2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) | 978-1-6654-2642-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/I-SMAC52330.2021.9641006

The study of the effect of the data collected during vaccination period on the prediction of the number of Covid-19 cases

1st Amir Ahmad
College of Information Technology
United Arab Emirates University
Al Ain, UAE
amirahmad@uaeu.ac.ae

2nd Santosh Kumar Ray
Department of Information Technology
Al Khawarizmi International College
Al Ain, UAE
santosh.ray@kic.ac.ae

3rd Ch. Aswani Kumar
Faculty of Engineering and IT
Vellore Institute of Technology
Vellore, India
cherukuri@acm.org

4th Apurva Anand
Mechanical Engineering Department, School of Engineering
Babu Banarasi Das University
Lucknow, India
apurva2050@yahoo.co.in

5th Muhsin Jabbar Cheratta
Independent researcher
Kannur, India
muhsinct@gmail.com

***Abstract*—Coronavirus disease (Covid-19) is a serious health problem for the world. Most of the countries are affected by this infectious disease. Many countries have started vaccination against Covid-19. The number of confirmed cases every day changes rapidly. Public health planners want to know these numbers in advance to arrange health facilities accordingly. Many machine learning models have been developed for the prediction of the number of Covid-infected people. The accuracy of these models depends upon the training data. Data collected during the period when there is no vaccination and data collected during the vaccination period have different properties. The models trained on different datasets perform differently. In this paper, we study the effect of the data collected during the vaccination period. The study will be helpful in generating more accurate prediction models for the vaccination period.**

***Index Terms*—Covid-19, confirmed cases, regression methods, vaccination.**

I. Introduction

Covid-19 is a novel corona virus that affects the respiratory system [1]. It was first detected in Wuhan, China [2] and declared a pandemic on 11^{th} March 2020 [4]. There is no specific medicine to cure it. Various vaccines have been developed to handle it [9], [12]. However, due to the scarcity of vaccines not many countries have been adequately vaccinated.

Public health experts in each country want to know the number of Covid-19 cases in future in order to prepare health facilities accordingly. There are different methods to accomplish this task [6], [19]. One of the approaches is converting the prediction problem into a regression problem using window approach then apply regression methods to forecast the number of Covid-19 cases in the future [5], [6], [14], [15]. This approach has shown excellent results.

Almost all the research works on this approach used the data collected during the period when there was no vaccination. It is important to know the effect of data collected during the vaccination period on the accuracy of this approach. In this paper, we study the effect of the data gathered during vaccination period on the accuracy of the approach. The data of four countries, Canada, Germany, the UAE and the USA, where vaccination drive has been started are used in the study. As per data recorded till 1st June 2021, more than 30 % population of these countries have taken at least one dose of vaccine [3].

The paper is organized in the following ways. The section 2 provides a literature survey. The experimental setup is discussed in Section 3. Section 4 includes experimental results and discussion. Conclusion and future work are presented in the last section.

II. Literature survey

There are many methods which have been employed to forecast the number of Covid-19 cases, some of them are discussed below.

1) **Susceptible-Infectious Recovered model** - Susceptible-Infectious Recovered (SIR) model and its extensions are based on ordinary differential equations. They have been used to model infectious diseases. SIR models have been employed to predict the number of Covid-19 cases [8], [23], [26], [28]
2) **Regression methods** - Regression methods are applied to predict the number of Covid-19 cases. Two approaches have been applied. In the first approach, a regression

x1	x2	x3	y
230	450	350	233
450	350	233	267
350	233	267	345
233	267	345	391
267	345	391	349
345	391	349	378
391	349	378	212

TABLE I: Dataset created by using the time series given in the Section 3.

model such as logistic curve is assumed and the parameters of the model are learnt on the basis of the number of Covid-19 cases time series [6], [7], [24]. Then the model is employed to forecast the number of Covid-19 cases in the future.

In the second approach, the number of confirmed cases time series is converted into a regression problem using moving the window method. Regression methods such as linear regression, neural networks, random forest regression etc. are used to learn the relationship [5], [6], [14], [15]. These trained regression models are employed to forecast the number of Covid-19 cases in the future. These approaches have been applied to forecast the number of Covid-19 cases [11], [13], [17], [27].

3) **Others** - Networks or graphs consisting of nodes are used to represent human to human interactions. A node represents a human whereas an edge represents the interaction between humans. Network analysis has been used to study the spread of Covid-19 [20], [21]. Social media such as Twitter has been investigated to forecast the Covid-19 spread [10], [18]. Internet search queries for Covid-19 related words are used to forecast the number of Covid-19 cases [22], [25].

In this paper, we apply window technique to convert the time-series into a regression problem. Then, regression methods are applied on it.

III. Experimental setup

The number of Covid-19 cases each day for a given period is a time series. This time series can be converted into a regression problem using the moving window approach. There are two parameters in this approach; width of the window (w) and number of days ahead for the prediction (k). We explain this approach with an example;

For a given time series

230, 450, 350, 233, 267, 345, 391, 349, 378, 212

Table 1 is created using moving window of size 3 and k = 1.

This can be considered as a regression problem with three input variables (x1, x2 and x3) and one output variable (y).

We will discuss the experiment setup used to analyze the effect of data collected during the vaccination period on the accuracy of the prediction models. In our case, we considered two time series, one collected during with vaccination days and the other gathered during the period without vaccination. The number of days without vaccination is N and V as the number of days with vaccination. A model is created using the data of N days and is used to predict the number of cases in the last 15 days of the vaccination period. It will be called Experiment-1. To compare the results, with the case when there was no vaccination, we divided the days without vaccination into two parts, N-V days and next V days (we have N days). The model is trained on the data created by using the data of N-V days and tested on the last 15 days of the next V days. This experiment will be called Experiment-2. The purpose is to have a similar experimental setup for both the cases, with vaccination and without vaccination. Another experiment (Experiment-3) was also carried out in which only the vaccination days data was used to create the prediction model.

Data from different countries are used in the experiments. The information about the data is provided in Table 2. A window size of 7 and k = 1 were used in the experiments. Root mean square error (RMSE) and Pearson correlation coefficient were used as the error measure in the experiments. The low values of RMSE and the high values of Pearson correlation coefficient are desirable.

Linear regression, Random forests regression and Support vector machines regression from the Weka machine learning tool [16] were used in the experiments. Default values of the parameters were used.

IV. Results and Discussion

Results for Experiment-1 and Experiment-2 are presented in Table 3 and Table 4. Table 3 shows the RMSE of Experiment-1 and Experiment-2, whereas Table 4 presents the Pearson correlation coefficient of these experiments. RMSE results suggest that except for Canada, RMSE values for Experiment-1 are less than those of Experiment-2. Pearson correlation coefficients are more in Experiment-1 than those of Experiment-2 with an exception for the UAE. As data collected during the vaccination period and the period without vaccination have different characteristics, it was expected that forecast of the number of Covid-19 cases in vaccination period using the data of the period without vaccination may not be as accurate as forecast of the number of Covid-19 cases in the period of without vaccination. The size of the training data is more for the prediction in the vaccination period. The larger size of the training data creates better model. That could be the reason for better prediction in the vaccination period.

RMSE results for Experiment-3 are presented in Table 5. Table 6 has Pearson correlation coefficient results of Experiment-3. Results of Experiment-1 and Experiment-3 are quite similar. In Experiment-3, only the data for vaccination period is used for the prediction. The size of the training data for Experiment-3 is less as compared to that of Experiment-1. However, models of Experiment-3 are as accurate as models of Experiment-1. Data for vaccination period and data without vaccination period have different characteristics. For the correct forecast of Covid-18 cases, it is important to understand this difference. We must employ different models trained on different time periods including one that is trained only on the data collected during vaccination period to get the best prediction models.

Country	Days without vaccination	N(days)	Days with vaccination	V(days)
Canada	From 26-01-2020 to 14-12-2020	324	From 15-12-2020 to 24-05-2021	161
Germany	From 27-01-2020 to 27-12-2020	336	From 28-12-2020 to 24-05-2021	148
UAE	From 29-01-2020 to 06-01-2021	344	From 07-01-2021 to 24-05-2021	138
USA	From 23-01-2020 to 20-12-2020	333	From 21-12-2020 to 24-05-2021	155

TABLE II: Datasets used in the experiments.

Country	Linear regression		Random forests regression		Support vector machine regression	
	Experiment-1	Experiment-2	Experiment-1	Experiment-2	Experiment-1	Experiment-2
Canada	1796	966	2193	1395	1317	1177
Germany	5101	10958	7679	12387	4688	10750
UAE	115	257	191	795	102	242
USA	6672	20748	5901	119737	6428	21502

TABLE III: RMSE of Experiment-1 and Experiment-2 for different countries.

Country	Linear regression		Random forests regression		Support vector machine regression	
	Experiment-1	Experiment-2	Experiment-1	Experiment-2	Experiment-1	Experiment-2
Canada	0.8827	-0.2067	0.8644	0.5253	0.8522	-0.5731
Germany	0.5985	0.4439	0.6665	0.4802	0.6159	0.4337
UAE	0.6363	0.7407	0.3133	-0.1733	0.7014	0.7417
USA	0.7826	0.5691	0.7562	0.6595	0.7191	0.5428

TABLE IV: Pearson correlation coefficient of Experiment-1 and Experiment-2 for different countries.

Country	Linear regression	Random forests regression	Support vector machine regression
Canada	1596	1149	1237
Germany	4422	3664	3827
UAE	113	116	98
USA	4701	8008	44925

TABLE V: RMSE of Experiment-3 for different countries

.

V. Conclusion

Covid-19 is a very serious public health problem. It is important to forecast the number of Covid-19 cases to prepare for the future. Various prediction models have been proposed for this purpose. Most of these models were trained on the data collected during the period without vaccination. In this paper, we studied the effect of the data collected during the vaccination period on the prediction accuracy. By using various regression methods, we showed that prediction results are different for these two periods. Therefore, it is important to understand that data from the vaccination period and data from the period without vaccination have different characteristics. Results suggest that models trained with data only from the vaccination period should be tested for accurate prediction.

In future, we will use more regression methods in the experiments. In this paper, data from four countries are used. Many countries have started vaccination against Covid-19. Data from more countries will be used in the future experiments.

References

[1] Naming the coronavirus disease (covid-19) and the virus that causes it. World Health Organization. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it. Accessed 26^{th} May 2020.

[2] Novel coronavirus in China. World Health Organization. https://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/. Accessed 26^{th} May 2020.

[3] Our World in Data. https://ourworldindata.org/grapher/share-people-fully-vaccinated-covid. Accessed 31^{th} May 2020.

[4] WHO Director-General's opening remarks at the media briefing on Covid-19. World Health Organization (press release). 11 march 2020. https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-marc Accessed 26^{th} May 2020.

[5] Mohamad Mahmoud Al Rahhal Fahad Raddah H. Albogamy Eslam Al Maghayreh Hussain AlSalman Abdu Gumaei, Mabrook Al-Rakhami. Prediction of covid-19 confirmed cases using gradient boosting regression method. *Computers, Materials & Continua*, 66(1):315–329, 2021.

[6] Amir Ahmad, Sunita Garhwal, Santosh Kumar Ray, Gagan Kumar, Sharaf Jameel Malebary, and Omar Mohammed Omar Barukab. The number of confirmed cases of covid-19 by using machine learning: Methods and challenges. *CoRR*, abs/2006.09184, 2020.

[7] M. Batista. Estimation of the final size of the covid-19 epidemic. *medRxiv*, 2020.

[8] Ian Cooper, Argha Mondal, and Chris G. Antonopoulos. A sir model assumption for the spread of covid-19 in different communities. *Chaos, Solitons and Fractals*, 139:110057, 2020.

[9] Lawrence Corey, John R. Mascola, Anthony S. Fauci, and Francis S. Collins. A strategic approach to covid-19 vaccine r&d. *Science*, 368(6494):948–950, 2020.

[10] A. D. Dubey. Twitter sentiment analysis during covid19 outbreak, 2020.

[11] S. J. Fong, N. Dey G. Li, R. G. Crespo, and E. Herrera-Viedma. Finding an accurate early forecasting model from small dataset: A case of 2019-ncov novel coronavirus outbreak. *International Journal of Interactive Multimedia and Artificial Intelligence*, 6:132–139, 2020.

[12] Guido Forni, Alberto Mantovani, Guido Forni, Alberto Mantovani, Lorenzo Moretta, Rino Rappuoli, Giovanni Rezza, Arnaldo Bagnasco, Giuseppina Barsacchi, Giovanni Bussolati, Massimo Cacciari, Pietro Cappuccinelli, Enzo Cheli, Renato Guarini, Massimo Livi Bacci, Marco Mancini, Cristina Marcuzzo, Maria Concetta Morrone, Giorgio Parisi, Gianfranco Pasquino, Carlo Patrono, Alberto Quadrio Curzio, Giuseppe Remuzzi, Alessando Roncaglia, Stefano Schiaffino, and Paolo Vineis. Covid-19 vaccines: where we stand and challenges ahead. *Cell Death and Differentiation*, 28(2), February 2021.

Country	Linear regression	Random forests regression	Support vector machine regression
Canada	0.7616	0.8452	0.8596
Germany	0.5978	0.7123	0.5957
UAE	0.6796	0.5285	0.6745
USA	0.815	0.9323	0.831

TABLE VI: Pearson correlation coefficient of Experiment-3 for different countries.

[13] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[14] Chaolin Gu, Jie Zhu, Yifei Sun, Kai Zhou, and Jiang Gu. The inflection point about covid-19 may have passed. *Science Bulletin*, 2020.

[15] R. Gupta, G. Pandey, P. Chaudhary, and S. K. Pal. Seir and regression model based covid-19 outbreak predictions in india. *medRxiv*, 2020.

[16] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18, 2009.

[17] Zixin Hu, Qiyang Ge, Shudi Li, Li Jin, and Momiao Xiong. Artificial intelligence forecasting of covid-19 in china, 2020.

[18] K. Jahanbin and V. Rahmanian. Using twitter and web news mining to predict covid-19 outbreak. *Asian Pacific Journal of Tropical Medicine*, March 2020.

[19] Firuz Kamalov, Aswani Cherukuri, Hana Sulieman, Fadi Thabtah, and Akbar Hossain. Machine learning applications for covid-19: A state-of-the-art review, 2021.

[20] Lixiang Li, Zihang Yang, Zhongkai Dang, Cui Meng, Jingze Huang, Haotian Meng, Deyu Wang, Guanhua Chen, Jiaxuan Zhang, Haipeng Peng, and Yiming Shao. Propagation analysis and prediction of the covid-19. *Infectious Disease Modelling*, 5:282 – 292, 2020.

[21] Ming Li, Jie Chen, and Youjin Deng. Scaling features in the spreading of covid-19, 2020.

[22] Dianbo Liu, Leonardo Clemente, Canelle Poirier, Xiyu Ding, Matteo Chinazzi, Jessica T Davis, Alessandro Vespignani, and Mauricio Santillana. A machine learning methodology for real-time forecasting of the 2019-2020 covid-19 outbreak using internet searches, news alerts, and estimates from mechanistic models, 2020.

[23] Mohd Hafiz Mohd and Fatima Sulayman. Unravelling the myths of r0 in controlling the dynamics of covid-19 outbreak: A modelling perspective. *Chaos, Solitons and Fractals*, 138:109943, 2020.

[24] Bohdan M. Pavlyshenko. Regression approach for modeling covid-19 spread and its impact on stock market, 2020.

[25] L. Qin, Q. Sun, Y. Wang, K. Wu, M. Chen, B. Shia, and S. Wu. Prediction of number of cases of 2019 novel coronavirus (covid-19) using social media search index. *International Journal of Environmental Research and Public Health*, 17(7), 2020.

[26] Diana M. Thomas, Rodney Sturdivant, Nikhil V. Dhurandhar, Swati Debroy, and Nicholas Clark. A primer on covid-19 mathematical models. *Obesity*, 28(8):1375–1377, 2020.

[27] A. Tomar and N. Gupta. Prediction for the spread of covid-19 in india and effectiveness of preventive measures. *Science of The Total Environment*, 728:138762, 2020.

[28] Jia Wangping, Han Ke, Song Yang, Cao Wenzhe, Wang Shengshu, Yang Shanshan, Wang Jianwei, Kou Fuyin, Tai Penggang, Li Jing, Liu Miao, and He Yao. Extended sir prediction of the epidemics trend of covid-19 in italy and compared with hunan, china. *Frontiers in Medicine*, 7:169, 2020.