Daily Covid-19 Deaths Prediction For Slovenia

David Susič
"Jožef Stefan" Institute
Ljubljana, Slovenia
david.susic@ijs.si

ABSTRACT

In this paper, models for predicting daily Covid-19 deaths for Slovenia are analysed. Two different approaches are considered. In the first approach, the models were trained on the fist wave dataset of state intervention plans, cases and country-specific static data for 11 other European countries. The models with the best performance in this case were the k-Nearest Neighbors regressor and the Random Forest regressor. In the second approach, a time-series analysis was performed. The models used in this case were Seasonal Autoregressive Integrated Moving Average Exogenous and Feed forward Neural Network. For comparison, all 4 models were tested on the second wave for Slovenia and the model with the best performance was Feed forward Neural Network, with a mean absolute error of 1.34 deaths.

KEYWORDS

Covid-19, deaths, predictions, machine learning

1 INTRODUCTION

The aim of this analysis is to find out whether we can predict Covid-19 deaths for Slovenia based on the characteristics of the epidemic in other European countries, and whether we can predict deaths based on a time series analysis of historical data (e.g. predicting for the second wave based on the first wave information). The main advantage of the first approach is that we do not need historical case and death data for the country for which we are making a prediction (in this case Slovenia), while the second approach is generally more accurate but relies on historical death data. The aim is also to find out which of the two approaches provides more accurate predictions. It is important to note that although this is a study for Slovenia, the results can be interpreted as a general assessment of the effectiveness of the methods described for predicting Covid-19 deaths and can be applied to any country for which the data are available.

The data used in this analysis are described in Section 2. Section 3 provides a description of the approaches and the models. Section 4 contains a discussion of the determination of the optimal parameters of the selected models. The results are given in Section 5. The conclusion, along with ideas for possible improvements, is given in Section 6.

2 DATA DESCRIPTION AND PREPARATION

The data used in this paper consist of daily Covid-19 related features at the country level. It contains 12 different Covid-19 related government interventions (school closing, workplace closing, cancel public events, restrictions on gatherings, close public transport, stay at home requirements, restrictions on internal movement, international travel controls, public information campaigns, testing policy, contact tracing, and facial coverings), Covid-19 related cases and deaths, and some static data, in particular the country's population, population density, median age, percentage of people over 65, percentage of people over 70, gdp per capita, cardiovascular death rate, diabetes prevalence, percentage of female and male smokers, hospital beds per thousand people, and life expectancy. To suppress anomalies in registered cases on Sundays and holidays, a 7-day moving average was used for both cases and deaths. The dataset covers the European countries of Slovenia, Italy, Hungary, Austria, Croatia, France, Germany, Poland, Slovak Republic, Bosnia and Herzegovina, and the Netherlands from January 22, 2020 to December 11, 2020. All of the countries chosen for this study are geographically next to one another and are thus expected to have similar course of epidemic. The data on government interventions, cases and deaths are derived from the "COVID-19 Government response tracker" database, collected by Blavatnik School of Government at Oxford University [4]. The intervention values range between 0-4 and represent their strictness, for example, if only some or all schools are closed. The static data are collected from a variety of sources (United Nations, World Bank, Global Burden of Disease, Blavatnik School of Government, etc.) [3]. The original data are publicly available online. The processed data used for the purpose of this study can be found online at https://repo.ijs.si/davidsusic/covid-seminar-data.

3 METHODS AND MODELS

Two different approaches were considered for the analysis. For the first part of the analysis, referred to as the country-specific approach, the models were trained on the data of government intervention plans, cases, deaths and country-specific static data for the 10 other European countries, with the aim of predicting deaths for Slovenia. In this case, the predictions were made for each day, disregarding the time order. For the second part of the analysis, a time series prediction was performed, using only the daily deaths for Slovenia as data.

3.1 Country-Specific Approach

In the country-specific approach, the selection of the base model was very important, as models that perform worse than the base model are not worthy of interpretation. The baseline was defined as

$$N_{\text{deaths}}(t) = N_{\text{cases}}(t - 14) \cdot M, \tag{1}$$

where $M = 0.023$ is the mortality rate factor of those infected, calculated as a weighted average of the mortality rates of the countries included in this study [2], and t denotes a specific day. This simple model implies, that the number of deaths on a given day t is equal to the number of new infections on the

Information Society 2021, 4–8 October 2021, Ljubljana, Slovenia

day $t - 14$, multiplied by the mortality rate factor. The regressor model that were tested are: Random Forest (RF), k-Nearest Neighbors (KNN), Stochastic Gradient Descent, Ridge, Lasso, and Epsilon-Support Vector. Description of all of the models can be found in the Python scikit-learn documentation [5]. The two that performed significantly better than the baseline were the KNN regressor and RF regressor. Other regression models performed the same or worse than the baseline model and were thus not used in the further analysis. All models were tested in the 10-fold cross-validation with the performance measures mean absolute error (MAE), mean squared error (MSE) and R^2 score on the data subset that does not include Slovenia. The measures are defined as:

$$\mathrm{MAE}(y, \hat{y}) = \frac{1}{n}\sum_{i=0}^{n-1} |y_i - \hat{y}_i|, \tag{2a}$$

$$\mathrm{MSE}(y, \hat{y}) = \frac{1}{n}\sum_{i=0}^{n-1} (y_i - \hat{y}_i)^2, \tag{2b}$$

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}, \tag{2c}$$

where $\hat{y}$ is the predicted value of the i-th sample, y_i is the corresponding true value, n is the sample size and $\bar{y}$ is the average true value $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_1$.

For each sample, additional features of the government interventions and cases were added for the previous days. The number of previous days was defined using the lookback parameter. Models were tested for lookback values between −28 and 0 days. The comparison is shown in Figure 1. It can be seen that the performance decreases in the range where the lookback is shorter than 14 days, but does not increase in the range where the lookback exceeds this value. The main reason for this is probably the fact that most deaths occur within the first 14 days of infection. A lookback of 14 days was used for further analysis as it was found to be the most appropriate.

3.2 Time-Series Approach

In the second approach, a time series analysis was performed. In this case, only daily deaths for Slovenia were used as data. The models used in this case were Seasonal Autoregressive Integrated Moving Average Exogenous (SARIMAX(p,d,q)(P,D,Q,m) [6] and Feed forward Neural Network (FFNN) [1].

The former is a combination of several different algorithms. The first is the autoregressive AR (p) model, which is a linear model that relies only on past p values to predict current values. The next is the moving average MA (q) model, which uses the residuals of the past q values to fit the model accordingly. The I(d) represents the order of integration. It represents the number of times we need to integrate the time series to ensure stationarity. The X stands for exogenous variable, i.e., it suggests adding a separate other external variable to measure the target variable. Finally, the S stands for seasonal, meaning that we expect our data to have a seasonal aspect. The parameters P, D, and Q are the seasonal versions of the parameters p, d, and q, and the parameter m represents the length of the cycle.

The FFNN structure included 10 input perceptrons - one for each death value in the last 10 days, a hidden layer of 64 perceptrons, and 1 output perceptron.

Since the future data of the time series contain the information about the past, a forward chaining approach was performed for n-fold cross validation. This means, that there is no random shuffling of the data. The test set must always be the final portion of the data - the final part of the date range. The concept of forward chaining is shown in Figure 2. The results of the 10-fold cross-validation of the predictions for 21 days are shown in Table 1.

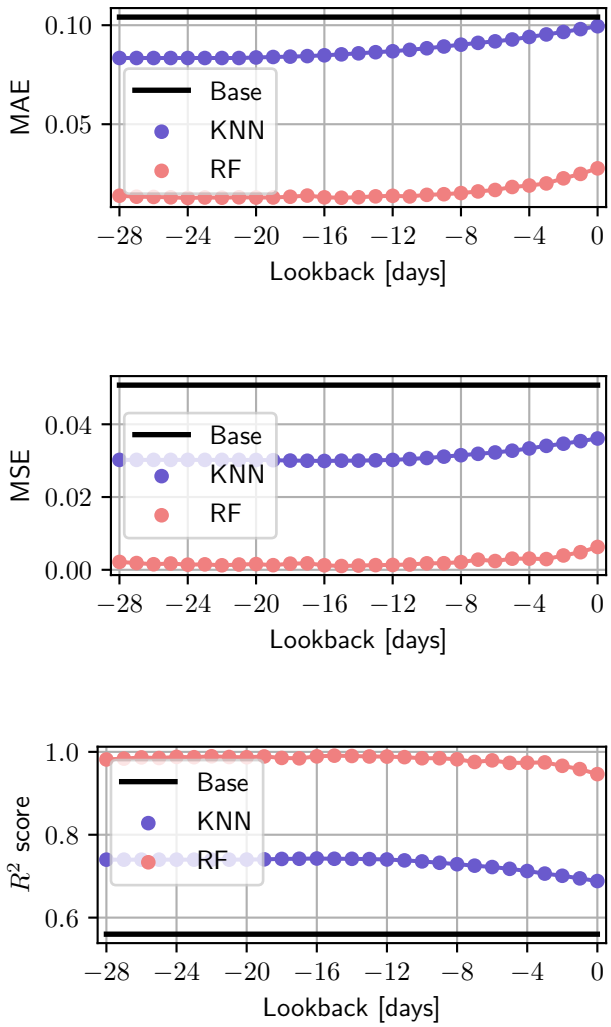

Figure 1: 10-fold cross validation performance measure of the models for different lookback parameter. The measures and its units are are: MAE [deaths/100k] (top), MSE [deaths^2/100k^2] (middle) and R^2 score (bottom)

Table 1: 10-fold cross-validation performance measures of the predictions for 21 days for SARIMAX and FFNN algorithms.

	MAE [deaths]	MSE [deaths^2]	R^2 score
SARIMAX	1.13	4.81	0.71s
FFNN	0.53	1.15	0.88

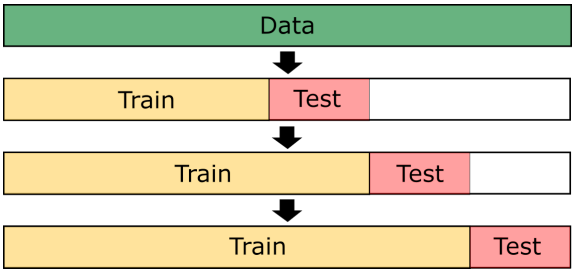

Figure 2: Forward chaining approach to time-series n-fold cross-validation.

4 MODELS' PARAMETERS SELECTION

The next step was to determine the optimal parameters of the selected models. For this purpose, the regressor models were trained on the same dataset used in the 10-fold cross-validation and tested on the data for Slovenia. For this particular case, different model parameters were tested to see which performed best. The MAE [deaths/100k] as a function of parameters K for the KNN and as a function of the number of trees for RF are shown in the Figures 3 and 4, respectively.

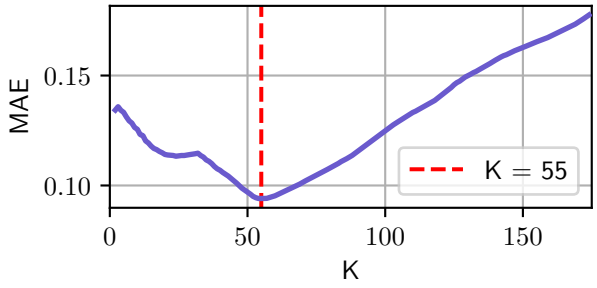

Figure 3: MAE of the KNN regressor as function of K.

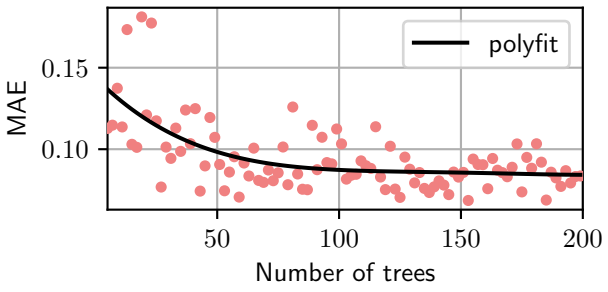

Figure 4: MAE of the RF regressor as a function of the number of trees.

For the KNN regressor, MAE has a minimum at $K = 55$, while for RF the fitting function shows that the appropriate number of trees is 100, since the model does not improve with additional trees at this point. It is important to note that since RF is random in the sense that it randomly selects a subset of features at each splitting decision, the results and hence the performance measures are also somewhat random. However, they do follow a certain trend that becomes apparent when a polyfit is applied. To reduce the randomness of the results, the average of 3 separate predictions was calculated for each number of trees.

To determine the best parameters of the SARIMAX model, the *auto_arima* algorithm from the Python *pmdarima* library was used [7]. The algorithm analyzes the given data and determines the best model and its parameters for that data. In this case, the selected model was SARIMAX(2, 1, 4)(4, 1, 1, 12).

In the case of FFNN, the parameter selection was omitted - the same model structure was always used.

5 RESULTS

With the optimal parameters selected, the graphs of the predictions can be plotted. The predictions of the country-specific approach are shown in Figure 5.

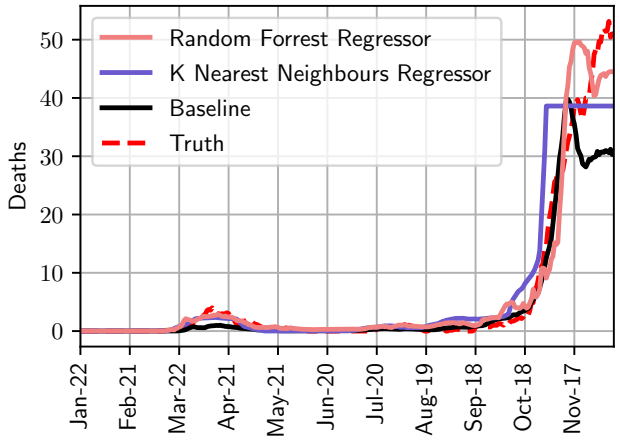

Figure 5: Deaths for Slovenia from 22.1.2020 to 11.12.2020. Models' predictions, compared to true values.

All models predicted the number of deaths for the first epidemic wave fairly accurately. As a result of the unrepresentative reporting of Covid-19 cases for the second wave, the base model predicts a much lower number of daily deaths. We can also see that the KNN regressor predicts the same value from a certain day forward. The reason for this is most probably that the algorithm always finds the same k=55 neighbors, thus always predicts the same value. To avoid this, a larger dataset would be required. MAE for RF, KNN and baseline are shown in Table 2.

Table 2: MAE comparison of the country-specific models for the interval from 22.1.2020 to 11.12.2020.

	RF	KNN	baseline
MAE [deaths]	5.41	5.39	5.48

The predictions for the time interval between 21.11.2020 and 11.12.2020 for the time-series approach are shown in Figure 6. MAE for FFNN and SARIMAX, shown in Table 3, are substantionally lower than MAE of the country-specific models. However, the accuracy decreases as the prediction time interval increases.

Table 3: MAE comparison of the time-series models for the interval from 21.11.2020 to 11.12.2020.

	FFNN	SARIMAX
MAE [deaths]	1.24	2.27

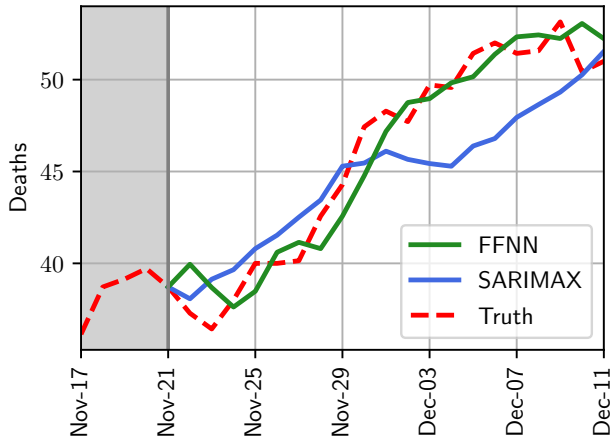

Figure 6: Slovenia deaths from 21.11.2020 to 11.12.2020. Time-series models' predictions, compared to true values.

To determine the overall best model for such predictions, all 4 models were tested on the second epidemic wave. The predictions are visualized in the Figure and the MAEs [deaths] are listed in the Table 4.

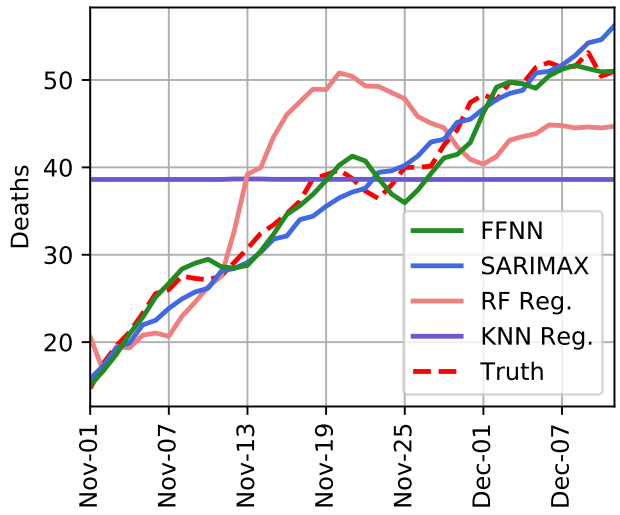

Figure 7: Slovenia deaths from 1.11.2020 to 11.12.2020. Models' predictions, compared to true values.

Table 4: MAE comparison of the models for the interval from 1.11.2020 to 11.12.2020.

	FFNN	SARIMAX	RF Reg.	KNN Reg.
MAE [deaths]	1.34	1.67	6.46	8.85

It can be seen that in this case the time-series approach is more accurate than the country-specific one. However, for longer time intervals, the country-specific approach is better because it does not rely on past data. It is important to note that the country-specific models' error are actually lower when making predictions from the start of the epidemic. The reason for this is that for the first 6 months, the numbers of deaths were very low as can be seen in the Figure 5.

The best performing model overall is the FFNN with the MAE of 1.34 deaths. The reason for the best performance of this model is probably that it had a relatively high number of input parameters. The input layer consisted of 10 perceptrons, i.e. each prediction was based on the values of the last 10 days.

6 CONCLUSION

In this paper, two different approaches to predicting Covid-19 deaths for Slovenia were tested. Both approaches turned out to be reliable. The main implications of the presented study are that for short time intervals the time series approach is much more accurate than the country-specific approach. The advantage of the country-specific approach is that it can predict the number of deaths for a given day, based on the number of cases, countermeasures and country-specific static data, without necessarily having information about the past. On the other hand, for the prediction of the second wave, where we already know the course of the epidemic in the first wave, the time series approach is better - at least for the prediction for Slovenia. In the future studies, predictions for the third and fourth waves will be analysed.

REFERENCES

[1] Francois Chollet et al. 2015. Keras. https://github.com/fchollet/keras.

[2] Ensheng Dong et al. 2020. An interactive web-based dashboard to track covid-19 in real time. *The Lancet Infectious Diseases*, 20, 5. DOI: 10.1016/S1473-3099(20)30120-1. http://doi.acm.org/10.1016/S1473-3099(20)30120-1.

[3] Thomas Hale et al. 2020. A cross-country database of covid-19 testing. *Scientific Data*, 7, 345. DOI: 10.1038/s41597-020-00688-8. http://doi.acm.org/10.1038/s41597-020-00688-8.

[4] Thomas Hale et al. 2021. A global panel database of pandemic policies (oxford covid-19 government response tracker). *Nature Human Behaviour*, 5, 3529–538. DOI: 10.1038/s41562-021-01079-8. http://doi.acm.org/10.1038/s41562-021-01079-8.

[5] Fabian Pedregosa et al. 2012. Scikit-learn: machine learning in python. *Journal of Machine Learning Research*, 12, (January 2012).

[6] Skipper Seabold and Josef Perktold. 2010. Statsmodels: econometric and statistical modeling with python. *Proceedings of the 9th Python in Science Conference*, 2010, (January 2010).

[7] Taylor G. Smith et al. 2017. pmdarima: arima estimators for Python. [Online; accessed 9.1.2021]. (2017). http://www.alkaline-ml.com/pmdarima.