IRSTI 27.41.17

R. Tabarek[1], B. Murat[2], A. Kaiyr[3], N. Smadyarov[4], D. Issa[5]
[1,2,3,4]Suleyman Demirel University, Kaskelen, Kazakhstan
[5]Google

TIME SERIES ANALYSIS TO FORECAST COVID-19 CASES IN CENTRAL ASIA

Abstract. According to the study, the cases are expected to increase in the upcoming days. An exponential rise in the number of cases is also noticeable in a time series analysis. The current forecast models are expected to help the government and medical professionals plan for future circumstances and enhance healthcare system readiness. The proposed study employs a support vector regression model for forecasting the overall number of deaths, recovered cases, cumulative number of reported cases, and regular case count. The starting information is retrieved from the 1st of March to the 30th of April, 2021 (61 Days). The model predicts deaths, recoveries, and the total number of confirmed cases with an accuracy of over 97 percent, and regular new cases with an accuracy of 87 percent. The findings point to a Gaussian reduction in the number of cases, which may take another 3 to 4 months to reach the bare minimum of no new cases registered.

Keywords: COVID19, Support vector regression, Data analysis, Central Asia.

***

Аңдатпа. Зерттеуге сәйкес, алдағы күндері істер көбейеді деп күтілуде. Аурулар санының экспоненциалды өсуі уақыт сериялары талдауда да байқалады. Ағымдағы болжам модельдері үкіметке және медициналық мамандарға болашақ жағдайларды жоспарлауға және денсаулық сақтау жүйесінің дайындығын арттыруға көмектеседі деп күтілуде. Ұсынылып отырған зерттеуде өлімнің жалпы санын, қалпына келтірілген жағдайларды, тіркелген жағдайлардың жиынтық санын және жүйелі жағдайларды болжау үшін тірек-векторлық регрессиялық модель қолданылады. Бастапқы ақпарат 2021 жылдың 1 наурызынан 30 сәуіріне дейін алынады (61 күн). Модель өлім-жітімді, қалпына келтіруді және расталған жағдайлардың жалпы санын 97 пайыздан жоғары дәлдікпен, ал әдеттегі жаңа жағдайларды 87 пайыздан болжайды. Зерттеулер нәтижелері Гауссияда ауру санының азаюына назар аударады, бұл жаңа тіркелмеген жағдайлардың минимумына жету үшін тағы 3-4 ай қажет болуы мүмкін.

Түйін сөздер: COVID19, Векторлық регрессияны қолдау, Деректерді талдау, Орталық Азия.

***

Аннотация. Согласно исследованию, в ближайшие дни ожидается рост числа случаев заболеваний COVID19. Экспоненциальный рост количества случаев также заметен при анализе временных рядов. Ожидается, что текущие модели прогнозов помогут правительству и медицинским специалистам планировать будущие обстоятельства и повысить готовность системы здравоохранения.
Предлагаемое исследование использует регрессионную модель опорного вектора для прогнозирования общего числа смертей, выздоровевших случаев, совокупного числа зарегистрированных случаев и регулярного подсчета случаев. Стартовая информация извлекается с 1 марта по 30 апреля 2021 года (61 день). Модель прогнозирует смерти, выздоровление и общее количество подтвержденных случаев с точностью более 97 процентов, а также регулярные новые случаи с точностью 87 процентов. Полученные данные указывают на сокращение числа случаев по Гауссу, которое может занять еще 3-4 месяца, чтобы достичь минимума, когда не зарегистрировано новых случаев.

Ключевые слова: COVID19, Поддержка векторной регрессии, Анализ данных, Центральная Азия.

Introduction

The particularly serious syndrome coronavirus (COVID-19) found in the City called of Wuhan in early December 2019 sparked a global coronavirus outbreak, with China being the epicenter of the infection [1]. The indications of this serious illness include fever, shortness of breath, and a dry cough [2].

After that, the disease has expanded to over 206 countries or areas of the world as a result of human travel, with the United States and Europe emerging as new major elements [3, 4]. On March 11th, 2020, the World Health Organization declared this outbreak to be a pandemic [5].

Central Asian economies are still vulnerable to risks similar to the ones generated by COVID-19. Long-term resilience will be hampered unless underlying problems along with an overdependence on commodities, migrant workforce (particularly in Tajikistan and Kyrgyzstan), small amounts of diversity, dual job markets, and weak social protection systems can indeed be resolved.

Within this regard, it is critical to develop models that are both algorithmically capable and practical in order to assist politicians, health workers, and the community at large. Characterizing the disease and offering a predicted estimate of the number of potential daily incidents will help the

healthcare system plan for the influx of new patients. Forecasting and monitoring the global disease hazard can be achieved with mathematical prediction models [6].

The study's main purpose is to predict potential COVID-19 cases in Central Asia using Time Series Forecasting method.

Literature Review

To predict the amount of COVID-19 cases during the next 10 days, the author used a flower pollination algorithm and the Salp Swarm Algorithm. A detailed overview of COVID-19's forecast future situation in Italy is given by A. Remuzzi and G. Remuzzi [8]. Perc et al. [7] proposed a simple iterative algorithm which only includes the everyday values of COVID-19 affirmed cases as inputs. The approach takes into account expected recoveries and deaths to assess the maximum daily growth rates that will lead to balanced and decreasing numbers rather than exponential growth. According to the forecasts, regular rates of growth should be held below 5% if we are to witness plateaus anytime soon. [9] examines the details available for the COVID-19 incidents in the 6 Western nations of Italy, Canada, Germany, the United Kingdom, France, and the United States using a segmented Poisson model. [10, 11, 12] have suggested several related studies.

They use a wide range of methods and boundaries for predicting. Regardless, assessing approaches go interdependent with overwhelming tasks (specialized and conventional). Their research looks at these issues and offers a number of recommendations to those fighting the global COVID-19 pandemic today (Table 1).

Table 1. The most recent research on COVID-19 time series data

#	Ref	Dataset	Methods	Region	Limitations
1	16	Selfgathered	Data mining (PNN+cf)	China	Detection of only suspected cases
2	17	CCDCP	Composite Monte-Carlo (CMC)	China	Focused on recommendation only
3	18	WHO	Logistic inference	Hybrid countries	Detection of only death cases
4	19	WHO	Modified auto encoders (MAE)	China + Hybrid countries	Measuring impact in qualitative way

5	20	DATA S-013	Gompertz model + Bertalanffy model	China	Detection of only death cases
6	21	ICD	Weibull equation and Hill equation	Github	Considered infection rate only
7	22	ICD	Charlson Comorbidity	Github	Measure risk rate

Methodology and Vizualization

The suggested forecast for COVID19 spread in Central Asian countries is discussed in this research, which is implemented in Python 3.6 employing support vector regression. The model's concepts are described in the methodology section, followed by an analysis. The findings are presented and discussed.

The Novel Coronavirus 2019 dataset .csv document can be downloaded from https://www.kaggle.com/sudalairajkumar/novel-coronavirus-virus-2019dataset. Just for Central Asian countries, a distinct .csv file is generated from the worldwide datasets. Overall Deaths, Overall Recovered, and Overall number of registered COVID19 patients are listed in the columns on a daily basis from March 1, 2021 to April 30, 2021. (61 days).

All of the material is presented in a cumulative format. Difference time series from the accumulated datasets to obtain values based on regular new case basis was estimated. As a result, paper introduces multiple columns to the set of data: three for total cases and three for regular new cases of deaths, rehabilitation, or verified COVID19 persons.

Both for linear and nonlinear regression types, support vector regression is a common option for forecasting and curve fitting. SVR is founded on support vector machine (SVM) elements, where support vectors are essentially closer spots towards the produced hyperplane in an n-dimensional feature space that distinguishes the pieces of data about the hyperplane.

More information about the SVR and SVM can be provided at [13, 14, 15]. The fitting is performed by the SVR model, as can be seen in Fig. 1. The modified hyperplane equation is $y = wX + b$, where w denotes weights and b denotes the intercept at $X = 0$. Epsilon (ε) is a symbol for the tolerance margin.

After that, the performance parameters of the model are assessed in order to verify whether they are reliable in forecasting the result. Table 2 calculates and displays the mean square error (MSE), root mean square error (RMSE), R^2 score, and percentage accuracy.

Table 2. The output parameters of the support vector regression approach with RBF kernel and a 10% fitting confidence interval

Data	MSE	RMSE	Reg. Score	% Accuracy
Total Deaths	0,00849	0,092142	0,986812	99%
Total Recovered	0,030289	0,174036	0,973437	97%
Daily Confirmed	0,109448	0,33083	0,8749	87%
Cumulative Confirmed	0,012856	0,113386	0,988613	99%
Daily Deaths	0,130847	0,361727	0,821829	82%

Results and Discussion

The forecasting of expected results for a time series entails several data manipulation steps to obtain the overall trend, which must correlate with the trend from the original dataset from the history. Although the historical data is in cumulative form, it is clear that the expected time series would follow a decreasing gaussian trend now that RBF kernels were included in this model. A transformation, as mentioned below, would maintain the decreasing trend continuing. Several steps were added to the algorithm that can provide assistance in achieving the main goal.

The projected time series to every case separately for the next 60 days were extracted, beginning on April 30th or the 61st day from the start. As a result, the 60-day forecast with the previous 61 days should be combined for better representation. The forecast column has declining values in it.

As a result, the time series difference was calculated and after utilization took place with the absolute values of the difference time series. The difference time series is inverted, resulting in an increasing trend that saturates after a certain point. Then the maximum value of the previous time series was added to the total number of the elements of the time series.

This provides the preservation of the pattern and its visualization in a cumulative way. Figures 3 and 4 demonstrate the plots of historical and forecasted values.

This transformation is not necessary for the forecasting of regular new cases time series prediction.

The results demonstrate that the model operated well when it came to matching with cumulative cases, but not so well when it came to fitting regular

cases. The daily data indicates that there are several spikes, reducing the model's precision and predictability.

If the present rate of daily new cases continues, the cumulative number of infected people could reach 56 000 by the second week of June, according to the model. Following the latest trends, the overall number of people who could die may exceed 1500 by the second week of June.

Furthermore, if there are more daily spikes in deaths and new cases, the overall number of infected people may increase, delaying the flattening of the slope. The spikes trigger non-stationarity in the model, making regression models challenging for forecasting accurately.

However, if the spikes are managed in the upcoming weeks with required physical distancing and containment steps, the curve may be flattened by the end of the second week of June.

Conclusion

Considering a comparative analysis with some Central Asian countries, the situation can still be regulated if reasonable precautions including quarantine and city disinfection are tightly pursued. The forecast models would aid the state and medical workforce in becoming more ready for future scenarios and increasing healthcare system readiness.

The suggested technique forecasts the overall number of COVID19 infected cases, daily new cases, overall number of deaths, and daily new deaths. The amount of people who have been recovered is also estimated. Future patterns have been forecast utilizing a robust machine learning model called support vector regression, judging by recent patterns.

The SVR was shown to overperform other linear, polynomial, and logistic regression models in terms of predictability consistency. The suggested approach addresses the dataset's variability. The model has an accuracy of over 97 percent in forecasting deaths, recovered cases, and the total number of reported cases, as well as an accuracy of 87 percent in forecasting regular new cases.

The disease has spread widely, and if adequate containment steps such as physical separation and hygiene are followed, the spikes in the datasets can be reduced and thereby limit the progression of COVID19.

References

1 Lu, H., Stratton, C.W., Tang, Y.W. Outbreak of pneumonia of unknown etiology in Wuhan, *China: the mystery and the miracle. J. Med Virol*. 92 (4), (2020): pp. 401–402.

2 Kelvin, A.A., Halperin, S. Covid-19 in children: the link in the transmission chain. Lancet Infect Dis. (2020): pp. 633-634.
3 Coronavirus: Europe now epicentre of the pandemic, says WHO Published13 March 2020. Bbc news. URL: https://www.bbc.com/news/world-europe-51876784.
4 The effects of virus variants on COVID-19 vaccines. Sky news (2020). URL: https://news.sky.com/story/coronavirus-how-the-us-isbecoming-the-new-epicentre-of-the-covid-19-pandemic-11964550.
5 Who director- general's opening remarks at the media briefing on covid19-11 March 2020. World health organization. 11 March 2020. URL: https://covid19.who.int/?gclid= EAIaIQobChMIxmkpZ2T6gIVRJ3VCh3ZVQeYEAAYASABEg LVg D BwE.
6 Hiteshi Tandon, Prabhat Ranjan, Tanmoy Chakraborty, Vandana Suhag Coronavirus (COVID-19): ARIMA based time-series analysis to forecast near future (2020): pp. 1-11.
7 Perc, M., Gori˘sek Miksi´c N, Slavinec M, Sto˘zer A. Forecasting covid19. *Frontiers in Physics* 8 (127), (2020): pp.1-5.
8 Remuzzi, A., Remuzzi, G. Covid-19 and Italy: what next? The Lancet. *Health Policy,* 395 (2020): pp.1225-1228.
9 Zhang, X., Ma, R. and Wang, L., 2020. Predicting turning point, duration and attack rate of COVID-19 outbreaks in major Western countries. Chaos, Solitons & Fractals, 135, p.109829.
10 Hussain, A.A., Bouachir, O., Al-Turjman, F. and Aloqaily, M., AI techniques for COVID-19. *IEEE Access*, 8, (2020): pp.128776-128795.
11 Rahman, M.A., Zaman, N., Asyhari, A.T., Al-Turjman, F., Bhuiyan, M.Z.A. and Zolkipli, M.F., 2020. Data-driven dynamic clustering framework for mitigating the adverse economic impact of Covid-19 lockdown practices. Sustainable cities and society, 62, p.102372.
12 Waheed, A., Goyal. M., Gupta, D., Khanna, A., Al-Turjman, F., Pinheiro, P.R. Covidgan: data augmentation using auxiliary classifier gan for improved covid-19 detection. *IEEE Access* 8 (2020): pp. 91916–91923.
13 Hastie TJ . The elements of statistical learning: data mining, inference, and pre- diction. New York: Springer; 2008. – 698 p.
14 Drucker, H . Support vector regression machines. In: Advances in neural informa- tion processing systems. MIT Press. p. 155–61 .
15 Sci-kit-learn. (2020). URL: https://scikit-learn.org/stable/auto _ examples/svm/ plot _ svm _ regression.html .
16 Fong, S.J., Li, G., Dey, N., Crespo, R.G., Herrera-Viedma, E. Finding an accurate early forecasting model from small dataset: a case of

2019ncov novel coronavirus outbreak. International Journal of Interactive Multimedia and Artificial Intelligence, 6 (1), (2020): pp. 132-140.

17 Fong, S.J., Li, G., Dey, N., Crespo, R.G. and Herrera-Viedma, E. Composite Monte Carlo decision making under high uncertainty of novel coronavirus epidemic using hybridized deep learning and fuzzy rule induction. *Applied soft computing*, 93, (2020): p.106282.

18 Batista, M. Estimation of the final size of the second phase of the coronavirus covid 19 epidemic by the logistic model. *Medrxiv* (2020): pp. 1-11.

19 Hu, Z., Ge, Q., Li, S., Jin, L., Xiong, M. Evaluating the effect of public health intervention on the global-wide spread trajectory of covid-19. *Medrxiv* (2020): pp.1-17.

20 Jia, L., Li, K., Jiang, Y., Guo, X., et al. Prediction and analysis of coronavirus disease 2019. arXiv:200305447. (2020): pp. 1-19.

21 Kumar, J., Hembram, K. Epidemiological study of novel coronavirus (covid-19). arXiv:200311376. (2020): pp.1-9.

22 DeCaprio, D., Gartner, J., Burgess, T., Kothari, S., Sayed, S. Building a covid-19 vulnerability index. arXiv:200307347. (2020): pp. 1-12.