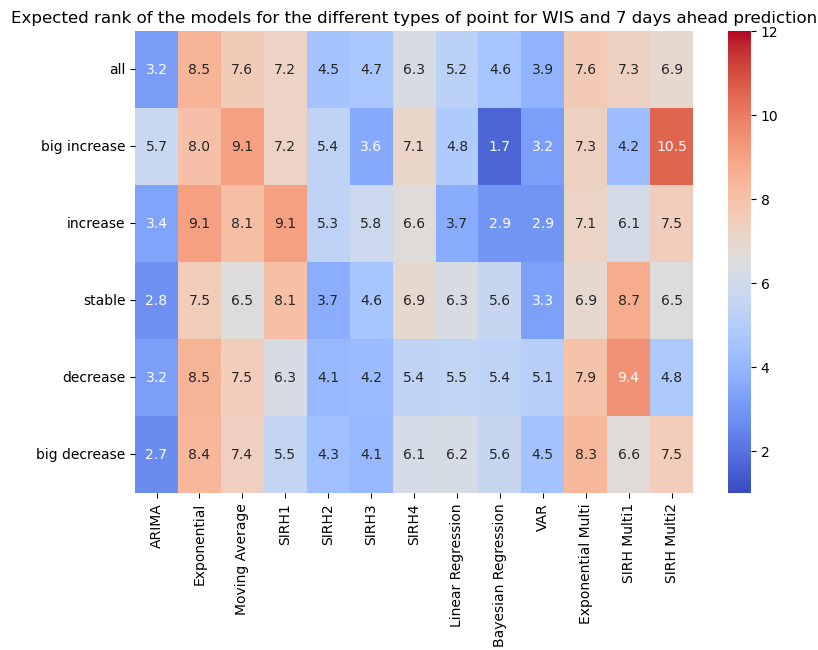


FIGURE 16 – Expected rank of the models for each type of point for 7-days ahead predictions and WIS loss

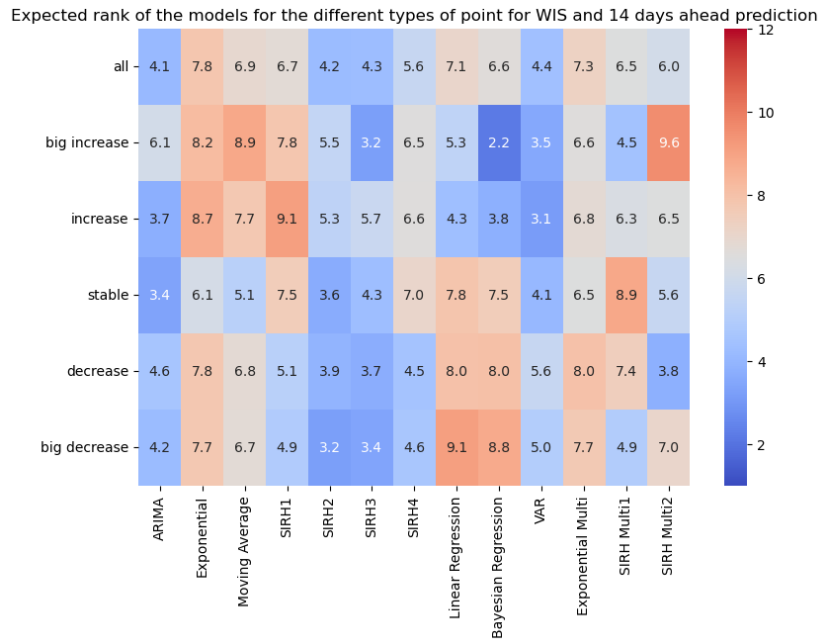


FIGURE 17 – Expected rank of the models for each type of point for 14-days ahead predictions and WIS loss

## Références

- [1] T BODINEAU. « Modélisation de phénomènes aléatoires : introduction aux chaînes de Markov et aux martingales ». In : *Ecole Polytechnique* (2015).
- [2] Estee Y CRAMER et al. « Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States ». In : *Proceedings of the National Academy of Sciences* 119.15 (2022), e2113561119.
- [3] Persi DIACONIS. « The markov chain monte carlo revolution ». In : *Bulletin of the American Mathematical Society* 46.2 (2009), p. 179-205.
- [4] Philip GERLEE et al. « Predicting regional COVID-19 hospital admissions in Sweden using mobility data ». In : *Scientific reports* 11.1 (2021), p. 24171.
- [5] Sharon K GREENE et al. « Nowcasting for Real-Time COVID-19 Tracking in New York City : An Evaluation Using Reportable Disease Data From Early in the Pandemic ». In : *JMIR Public Health Surveill* 7.1 (jan. 2021), e25538.
- [6] Emily HOWERTON et al. « Evaluation of the US COVID-19 Scenario Modeling Hub for informing pandemic response under uncertainty ». In : *Nature communications* 14.1 (2023), p. 7260.
- [7] Cliff C KERR et al. « Covasim : an agent-based model of COVID-19 dynamics and interventions ». In : *PLOS Computational Biology* 17.7 (2021), e1009149.
- [8] Tadeusz KUFEL. « ARIMA-based forecasting of the dynamics of confirmed Covid-19 cases for selected European countries ». In : *Equilibrium. Quarterly Journal of Economics and Economic Policy* 15.2 (2020), p. 181-204.
- [9] Juliette PAIREAU et al. « An ensemble model based on early predictors to forecast COVID-19 health care demand in France ». In : *Proceedings of the National Academy of Sciences* 119.18 (2022), e2103302119.
- [10] James POWELL. « Asymptotics for least squares ». In : ().

- [11] Hazhir RAHMANDAD, Ran XU et Navid GHAFARZADEGAN. « Enhancing long-term forecasting : Learning from COVID-19 models ». In : *PLoS computational biology* 18.5 (2022), e1010100.
- [12] Nicholas G REICH et al. « Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the US ». In : *PLoS computational biology* 15.11 (2019), e1007486.
- [13] Andreas RUCKSTUHL. « Introduction to nonlinear regression ». In : *IDP Institut für Datenanalyse und Prozessdesign, Zürcher Hochschule für Angewandte Wissenschaften* (2010), p. 365.
- [14] Aaron C SHANG, Kristen E GALOW et Gary G GALOW. « Regional forecasting of COVID-19 caseload by non-parametric regression : a VAR epidemiological model ». In : *AIMS public health* 8.1 (2021), p. 124.
- [15] Charles SUQUET. « Lois des grands nombres ». In : *Université des Sciences et Technologies de Lille* (2004).
- [16] Aad W VAN DER VAART. *Asymptotic statistics*. T. 3. Cambridge university press, 2000.
- [17] Cécile VIBOUD et al. « The RAPIDD ebola forecasting challenge : Synthesis and lessons learnt ». In : *Epidemics* 22 (2018). The RAPIDD Ebola Forecasting Challenge, p. 13-21. ISSN : 1755-4365.