International Journal of Educational Excellence
(2020) Vol. 6, No. 1, 71-94
ISSN 2373-5929
DOI: 10.18562/IJEE.054

Artificial Intelligence Forecasting of Covid-19 in China

Zixin Hu [a], Qiyang Ge [b], Shudi Li [c], & Momiao Xiong [d]

Received: 17 February 2020 • Accepted: 03 March 2020

Abstract: Background: An alternative to epidemiological models for transmission dynamics of Covid-19 in China, we propose the artificial intelligence (AI)-inspired methods for real-time forecasting of Covid-19 to estimate the size, lengths and ending time of Covid-19 across China. Methods: We developed a modified stacked auto-encoder for modeling the transmission dynamics of the epidemics. We applied this model to real-time forecasting the confirmed cases of Covid-19 across China. The data were collected from January 11 to February 27, 2020 by WHO. We used the latent variables in the auto-encoder and clustering algorithms to group the provinces/cities for investigating the transmission structure. Results: We forecasted curves of cumulative confirmed cases of Covid-19 across China from Jan 20, 2020 to April 20, 2020. Using the multiple-step forecasting, the estimated average errors of 6-step, 7-step, 8-step, 9-step and 10-step forecasting were 1.64%, 2.27%, 2.14%, 2.08%, 0.73%, respectively. We predicted that the time points of the provinces/cities entering the plateau of the forecasted transmission dynamic curves varied, ranging from Jan 21 to April 19, 2020. The 34 provinces/cities were grouped into 9 clusters. Conclusions: The accuracy of the AI-based methods for forecasting the trajectory of Covid-19 was high. We predicted that the epidemics of Covid-19 will be over by the middle of April. If the data are reliable and there are no second transmissions, we can accurately forecast the transmission dynamics of the Covid-19 across the provinces/cities in China. The AI-inspired methods are a powerful tool for helping public health planning.

Key-words: COVID-19, Artificial Intelligence, Transmission Dynamics, Forecasting, Time Series, Auto-Encoder.

---

[a] The School of Life Sciences, Fudan University, Shanghai (China) ORCID 0000-0002-2359-744X, [b] Human Phenome Institute, Fudan University, Shanghai (China), [c] The School of Mathematic Sciences, Fudan University, Shanghai (China, [d] The School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030 (USA) ORCID 0000-0003-0635-5796. Correspondence: Dr. Momiao Xiong, Department of Biostatistics and Data Science, School of Public Health, The University of Texas Health Science Center at Houston, P.O. Box 20186, Houston, Texas 77225 (USA). Momiao.Xiong@uth.tmc.edu

1. Introduction

In the beginning of December, 2019, Covid-19 virus that slipped from animals to humans in Wuhan city, China caused an outbreak of respiratory illness. A number of the statistical, dynamic and mathematical models of the Covid-19 outbreak including the SEIR model have been developed to analyze its transmission dynamics (Li, et al., 2020; Wu, Leung, & Leung, 2020; Zhao, et al., 2020; Kucharski, Russell, Diamond,, & Liu; 2020; Tuite, & Fisman, 2020). Although these epidemiological models are useful for estimating the dynamics of transmission, targeting resources and evaluating the impact of intervention strategies, the models require parameters and depend on many assumptions. Unlike system identification in engineering where the parameters in the models are estimated using real data, at the outbreak, estimated parameters using real time data are not readily available (Funk, Camacho, Kucharski, Eggo, y Edmunds, 2018; Johansson et al., 2019). Most analyses used hypothesized parameters and hence do not fit the data very well. The accuracy of forecasting the future cases of Covid-19 using these models may not be very high. Timely interventions are needed to control the serious impacts of Covid-19 on health.

To overcome limitations of the epidemiological model approach, and assist public health planning and policy making, we develop an AI based method for real time forecasting of the new and cumulative confirmed cases of Covid-19 in total and provinces/ cities across China. We also forecast the possible trend and plateau of Covid-19 transmission in China and group the provinces/cities into clusters according to the dynamic patterns of Covid-19 transmission. The analysis is based on the surveillance data of the confirmed Covid-19 cases in China up to February 18, 2020.

2. Methods

2.1. Data Sources

Data on the confirmed cases of Covid-19 from January 11, 2020 to January 20, 2020, and from January 21, 2020 to February 27, 2020, were from Surging News Network (https://www.thepaper.cn/) and WHO (https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports), respectively. WHO took the lab confirmed case as the confirmed cases from January 21, 2020 to February 13, 2020 and took the sum of the number of the clinical confirmed and lab conformed cases as the number of the confirmed cases after February 14, 2020. The numbers of the clinical confirmed and the lab confirmed cases in Hubei Province, China on February 14, was 15,384 and 36,602, respectively. Therefore, the number of the confirmed

cases in Hubei Province before February 14, 2020 was adjusted by the formula:

the number of the confirmed cases = the number of the lab confirmed cases $\times \frac{(15384+36602)}{36602}$.

Data included the total numbers of the accumulated and new confirmed cases in all of China and the numbers of the accumulated and new confirmed cases across 31 Provinces/Cities in mainland China and three other regions (Hong Kong, Macau and Taiwan) in China. The data were organized in a matrix with the rows representing the whole China and province/city and columns representing the number of the new confirmed cases of each day.

The confirmed cases of each province/city were a time series. Let t_{ij} be the number of the confirmed cases of the j^{th} day within the i^{th} province/city. Let Z be a $34 \times m$ dimensional matrix. The element Z_{ij} is the number of the confirmed new cases of Covid-19 on the j^{th} day, starting with January 11, 2020 in the i^{th} city.

2.2. Modified Auto-encoder for Modeling Time Series

Modified auto-encoders, MAE) (Charte, Charte, García, Jesus, & Herrera, 2018; Yuan, Huang, Wang, Yang, & Gui, 2018) were used to forecast the number of the accumulative and new confirmed cases of Covid-19. Unlike the classical auto-encoder where the number of nodes in the layers usually decreases from the input layer to the latent layers, the numbers of the nodes in the input, the first latent layer, the second latent layer and output layers in the MAE were 8, 32, 4 and 1, respectively (Figure 1). We view a segment of time series with 8 days as a sample of data and take 128 segments of time series as the training samples. One element from the dada matrix Z is randomly selected as a start day of the segment and select its 7 successive days as the other days to form a segment of time series. Let i be the index of the segment and j_i be the column index of the matrix Z that was selected as the starting day. The i^{th} segment time series can be represented as $\{Z_{j_i}, Z_{j_i+1}, \ldots, Z_{j_i+7}\}$. Data were normalized to $X_{j_i+k} = \frac{Z_{j_i+k}}{S}, k = 0, 1, \ldots, 7$, where $S = \frac{1}{8}\sum_{k=0}^{7} Z_{j_i+k}$. Let $Y_i = \frac{Z_{j_i+8}}{S}$ be the normalized number of cases to forecast. If $S = 0,$ then set $Y_i = 0$. The loss function was defined as

$$L = \sum_{i=1}^{128} W_i (Y_i - \hat{Y}_i)^2,$$

where Y_i was the observed number of the cases in the forecasting day of the i^{th} segment time series and $\hat{Y}_i$ was its forecasted number of cases by the MAE, and W_i were weights. If j_i was in the interval [1, 12], then $W_i = 1$. If j_i was in the interval [13, 24], then $W_i = 2$, etc. The back propagation algorithm was used to estimate the weights and bias in the MAE. Repeat training processes 5 times. The average forecasting $\hat{Y}_i, i = 1, \ldots, 34$ will be taken as a final forecasted number of the accumulated confirmed cases for each province/city.

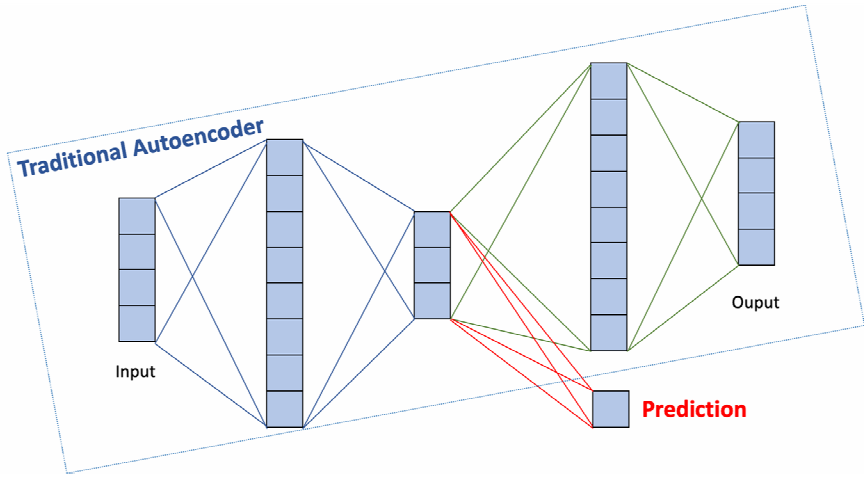


Figure 1. *Architecture of a MAE.*

2.3. Forecasting Procedures

The trained MAE was used for forecasting the future number of the confirmed cases of Covid-19 for each province/city. Consider the i^{th} province/city. Assume that the number of new confirmed cases of Covid-19 on the j^{th} day that needs to be forecasted is x_{ij}. Let H be a 34×8 dimensional matrix and $h_{il} = x_{ij-9+l}, i = 1, \ldots, 34, \text{ and } l = 1, \ldots, 8$. Let $g_i = \frac{1}{8}\sum_{l=1}^{8} h_{il}, i = 1, \ldots, 34$ be the average of the i^{th} row of the matrix H. Let U be the normalized matrix of H where $u_{il} = \frac{h_{il}}{g_i}, i = 1, \ldots, 34, \text{and } l = 1, \ldots, 8$. The output of the MAE is the forecasted number of the new confirmed cases and is denoted as

$\hat{v}_i = f(u_{i1}, \ldots, u_{i8}, \theta), i = 1, \ldots, 34$, where θ represented the estimated parameters in the trained MAE. The one-step forecasting of the number of the new confirmed cases of Covid-19 for each city is given by $\hat{Y}_i = \hat{v}_i g_i, i = 1, \ldots, 34$.

The recursive multiple-step forecasting involved using a one-step model multiple times where the prediction for the preceding time step was used as an input for making a prediction on the following time step. For example, for forecasting the number of the new confirmed cases for the one more next day, the predicted number of new cases in one-step forecasting would be used as an observational input in order to predict day 2. Repeat the above process to obtain the two-step forecasting. The summation of the final forecasted number of the new confirmed cases for each province/city was taken as the prediction of the total number of the new confirmed cases of Covid-19 in China.

2.4. Clustering

The values of the latent variables in the second latent layer of the MAE for each province/city were extracted. For each province/city, a 34×4 dimensional latent matrix A were formed. The largest single value λ of the latent matrix A was obtained via single value decomposition. We performed five-time trainings and obtained five largest single values. For each province/city, we formed a feature vector that consisted of the five largest single values λs, the starting day and the forecasted end day of the Covid-19 outbreak, the day, the number of new confirmed cases reaching the maximum, the largest number of the forecasted new confirmed cases and the number of the forecasted accumulated confirmed cases of Covid-19 in the respective province/city. The k-means algorithms were performed on the 34 feature vectors to group provinces/cities into clusters.

5. Results

Figure 2 plotted the total number curves of the reported and forecasted cumulative and new confirmed cases of Covid-19 in China as a function of days. The reported cases were from January 11, 2020 to February 27, 2020. A total number of 47 days' data were available. We began to forecast on February 20, 2020. Figure 2 showed that the forecasting curve was close to the reported curve. From Figure 2 it can be observed that the curve of the new confirmed cases reached the 5,236 peak on February 5, 2020 and decreased to zero on April 20 (forecasting). The potential cumulative confirmed cases of Covid-19 in China was observed to reach the plateau (83,401) on April 20, 2020. Figure S1 plotted the national reported and fitted curves of the

cumulative confirmed cases in China from January 20, 2020 to February 27, 2020. To further examine the accuracy of forecasting, Table 1 where the data from January 20, 2020 to February 27, 2020 were used to fit the MAE model. In the table, we listed 1 day-step to 10 day-step forecasting errors, respectively, i.e., the errors of using the current reported cases to forecast future s day cases. Table 1 showed that the average errors of the forecasting did not strictly increase as the number-steps for forecasting increased due to fluctuations of the data. Forecasting accuracy was very high.

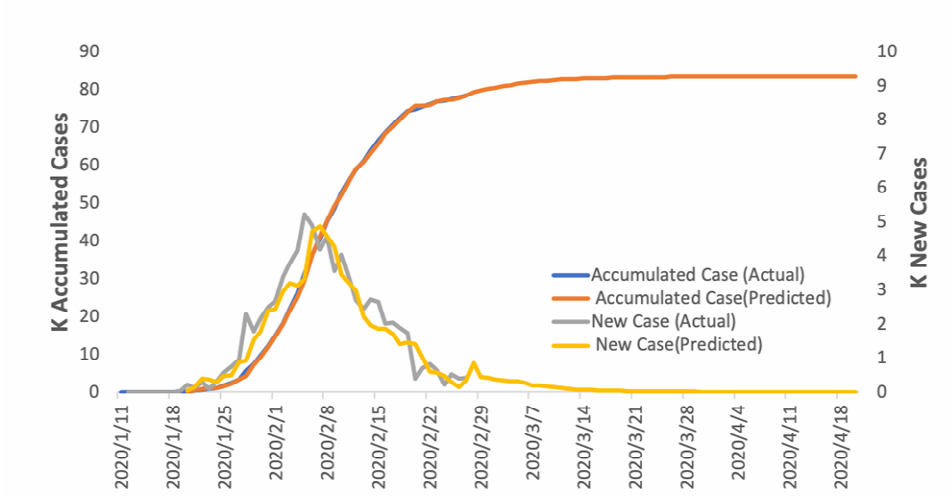


Figure 2. *The national reported and forecasted curves of the cumulative and new confirmed cases of Covid-19 in China as a function of days from January 11, 2020 to April 20, 2020.*

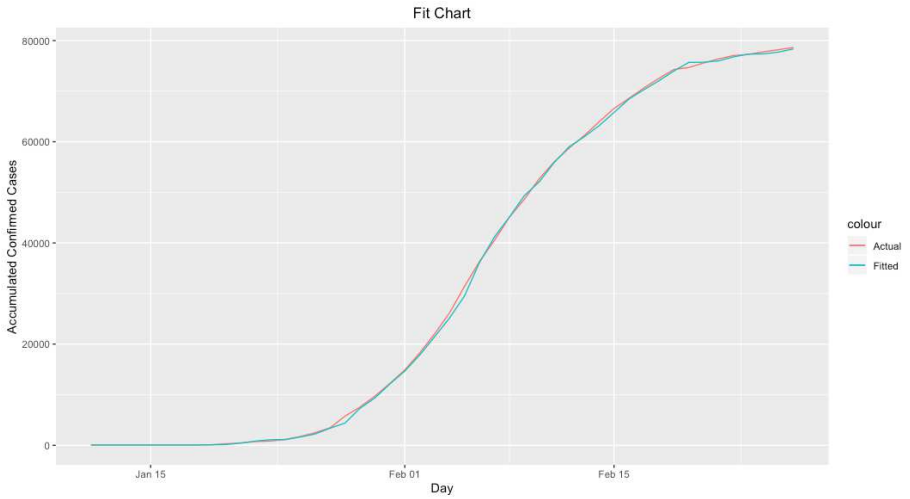


Figure S1. *The national reported and fitted curves of the cumulative confirmed cases of Covid-19 in China from January 11, 2020 to February 27, 2020, where the red curve was the reported and green curve was the fitted.*

Figure 3 presented the forecasted curves of cumulative confirmed cases of Covid-19 across 31 province/cities in mainland China and three other regions (Hong Kong, Macau and Taiwan) in China as a function of days from January 11, 2020 to April 20, 2020. We observed from Figure 3 that the times of different provinces/cities entering the plateau will be different, ranging from Jan 21, 2020 to April 20, 2020. Xizang first entered the plateau at Jan 21 followed by Macao on Jan 26 and Qinghai on Jan 28. Hubei province will be one of the last to enter the plateau (70,019) around April 20, 2020, most provinces/cities will enter the plateau around middle of March. We can also observe the different shapes of the curves among the provinces/cities, which imply different dynamic patterns of the transmission of Covid-19 across the provinces/cities.

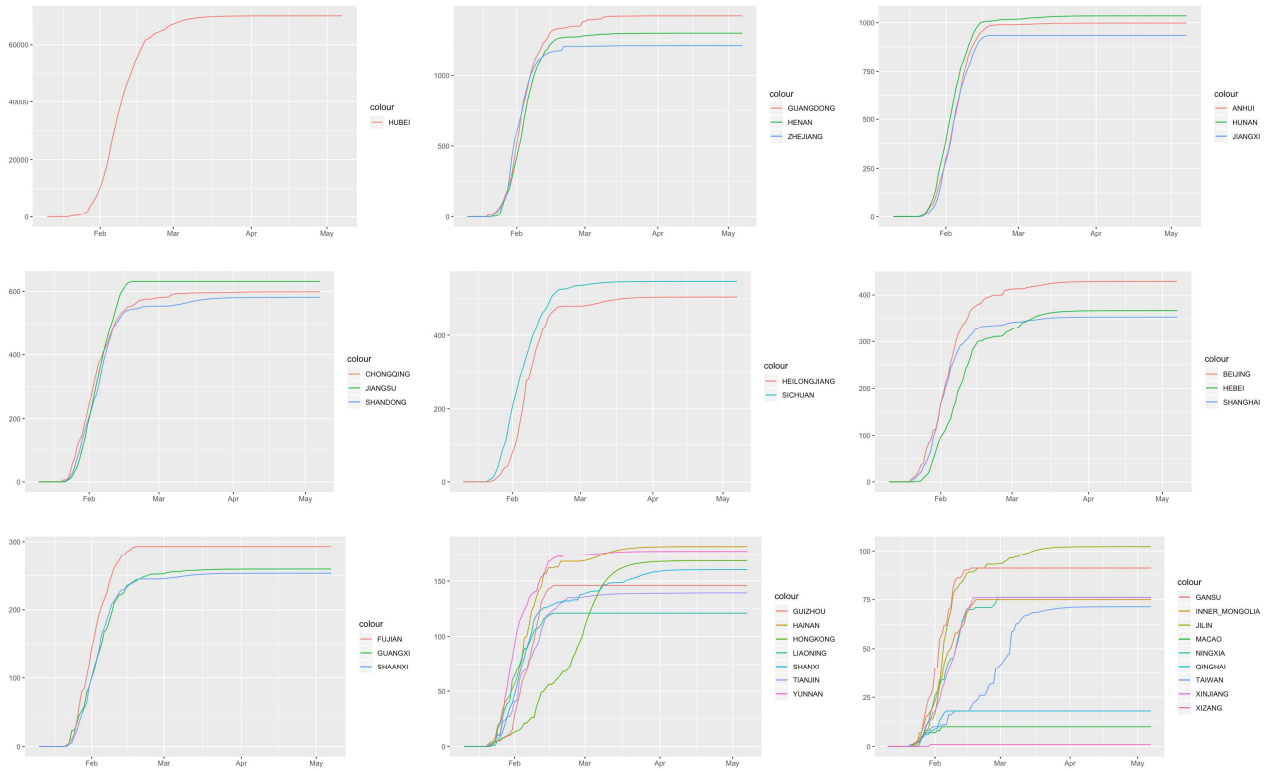
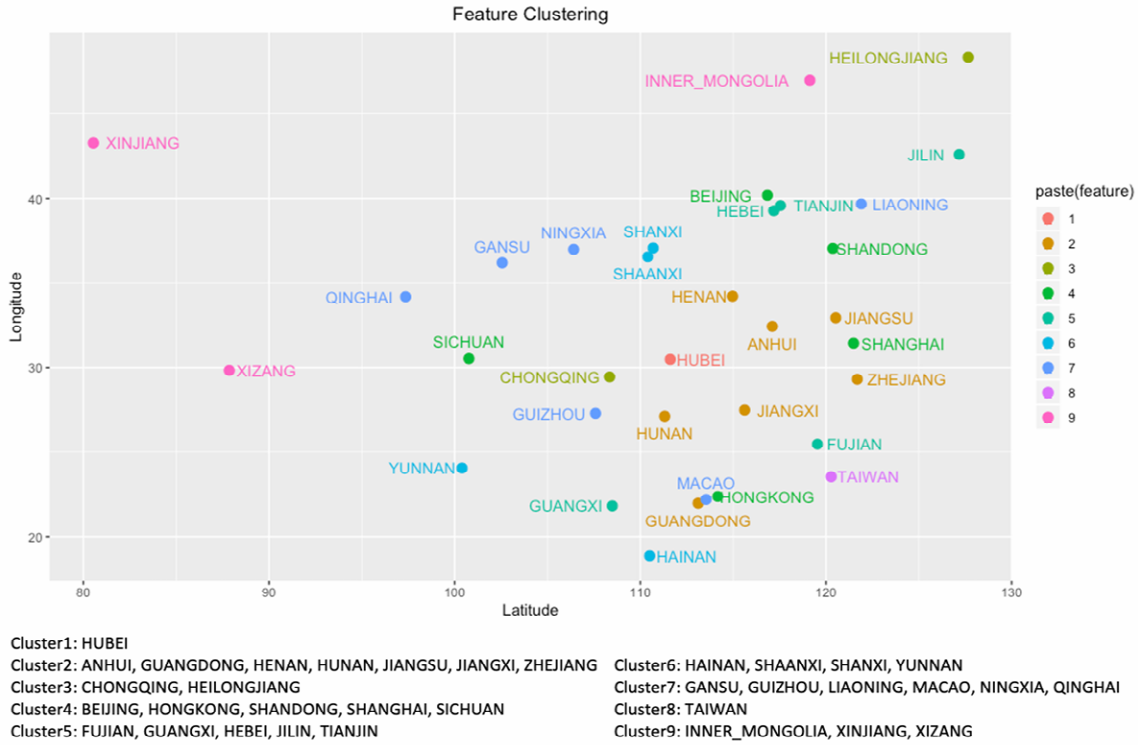


Figure 3. *The forecasted curves of the cumulative confirmed cases of Covid-19 across 34 province/cities in China as a function of days from January 11, 2020 to April 20, 2020.*

Many factors such as size of outflow population from Wuhan to each affected provinces/cities, interventions, geographic locations, economic and social activities, environmental heterogeneity and healthcare facility affect

disease transmission dynamics across the country. Clustering its temporal dynamics will provide numerous insights on patterns of propagation of Covid-19. To further capture the dynamic pattern of Covid-19 spread across the provinces/cities, we presented Figure 4 where 34 provinces/cities formed nine clusters. Since the spread of Covid-19 was influenced by the pattern of contacts between individuals, clusters partially showed geographic structure.



Cluster1: HUBEI
Cluster2: ANHUI, GUANGDONG, HENAN, HUNAN, JIANGSU, JIANGXI, ZHEJIANG
Cluster3: CHONGQING, HEILONGJIANG
Cluster4: BEIJING, HONGKONG, SHANDONG, SHANGHAI, SICHUAN
Cluster5: FUJIAN, GUANGXI, HEBEI, JILIN, TIANJIN
Cluster6: HAINAN, SHAANXI, SHANXI, YUNNAN
Cluster7: GANSU, GUIZHOU, LIAONING, MACAO, NINGXIA, QINGHAI
Cluster8: TAIWAN
Cluster9: INNER_MONGOLIA, XINJIANG, XIZANG

Figure 3. *The clusters that were grouped by features extracted from the MAE and the cumulative confirmed case time series of Covid-19 across 31 provinces/cities in mainland China and three other regions in China formed 9 clusters.*

For example, Hubei province, a major source of Covid-19, formed a cluster. Followed by the Anhui, Guangdong, Henan, Hunan, Jiangsu, Jiangxi and Zhejiang, which are in areas surrounding pathogen sources (Hubei province). The Inner Mongolia, Xizang and Qinghai cluster had the lowest number of cases and are in areas far away from the sources of Covid-19. In addition to the geographic locations that affected the transmission of Covid-19, the healthcare resources and economic and social activities may also affect the transmission dynamics of Covid-19 as the cluster of Shanghai, Hongkong, Beijing, Shandong and Sichuan. And Fujian Guangxi, Hebei, Jilin and Tianjin formed a cluster, but these five provinces/cities are not located in the same

geographic area. They may have similar economic relationships with Wuhan, healthcare resources and take similar interventions to control of the spread of Covid-19.

6. Discussion

As an alternative to epidemiologic transmission model, we used MAE to forecast the real-time trajectory of the transmission dynamics and generate the real-time forecasts of Covid-19 across the provinces/cities in China. The results showed that the accuracies of prediction and subsequently multiple-step forecasting were high. Our experience revealed that forecasting improves when the training time was longer. We estimated the potential time points of decreasing growth of new confirmed case curves across 34 provinces/cities, the lengths of the Cov-19 epidemics across China, and the times when the number of accumulated confirmed cases of Covid-19 would reach the plateau of their accumulate case curves. If the data are reliable and there will be no second transmission, the MAE models predicted that the Covid-19 outbreak in China might be over in the middle of April. In this study, we used cluster analysis to group 34 provinces/cities into 9 clusters and explore their geographic and healthcare resource structures.

The MAE models allow inputting the interventions information and investigating the impact of interventions on the size of the virus outbreak and end time of the virus outbreak. However, we have not explored such functions of the MAE due to lack of data. Similar to epidemiologic transmission dynamic models, the MAE can also be used for simulations. The trained MAE can well approximate many dynamic processes. Using the hypothesized initial sizes of the epidemic outbreak, we can use the MAE with known parameters and architecture to estimate the sizes of outbreak in the future and simulate the impact of the interventions on the sizes and severity of the epidemics. Complimentary to a model approach to transmission dynamics of virus outbreaks, the data driven AI-based methods provide real time forecasting tools for tracking, estimating the trajectory of epidemics, assessing their severity, predicting the lengths of epidemics and assisting government and health workers to make plan and good decisions.

Appendix: Table 1.

Date	Actual	1-step prediction	1-step error	2-step error	3-step error	4-step error	5-step error	6-step error	7-step error	8-step error	9-step error	10-step error
18/02/2020	72.528	71.757	-1,06%									
19/02/2020	74.280	74.005	-0,37%	-1,34%								
20/02/2020	74.675	75.564	1,19%	1,31%	-0,06%							
21/02/2020	75.569	76.685	1,48%	1,36%	1,49%	0,12%						
22/02/2020	76.392	76.305	-0,11%	1,49%	1,41%	1,81%	0,15%					
23/02/2020	77.042	77.827	1,02%	0,02%	1,58%	1,39%	2,30%	0,13%				
24/02/2020	77.262	79.837	3,33%	2,18%	0,83%	2,53%	2,46%	3,30%	1,08%			
25/02/2020	77.780	79.155	1,77%	3,01%	2,14%	0,47%	2,28%	2,36%	3,15%	0,76%		
26/02/2020	78.191	77.957	-0,30%	1,61%	2,70%	2,15%	0,28%	2,20%	2,49%	3,21%	0,80%	
27/02/2020	78.630	78.646	0,02%	-0,53%	1,49%	2,47%	2,22%	0,21%	2,37%	2,46%	3,37%	0,73%
Average Absolute Error			1,07%	1,43%	1,46%	1,56%	1,62%	1,64%	2,27%	2,14%	2,08%	0,73%

Table 1. *Errors of forecasting the national cumulative confirmed cases in China.*

References

Charte, D., Charte, F., García, S., del Jesus, M. J., & Herrera, F. (2018). A practical tutorial on autoencoders for nonlinear feature fusion: Taxonomy, models, software and guidelines. *Information Fusion*, 44, 78-96.

Funk, S., Camacho, A., Kucharski, A. J., Eggo, R. M., & Edmunds, W. J. (2018). Real-time forecasting of infectious disease dynamics with a stochastic semi-mechanistic model. *Epidemics*, 22, 56-61.

Johansson, M. A., Apfeldorf, K. M., Dobson, S., Devita, J., Buczak, A. L., Baugher, B. et al. (2019). An open challenge to advance probabilistic forecasting for dengue epidemics. *Proceedings of the National Academy of Sciences*, 116(48), 24268-24274.

Kucharski, A., Russell, T., Diamond, C., Liu, Y, (2020). CMMID nCoV working group. In J. Edmunds, S. Funk, & R. Eggo (ed.). *Analysis and projections of transmission dynamics of nCoV in Wuhan*. Retrieved from https://cmmid.github.io/ncov/wuhan_early_dynamics/index.html.

Li, Q., Guan, X., Wu, P., *et al.* (2020) Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*, 2020 Jan 29. [Epub ahead of print] doi: 10.1056/NEJMoa2001316.

Tuite, A. R., & Fisman, D. N. (2020). Reporting, epidemic growth, and reproduction numbers for the 2019 novel coronavirus (2019-nCoV) epidemic. *Annals of Internal Medicine*, 2020 Feb 5. [Epub ahead of print] doi: 10.7326/M20-0358..

Wu, J.T., Leung, K., Leung, G.M. (2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *The Lancet*. 2020 Jan 31. [Epub ahead of print] pii:S0140-6736(20)30260-9. doi: 10.1016/S0140-6736(20)30260-9.

Yuan, X., Huang, B., Wang, Y., Yang, C., & Gui, W. (2018). Deep learning-based feature representation and its application for soft sensor modeling with variable-wise weighted SAE. *IEEE Transactions on Industrial Informatics*, 14(7), 3235-3243.

Zhao, S., Musa, S. S., Lin, Q., Ran, J., Yang, G., Wang, W, et. al. (2020). Estimating the unreported number of novel coronavirus (2019-nCoV) cases in China in the first half of January 2020: a data-driven Modelling analysis of the early outbreak. *Journal of Clinical Medicine*, 9(2), 388.

International Journal of Educational Excellence
(2020) Vol. 6, No. 1, 71-94
ISSN 2373-5929
DOI: 10.18562/IJEE.054

Pronósticos con Inteligencia Artificial para COVID-19 en China

Zixin Hu [a], Qiyang Ge [b], Shudi Li [c], & Momiao Xiong [d]

Received: 17 February 2020 • Accepted: 03 March 2020

Resumen: Antecedentes: Como alternativa a los modelos epidemiológicos para la dinámica de transmisión del COVID-19 en China, proponemos métodos basados en la inteligencia artificial (IA) para la predicción a tiempo real del COVID-19 y con el fin de estimar el tamaño, la duración y conclusión del COVID-19 en China. Métodos: Desarrollamos un auto-codificador apilado modificado para modelar la dinámica de transmisión de las epidemias. Aplicamos este modelo para pronosticar en tiempo real los casos confirmados de COVID-19. Los datos fueron los recopilados del 11 de enero al 27 de febrero de 2020 por la WHO. Usamos las variables latentes en los algoritmos de autocodificador y clustering para agrupar las provincias con el objetivo de investigar la estructura de transmisión. Resultados: llevamos a cabo un pronóstico mediante curvas de casos confirmados y acumulativos de COVID-19 desde el 20/01/2020 hasta el 20/04/2020. Los errores promedio estimados en estos pronósticos, de 6, 7, 8, 9 y 10 pasos, fueron 1,64%, 2,27%, 2,14%, 2,08% y 0,73%, respectivamente. Predijimos que los puntos temporales de las provincias que entraban en la meseta de las curvas dinámicas de transmisión pronosticadas variarían desde el 21/01/20 al 19/04/2020. Las 34 provincias fueron agrupadas en 9 grupos. Conclusiones: La precisión de los métodos basados en la IA para pronosticar la trayectoria de COVID-19 fue alta. Estimamos que la epidemia terminará a mediados de abril. Si los datos son fiables y no hay segundas transmisiones, podemos pronosticar de manera precisa la dinámica de transmisión de COVID-19 a través de las provincias chinas. Los métodos inspirados en la IA resultan una poderosa herramienta para ayudar en la planificación de la salud pública.

Plabras Clave: COVID-19, Inteligencia Artificial, Dinámica de Transmisión, Previsión, Series Temporales, Auto-Encoder.

---

[a] The School of Life Sciences, Fudan University, Shanghai (China) ORCID 0000-0002-2359-744X, [b] Human Phenome Institute, Fudan University, Shanghai (China), [c] The School of Mathematic Sciences, Fudan University, Shanghai (China, [d] The School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030 (USA) ORCID 0000-0003-0635-5796. Correspondence: Dr. Momiao Xiong, Department of Biostatistics and Data Science, School of Public Health, The University of Texas Health Science Center at Houston, P.O. Box 20186, Houston, Texas 77225 (USA). Momiao.Xiong@uth.tmc.edu

1. Introducción

A principios de diciembre de 2019, el virus Covid-19 pasó de los animales a los seres humanos en la ciudad de Wuhan, China, causando una enfermedad respiratoria. Se han desarrollado varios modelos estadísticos, dinámicos y matemáticos de este brote de Covid-19, incluido el modelo SEIR, para analizar su dinámica de transmisión (Li, et al., 2020; Wu, Leung, & Leung, 2020; Zhao, et al., 2020; Kucharski, Russell, Diamond, & Liu; 2020; Tuite, & Fisman, 2020). Aunque estos modelos epidemiológicos son útiles para estimar la dinámica de la transmisión, orientar los recursos y evaluar el impacto de las estrategias de intervención, los modelos requieren parámetros y dependen de muchos supuestos. Los parámetros estimados utilizando datos en tiempo real, en el estallido del brote, no están disponibles fácilmente a diferencia de la identificación del sistema en ingeniería donde los parámetros de los modelos se estiman utilizando datos reales, al estallar, (Funk, Camacho, Kucharski, Eggo, y Edmunds, 2018; Johansson et al., 2019). La mayoría de los análisis utilizaron parámetros hipotéticos y, por lo tanto, no se ajustan muy bien a los datos. La precisión de la predicción de los casos futuros de Covid-19 utilizando estos modelos puede no ser muy alta. Se necesitan intervenciones oportunas para controlar los graves impactos de Covid-19 en la salud.

Para superar las limitaciones del enfoque modelo epidemiológico y ayudar a la planificación de la salud pública y la formulación de políticas, desarrollamos un método basado en la IA para la previsión en tiempo real de los nuevos y acumulados casos confirmados de Covid-19 en total y provincias / ciudades de toda China. También pronosticamos la posible tendencia y meseta de la transmisión de Covid-19 en China y agrupamos las provincias/ciudades en clusters según las pautas dinámicas de la transmisión de Covid-19. El análisis se basa en los datos de vigilancia de los casos confirmados de Covid-19 en China hasta el 18 de febrero de 2020.

2. Métodos

2.1. Fuentes de datos

Los datos sobre los casos confirmados de Covid-19 del 11 de enero de 2020 al 20 de enero de 2020 y del 21 de enero de 2020 al 27 de febrero de 2020 procedían de Surging News Network (https://www.thepaper.cn/) y la WHO (https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports). La OMS tomó el caso confirmado por el laboratorio como los casos confirmados desde el 21 de enero de 2020 hasta el 13 de febrero de 2020 y tomó la suma del número de los casos clínicos confirmados y conformados por el laboratorio como el número de los casos confirmados

después del 14 de febrero de 2020. Los números de los casos clínicos confirmados y los casos confirmados por el laboratorio en la provincia de Hubei (China) el 14 de febrero fueron 15.384 y 36.602, respectivamente. Por lo tanto, el número de los casos confirmados en la Provincia de Hubei antes del 14 de febrero de 2020 fue ajustado por la fórmula: Nnúmero de casos confirmados: el número de casos confirmados en el laboratorio $\times \frac{(15384+36602)}{366602}$.

Los datos incluían las cifras totales de los casos confirmados acumulados y nuevos en toda China y las cifras de los casos confirmados acumulados y nuevos en 31 provincias/ciudades de China continental y otras tres regiones (Hong Kong, Macao y Taiwán) de China. Los datos se organizaron en un matríxico en el que las filas representaban toda China y las provincias/ciudades y las columnas representaban el número de nuevos casos confirmados de cada día.

Los casos confirmados de cada provincia/ciudad eran una serie temporal. Que t_{ij} ser el número de los casos confirmados de la j^{th} día dentro del i^{th} provincia/ciudad. Deje que Z ser un $34 \times m$ matriz dimensional. El elemento Z_{ij} es el número de los nuevos casos confirmados de Covid-19 en el j^{th} a partir del 11 de enero de 2020 en el i^{th} ciudad.

2.2. Auto-codificador modificado para modelar series temporales

Se utilizaron autocodificadores modificados, MAE) (Charte, Charte, García, Jesus, & Herrera, 2018; Yuan, Huang, Wang, Yang, & Gui, 2018) para pronosticar el número de los casos acumulados y nuevos casos confirmados de Covid-19. A diferencia del autocodificador clásico, en el que el número de nodos de las capas suele disminuir de la capa de entrada a las capas latentes, los números de los nodos de las capas de entrada, la primera capa latente, la segunda capa latente y las capas de salida en el MAE fueron 8, 32, 4 y 1, respectivamente (Figura 1). Vemos un segmento de serie temporal con 8 días como muestra de datos y tomamos 128 segmentos de serie temporal como muestras de formación. Un elemento de la matriz dada Z es seleccionado al azar como día de inicio del segmento y selecciona sus 7 días sucesivos como los otros días para formar un segmento de la serie de tiempo. Dejemos que i ser el índice del segmento y j_i ser el índice de la columna de la matriz Z que fue seleccionado como el día de inicio. El i^{th} Las series temporales de segmentos pueden representarse como $\{Z_{j_i}, Z_{j_i+1}, \dots, Z_{j_i+7}\}$. Los datos se normalizaron para $X_{j_i+k} = \frac{Z_{j_i+k}}{S}, k = 0, 1, \dots, 7$, en donde $S = \frac{1}{8}\sum_{k=0}^{7} Z_{j_i+k}$.

Let $Y_i = \frac{z_{j_i+s}}{s}$ ser el número normalizado de casos a pronosticar. Si $S = 0$, entonces $Y_i = 0$. La función de pérdida se definió como

$$L = \sum_{i=1}^{128} W_i (Y_i - \hat{Y}_i)^2,$$

donde Y_i fue el número de casos observados en el día de la previsión de la i^{th} series temporales de segmentos y $\hat{Y}_i$ fue su número de casos previsto por el MAE, y W_i eran valoradas. Si j_i fue en el intervalo [1, 12], entonces $W_i = 1$. Si j_i fue en el intervalo [13, 24], entonces $W_i = 2$, etc. El algoritmo de retropropagación se utilizó para estimar los pesos y el sesgo en el MAE. Repita los procesos de entrenamiento 5 veces. La previsión media $\hat{Y}_i, i = 1, \ldots, 34$ se tomará como un número final previsto de los casos confirmados acumulados para cada provincia/ciudad.

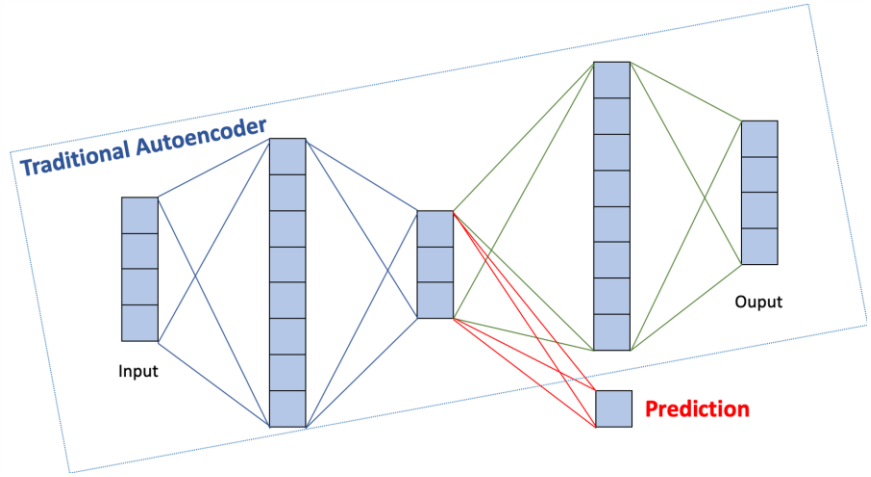


Figura 1. *Arquitectura de un MAE.*

2.3. Procedimientos de previsión

El MAE entrenado se utilizó para pronosticar el número futuro de los casos confirmados de Covid-19 para cada provincia/ciudad. Considere el i^{th} provincia/ciudad. Supongamos que el número de nuevos casos confirmados de Covid-19 en el j^{th} día que hay que pronosticar es x_{ij}. Que H sea una 34×8 matriz dimensional y $h_{il} = x_{ij-9+l}, i = 1, \ldots, 34, \text{ and } l = 1, \ldots, 8$. Que $g_i = \frac{1}{8}\sum_{l=1}^{8} h_{il}, i = 1, \ldots, 34$ ser el promedio de la i^{th} fila de la matriz H. Que U es la matriz normalizada de H donde

$u_{il} = \frac{h_{il}}{g_i}, i = 1, \dots, 34, \text{and } l = 1, \dots, 8$. El resultado del MAE es el número previsto de los nuevos casos confirmados y se denota como $\hat{v}_i = f(u_{i1}, \dots., u_{i8}, \theta), i = 1, \dots, 34$, donde θ representaban los parámetros estimados en el MAE entrenado. La previsión en un solo paso del número de nuevos casos confirmados de Covid-19 para cada ciudad viene dada por $\hat{Y}_i = \hat{v}_i g_i, i = 1, \dots, 34.$

La predicción recursiva de múltiples pasos implicaba el uso de un modelo de un solo paso en múltiples ocasiones en el que la predicción para el paso de tiempo precedente se utilizaba como insumo para hacer una predicción sobre el paso de tiempo siguiente. Por ejemplo, para pronosticar el número de los nuevos casos confirmados para el día siguiente, el número previsto de nuevos casos en el pronóstico de una etapa se utilizaría como dato de observación para predecir el día 2. Se repite el proceso anterior para obtener la previsión en dos etapas. La suma del número final previsto de los nuevos casos confirmados para cada provincia/ciudad se tomó como la predicción del número total de los nuevos casos confirmados de Covid-19 en China.

2.4. Clustering

Se extrajeron los valores de las variables latentes en la segunda capa latente del MAE para cada provincia/ciudad. Para cada provincia/ciudad, un 34×4 matrices latentes dimensionales A fueron formadas. El mayor valor individual λ de la matriz latente A se obtuvo a través de la descomposición de un solo valor. Realizamos cinco entrenamientos y obtuvimos cinco valores únicos más grandes. Para cada provincia/ciudad, formamos un vector de características que consistía en los cinco valores individuales más grandes λs, el día de inicio y el día previsto de finalización del brote de Covid-19, el día, el número de nuevos casos confirmados que alcanza el máximo, el mayor número de los nuevos casos confirmados previstos y el número de los casos confirmados acumulados previstos de Covid-19 en la provincia/ciudad respectiva. Los algoritmos de k-means se realizaron en el 34 presentan vectores para agrupar las provincias/ciudades en conglomerados.

5. Resultados

En la figura 2 se trazaron las curvas de número total de los casos acumulados y nuevos casos confirmados de Covid-19 notificados y pronosticados en China en función de los días. Los casos notificados fueron del 11 de enero de 2020 al 27 de febrero de 2020. Se disponía de un total de 47 días de datos. Comenzamos a hacer pronósticos el 20 de febrero de 2020. La figura 2 mostraba que la curva de pronóstico era cercana a la curva

reportada. De la Figura 2 se puede observar que la curva de los nuevos casos confirmados alcanzó el pico de 5.236 el 5 de febrero de 2020 y disminuyó a cero el 20 de abril (pronóstico). Se observó que los posibles casos confirmados acumulados de Covid-19 en China alcanzaron la meseta (83.401) el 20 de abril de 2020. En la figura S1 se trazaron las curvas nacionales notificadas y ajustadas de los casos confirmados acumulados en China desde el 20 de enero de 2020 hasta el 27 de febrero de 2020. Para examinar más a fondo la exactitud del pronóstico, la Tabla 1 donde se utilizaron los datos del 20 de enero de 2020 al 27 de febrero de 2020 para ajustar el modelo MAE. En el cuadro, se enumeraron los errores de previsión de 1 a 10 días, respectivamente, es decir, los errores de utilización de los casos notificados actualmente para prever los casos de días futuros. El cuadro 1 mostró que los errores medios de la previsión no aumentaron estrictamente, ya que el número de pasos para la previsión aumentó debido a las fluctuaciones de los datos. La precisión de las previsiones era muy alta.

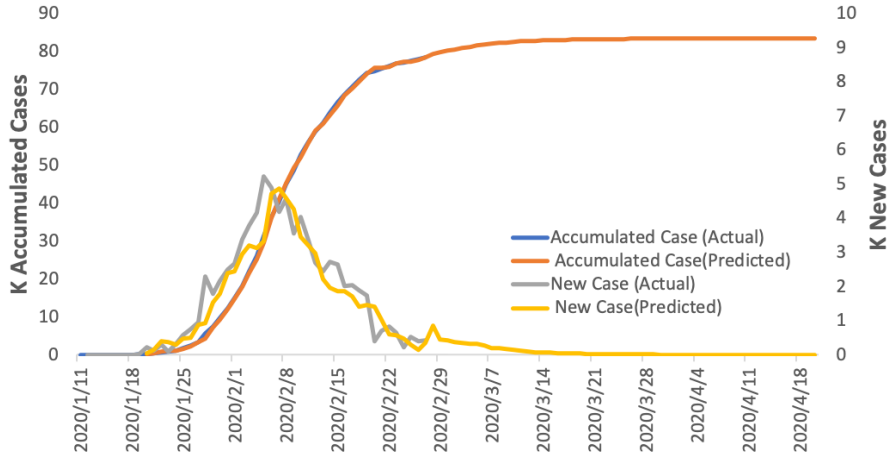


Figura 2. *Las curvas nacionales notificadas y pronosticadas de los casos acumulados y nuevos confirmados de Covid-19 en China en función de los días comprendidos entre el 11 de enero de 2020 y el 20 de abril de 2020.*

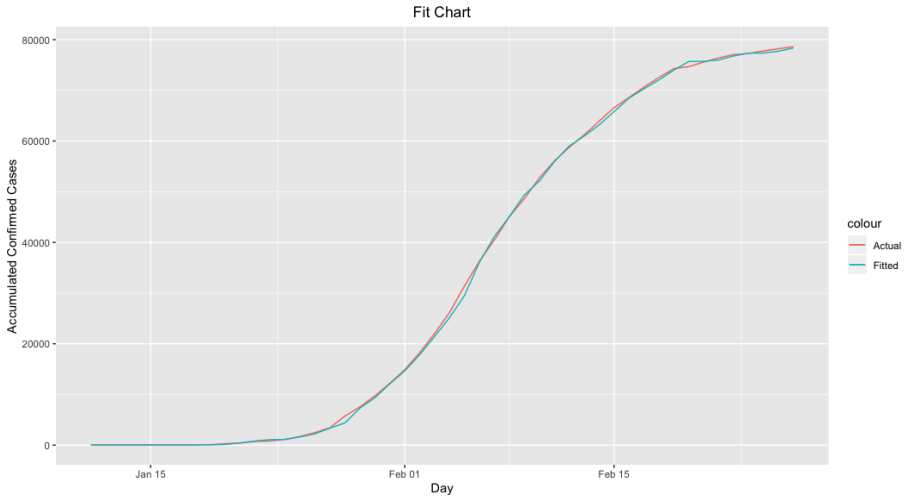


Figura S1. *Las curvas nacionales notificadas y ajustadas de los casos acumulados confirmados de Covid-19 en China desde el 11 de enero de 2020 hasta el 27 de febrero de 2020, en las que la curva roja era la notificada y la verde la ajustada.*

La figura 3 presentaba las curvas previstas de los casos confirmados acumulados de Covid-19 en 31 provincias/ciudades de China continental y otras tres regiones (Hong Kong, Macao y Taiwán) de China en función de los días comprendidos entre el 11 de enero y el 20 de abril de 2020. Observamos en la figura 3 que los tiempos de las diferentes provincias/ciudades que entran en la meseta serán diferentes, oscilando entre el 21 de enero de 2020 y el 20 de abril de 2020. Xizang entró en la meseta por primera vez el 21 de enero, seguido de Macao el 26 de enero y Qinghai el 28 de enero. La provincia de Hubei será una de las últimas en entrar en la meseta (70.019) alrededor del 20 de abril de 2020, la mayoría de las provincias/ciudades entrarán en la meseta a mediados de marzo. También podemos observar las diferentes formas de las curvas entre las provincias/ciudades, que implican diferentes patrones dinámicos de la transmisión de Covid-19 a través de las provincias/ciudades.

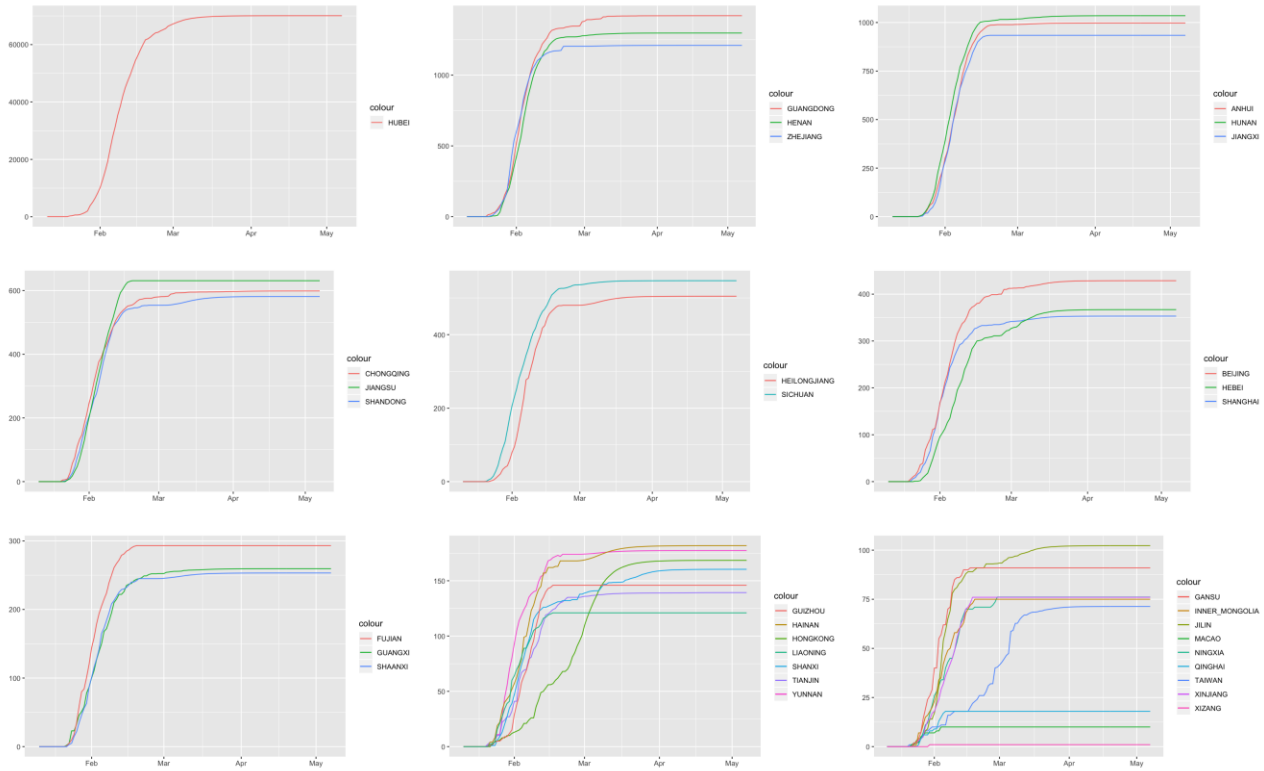


Figura 3. *Las curvas previstas de los casos confirmados acumulados de Covid-19 en 34 provincias/ciudades de China en función de los días comprendidos entre el 11 de enero de 2020 y el 20 de abril de 2020.*

Muchos factores, como el tamaño de la población que sale de Wuhan hacia cada provincia/ciudad afectada, las intervenciones, la ubicación geográfica, las actividades económicas y sociales, la heterogeneidad ambiental y las instalaciones de atención de la salud afectan a la dinámica de la transmisión de enfermedades en todo el país. La agrupación de su dinámica temporal proporcionará numerosos conocimientos sobre los patrones de propagación del Covid-19. Para captar mejor el patrón dinámico de la propagación de Covid-19 a través de las provincias/ciudades, presentamos la Figura 4, donde 34 provincias/ciudades formaron nueve conglomerados. Dado que la propagación de Covid-19 estaba influenciada por el patrón de contactos entre individuos, los cúmulos mostraban parcialmente la estructura geográfica.

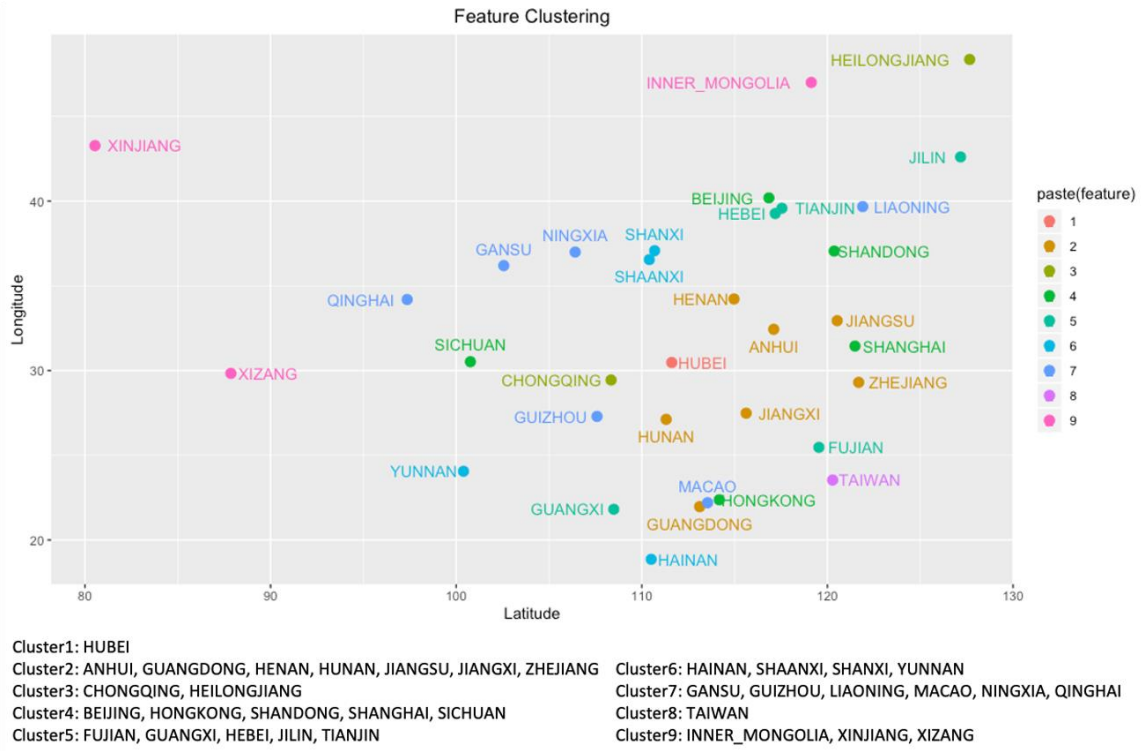


Figura 4. *Los grupos que se agruparon por características extraídas del MAE y las series cronológicas acumuladas de casos confirmados de Covid-19 en 31 provincias/ciudades de China continental y otras tres regiones de China formaron 9 grupos.*

Por ejemplo, la provincia de Hubei, una de las principales fuentes de Covid-19, formó un grupo. Seguido por los de Anhui, Guangdong, Henan, Hunan, Jiangsu, Jiangxi y Zhejiang, que se encuentran en las zonas que rodean las fuentes de patógenos (provincia de Hubei). El grupo de Mongolia Interior, Xizang y Qinghai tuvo el menor número de casos y se encuentran en zonas muy alejadas de las fuentes de Covid-19. Además de las ubicaciones geográficas que afectaron a la transmisión del Covid-19, los recursos sanitarios y las actividades económicas y sociales también pueden afectar a la dinámica de la transmisión del Covid-19 como el grupo de Shanghai, Hongkong, Beijing, Shandong y Sichuan. Y Fujian Guangxi, Hebei, Jilin y Tianjin formaron un conglomerado, pero estas cinco provincias/ciudades no están ubicadas en la misma área geográfica. Pueden tener relaciones económicas similares con Wuhan, recursos sanitarios y tomar intervenciones similares para controlar la propagación de Covid-19.

6. Discusión

Como alternativa al modelo de transmisión epidemiológica, utilizamos el MAE para pronosticar la trayectoria en tiempo real de la dinámica de la transmisión y generar los pronósticos en tiempo real de Covid-19 en todas las provincias/ciudades de China. Los resultados mostraron que la precisión de la predicción y, por consiguiente, de la previsión en múltiples etapas era alta. Nuestra experiencia reveló que la previsión mejora cuando el tiempo de formación es más largo. Calculamos los posibles puntos temporales de la disminución del crecimiento de las curvas de nuevos casos confirmados en 34 provincias/ciudades, la duración de las epidemias de Covid-19 en toda China y los momentos en que el número de casos confirmados acumulados de Covid-19 alcanzaría la meseta de sus curvas de casos acumulados. Si los datos son fiables y no habrá una segunda transmisión, los modelos del MAE predijeron que el brote de Covid-19 en China podría terminar a mediados de abril. En este estudio, utilizamos el análisis de conglomerados para agrupar 34 provincias/ciudades en 9 conglomerados y explorar sus estructuras de recursos geográficos y sanitarios.

Los modelos del MAE permiten introducir la información de las intervenciones e investigar el impacto de las mismas en el tamaño del brote del virus y el tiempo final del mismo. Sin embargo, no hemos explorado tales funciones del MAE debido a la falta de datos. De manera similar a los modelos dinámicos de transmisión epidemiológica, el MAE también puede ser usado para simulaciones. El MAE entrenado puede aproximarse bien a muchos procesos dinámicos. Utilizando los tamaños iniciales del brote epidémico, podemos usar el MAE con parámetros y arquitectura conocidos para estimar los tamaños del brote en el futuro y simular el impacto de las intervenciones en los tamaños y la severidad de las epidemias. Como complemento de un enfoque modelo de la dinámica de la transmisión de los brotes de virus, los métodos basados en la IA impulsados por los datos proporcionan herramientas de previsión en tiempo real para el seguimiento, la estimación de la trayectoria de las epidemias, la evaluación de su gravedad, la predicción de la duración de las epidemias y la ayuda al gobierno y a los trabajadores sanitarios para que puedan planificar y tomar decisiones acertadas.

Apéndice: Tabla 1.

Date	Actual	1-step prediction	1-step error	2-step error	3-step error	4-step error	5-step error	6-step error	7-step error	8-step error	9-step error	10-step error
18/02/2020	72.528	71.757	-1,06%									
19/02/2020	74.280	74.005	-0,37%	-1,34%								
20/02/2020	74.675	75.564	1,19%	1,31%	-0,06%							
21/02/2020	75.569	76.685	1,48%	1,36%	1,49%	0,12%						
22/02/2020	76.392	76.305	-0,11%	1,49%	1,41%	1,81%	0,15%					
23/02/2020	77.042	77.827	1,02%	0,02%	1,58%	1,39%	2,30%	0,13%				
24/02/2020	77.262	79.837	3,33%	2,18%	0,83%	2,53%	2,46%	3,30%	1,08%			
25/02/2020	77.780	79.155	1,77%	3,01%	2,14%	0,47%	2,28%	2,36%	3,15%	0,76%		
26/02/2020	78.191	77.957	-0,30%	1,61%	2,70%	2,15%	0,28%	2,20%	2,49%	3,21%	0,80%	
27/02/2020	78.630	78.646	0,02%	-0,53%	1,49%	2,47%	2,22%	0,21%	2,37%	2,46%	3,37%	0,73%
Average Absolute Error			1,07%	1,43%	1,46%	1,56%	1,62%	1,64%	2,27%	2,14%	2,08%	0,73%

Tabla 1. *Errores en el pronóstico de los casos nacionales confirmados acumulados en China*..

Referencias

Charte, D., Charte, F., García, S., del Jesus, M. J., & Herrera, F. (2018). A practical tutorial on autoencoders for nonlinear feature fusion: Taxonomy, models, software and guidelines. *Information Fusion*, 44, 78-96.

Funk, S., Camacho, A., Kucharski, A. J., Eggo, R. M., & Edmunds, W. J. (2018). Real-time forecasting of infectious disease dynamics with a stochastic semi-mechanistic model. *Epidemics*, 22, 56-61.

Johansson, M. A., Apfeldorf, K. M., Dobson, S., Devita, J., Buczak, A. L., Baugher, B. et al. (2019). An open challenge to advance probabilistic forecasting for dengue epidemics. *Proceedings of the National Academy of Sciences*, 116(48), 24268-24274.

Kucharski, A., Russell, T., Diamond, C., Liu, Y, (2020). CMMID nCoV working group. In J. Edmunds, S. Funk, & R. Eggo (ed.). *Analysis and projections of transmission dynamics of nCoV in Wuhan*. Retrieved from https://cmmid.github.io/ncov/wuhan_early_dynamics/index.html.

Li, Q., Guan, X., Wu, P., *et al.* (2020) Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*, 2020 Jan 29. [Epub ahead of print] doi: 10.1056/NEJMoa2001316.

Tuite, A. R., & Fisman, D. N. (2020). Reporting, epidemic growth, and reproduction numbers for the 2019 novel coronavirus (2019-nCoV) epidemic. *Annals of Internal Medicine*, 2020 Feb 5. [Epub ahead of print] doi: 10.7326/M20-0358..

Wu, J.T., Leung, K., Leung, G.M. (2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *The Lancet*. 2020 Jan 31. [Epub ahead of print] pii:S0140-6736(20)30260-9. doi: 10.1016/S0140-6736(20)30260-9.

Yuan, X., Huang, B., Wang, Y., Yang, C., & Gui, W. (2018). Deep learning-based feature representation and its application for soft sensor modeling with variable-wise weighted SAE. *IEEE Transactions on Industrial Informatics*, 14(7), 3235-3243.

Zhao, S., Musa, S. S., Lin, Q., Ran, J., Yang, G., Wang, W, et. al. (2020). Estimating the unreported number of novel coronavirus (2019-nCoV) cases in China in the first half of January 2020: a data-driven Modelling analysis of the early outbreak. *Journal of Clinical Medicine*, 9(2), 388.