# How does the outbreak of 2019-nCoV spread in mainland China? A retrospective analysis of the dynamic transmission routes*

Xiandeng Jiang† Le Chang‡ Yanlin Shi§**

March 1, 2020

**Abstract**

The fourth outbreak of the Coronaviruses, known as the 2019-nCoV, has occurred in Wuhan city of Hubei province in China in December 2019. We propose a time-varying sparse vector autoregressive (VAR) model to retrospectively analyze and visualize the dyamic transmission routes of this outbreak in mainland China over January 31 – February 19, 2020. Our results demonstrate that the influential inter-province routes from Hubei have become unidentifiable since February 4, whereas the self-transmission in each province was accelerating over February 4–15. From February 16, all routes became less detectable, and no influential transmissions could be identified on February 18 and 19. Such evidence supports the effectiveness of government interventions, including the travel restrictions in Hubei. Implications of our results suggest that in addition to the origin of the outbreak, virus preventions are of crucial importance in provinces with the largest migrant workers percentages (e.g., Jiangxi, Henan and Anhui) to controlling the spread of 2019-nCoV.

*This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

** Corresponding author.

†The School of Public Finance and Taxation, Southwestern University of Finance and Economics, Chengdu, Sichuan, P.R. China, 611130.

‡ Research School of Finance, Actuarial Studies, and Statistics, Australian National University, Canberra, ACT, Australia, 2601.

§ Department of Actuarial Studies and Business Analytics, Macquarie University, Sydney, NSW, Australia, 2109.

# 1 Introduction

Coronaviruses are single-stranded, enveloped and positive-sense RNA viruses, which are spherical in shape and have petal-like spines [1]. Firstly discovered and identified in 1965 [2], coronaviruses have not caused large-scale outbreaks until the 2003 SARS epidemic in China, followed by 2012 MERS in Saudi Arabia and 2015 MERS in South Korea [3]. Although the exact origin remains debatable [4], the fourth outbreak has taken place in Hubei province of China in December 2019 and rapidly spread out nationally [5–10]. On January 10, 2020, the WHO officially named this new coronavirus as the 2019 novel coronavirus, or 2019-nCoV, and released a comprehensive interim guidance on dealing with this new virus for all countries [11]. As of February 19, there are 75,101 confirmed cases (including 2,121 death report) in China, among which over 80% are from Hubei and over 50% are from Wuhan, the capital city of Hubei [12].

To combat against the rapid spread of the 2019-nCoV, since mid-January 2020, the central government of China and all local governments have implemented intensive preventions. Examples include tracing close contacts and quarantining infected cases, promoting social consensus on self-protection like wearing face mask in public area, among others [13]. With the unexpectedly rapidly growing number of confirmed cases, more extreme and unprecedented measures have taken places. On 23 January, the Chinese authorities introduced travel restrictions on five cities (Wuhan, Huanggang, Ezhou, Chibi and Zhijiang) of Hubei, shutting down the movement of more than 40 million people [14]. Among existing research, most argues that those interventions have effectively halted the spread of the 2019-nCoV [14–24].

During this anti-epidemic war, statistic and mathematical modeling plays a non-negligible role. Among the emerging large volume of studies, the classical susceptible exposed infectious recovered (SEIR) model wth its various extensions is the most popular method [25–38]. SEIR family models are effective in exploring the epidemic characteristics of the outbreak, forecasting the inflection point and ending time, and deciding the measures to curb the spreading. Despite this, they are less appropriate in identifying transmission routes of the 2019-nCoV outbreak, which is also not thoroughly investigated in existing literature.

In this paper, we fill in this gap and perform a retrospective analysis using the publicly

available data [12]. Rather than employing the SEIR, we develop a time-varying coefficient sparse vector autoregressive (VAR) model. Using the least absolute shrinkage and selection operator (lasso) [39, 40] and the local constant kernel smoothing estimator [41], our model is capable of estimating the dynamic high-dimensional Granger causality coefficient matrices. This enables the detection and visualization of time-varying inter-province and self-transmission routes of the 2019-nCov on the daily basis. The resulting "road-map" can help policy-markers and public-health officers retrospectively evaluate both the effectiveness and unexpected outcomes of their interventions. Such an evaluation is critical to winning the current battle against 2019-nCoV in China, providing useful experience for other countries facing the emerging threat of this new coronavirus, and saving lives when a new epidemic occurs in the future.

# 2 Methods

## 2.1 Model

Throughout this study, we are interested in the growth rate $y_{i,t}$ such that:

$$y_{i,t+1} = \ln(x_{i,t+1}) - \ln(x_{i,t}), \tag{1}$$

where $x_{i,t}$ is the accumulated confirmed cases in province $i$ on day $t$ ($i = 1, \ldots, N$ and $t = 1, \ldots, T$). $T$ and $N$ define the number of days and number of provinces under consideration, respectively. We then define $\boldsymbol{y}_t = (y_{1,t}, \ldots, y_{N,t})'$, an $N \times 1$ vector of the growth rate on day $t$. To investigate a dynamic direct transmission of the growth rate among provinces, we propose a time-varying coefficient sparse VAR model, namely the tvSVAR model, which assumes that Granger causality coefficients are functions of time, such that:

$$\boldsymbol{y}_{t+1} = \boldsymbol{\alpha}_t + \boldsymbol{B}_t \boldsymbol{y}_t + \boldsymbol{\epsilon}_t, \tag{2}$$

where $\boldsymbol{\alpha}_t$ is an $N$-dimensional intercept vector at time $t$. $\boldsymbol{B}_t$ is an $N \times N$ Granger causality matrix at time $t$ with a dynamic sparse structure, for which entries can be exactly zero and the locations of zeros can vary with time. $\boldsymbol{\epsilon}_t$ is an $N \times 1$ vector of error terms. The

sparsity of $\boldsymbol{B}_t$ is assumed because $N$ could be even larger than $T$ in our case, which leads to very unstable estimations and problematic interpretations of $\boldsymbol{B}_t$.

One important benefit of using the proposed tvSVAR to model the transmissions is that the Granger causality matrix, $\boldsymbol{B}_t$, can provide both the direction and strength of the route on day $t$. For example, the $ij^{th}$ entry in $\boldsymbol{B}_t$ measures the strength of the transmission from province $i$ to province $j$ on day $t$. The $i^{th}$ diagonal of $\boldsymbol{B}_t$ represents the self-transmission in province $i$ that captures the relationship between the growth rate in the current and previous days. More critically, the sparse structure eases the interpretation of $\boldsymbol{B}_t$ because many weak transmissions may be of a random nature. The corresponding coefficients, therefore, can be treated as noises and are shrunk to zeros exactly. Moreover, a time-varying design of $\boldsymbol{B}_t$ allows us to investigate changes in the identified transmissions over time.

To capture both dynamic and sparse structure of the Granger causality coefficients, we solve the following optimization problem:

$$(\widehat{\boldsymbol{\alpha}}_t, \widehat{\boldsymbol{B}}_t) = \arg\min_{\boldsymbol{\alpha}_t, \boldsymbol{B}_t} \left\{ \sum_{s=1}^{T-1} \left(\boldsymbol{y}_{s+1} - \boldsymbol{\alpha}_t - \boldsymbol{B}_t \boldsymbol{y}_s\right)^T \boldsymbol{W}_s \left(\boldsymbol{y}_{s+1} - \boldsymbol{\alpha}_t - \boldsymbol{B}_t \boldsymbol{y}_s\right) + \lambda \sum_{i,j}^{N} w_{i,j,t} |\beta_{i,j,t}| \right\} \tag{3}$$

where $\boldsymbol{W}_s = \text{diag}\left(K_{b_1}(\tau_s - \tau), \ldots, K_{b_N}(\tau_s - \tau)\right)$ is the matrix of kernel weights calculated based on the bandwidth $b_i$, $i = 1, \ldots, N$, and $K_{b_i}(\tau_s - \tau) = K(\frac{\tau_s - \tau}{b_i})/b_i$ with $\tau_s$ defined as a scaled time $\frac{s}{T-1}$. We use the Epanechnikov kernel $K(x) = 0.75(1 - x^2)_+$ and a unified bandwidth for each $i$ ($b_i \equiv b$) to avoid a large number of tuning parameters. The coefficients $\beta_{i,j,t}$ denotes the $ij^{th}$ entry of the Granger causality matrix $\boldsymbol{B}_t$, and $\lambda$ is the tuning parameter that aims to shrink insignificant $\beta_{i,j,t}$ to zero and thus controls the sparsity of $\boldsymbol{B}_t$. Another essential feature of our proposed model is that the adaptive weights $w_{i,j,t}$ are employed to penalize $\beta_{i,j,t}$ differently in the lasso (L1) penalty [39, 40]. The choice of weights $w_{i,j,t}$ takes account of the prior knowledge about the transmissions and can be specified by the users. In this study, we consider $w_{i,j,t}$ as the reciprocal of the accumulated confirmed in province $i$ on day $t-1$. That is, the growth rate of a province with a smaller accumulated confirmed cases is less likely to influence the growth rates of others, and thus,

more likely to be shrunk to zero. The final sparsity structure of $\boldsymbol{B}_t$ is still data-driven.

The estimators as in (3) can also be viewed as a penalized version of local constant kernel smoothing estimator [41]. We utilize a modified version of the fast iterative soft thresholding algorithm (FISTA) [42] to solve the optimization problem (3).

Given a bandwidth $b$ and a penalty parameter $\lambda$, we can find the estimator $(\widehat{\boldsymbol{\alpha}}_t, \widehat{\boldsymbol{B}}_t)$ for each day $t$ and observe the dynamic patterns of the transmission over time $t$ for each pair of provinces. The bandwidth and penalty parameter are chosen to optimize the cross-validated forecasting accuracy, as measured by the one-step-ahead root mean squared forecast error (RMSFE), such that

$$\text{RMSFE}(b,\lambda) = \sqrt{\frac{1}{N(T_1 - T_0 - 1)} \sum_{i=1}^{N} \sum_{t=T_0}^{T_1-1} \left( \widehat{y}_{i,t+1}^{(b,\lambda)} - y_{i,t+1} \right)^2}, \tag{4}$$

where $[T_0, T_1]$ is the evaluation period, which is given by the last third of the data in our study, $\widehat{y}_{i,t+1}^{(b,\lambda)}$ denotes the one-step-ahead forecast for province $i$ based on the data up to day $t$, and $y_{i,t+1}$ defines the observed growth rate at day $t+1$ for province $i$.

## 2.2 Code availability

The R code that supports the findings of this study is available from the author on request.

# 3 Data and Results

## 3.1 Data

The data studied in this paper include confirmed 2019-nCoV cases which occurred in mainland China from the National Health Commission of the People's Republic of China [12]. The data-coverage ranges from January 29, 2020 to February 19, 2020, during which no missing data were recorded at province-level. The accumulated cases and the associated growth rates, grouped by the total national number, cases in Hubei province and cases in all other provinces, are plotted in Figure 1 (a) and (b), respectively. The total national (Hubei province) accumulated confirmed cases increased rapidly from 7,736 (4,586) on January 31

to 75,101 (62,457) on February 19. Note that on February 12, confirmed cases in Hubei included those confirmed by both laboratory and clinical diagnosis, leading to a one-time hump of the accumulated number. Compared to those of Huber, confirmed cases of other provinces took up a smaller proportion of the total national number, ranging from 40.7% on January 29 to 16.8% on February 19. This suggests that the growth rate of other provinces should be lower than that of Hubei, which is consistent with Figure 1 (b). Throughout our investigation period, except for the one-time hump on February 12, growth rates of Hubei and the rest steadily declined, from 33% and 25% to 5% and 1%, respectively.

## 3.2 Estimation results: Transmission routes

By taking the difference of the logged accumulated cases and applying one lag, our estimated transmission routes are available from January 31 to February 19 (two observations are lost). To avoid potential noises caused by small numbers, we only include data of provinces, which had at least 150 accumulated confirmed cases as of February 19. Altogether, our modeled sample contains 20 province-level confirmed cases. A non-zero estimate of $\beta_{i,j,t}$, the $ij^{th}$ entry of $\boldsymbol{B_t}$ in (2), indicates that on the $t^{th}$ day, the growth rate of province $j$ is Granger caused by that of province $i$. In other words, there is a transmission route from province $i$ to province $j$. Among the 20-day results, we noticed that the estimated transmission routes on days 1–5 changed considerably on daily basis. From the sixth day onwards, however, those estimated routes were more steady. Hence, we plot the estimates on days 1–5 and those on the every fifth day thereafter, on Figure 2. Be noted that estimates smaller than 0.2 (none-influential) are not presented to avoid inevitable estimation noises and for a better illustration.

In Figure 2, we use color of light orange (small) to dark red (large) indicating the accumulated confirmed cases in each province, up to time $t$. Estimated transmission routes are colored in blue. Self-transmissions are denoted by dots, and a larger size of dot suggests a larger estimated $\beta_{i,i,t}$. Inter-province transmission is represented by arrows, with the transparency indicating the magnitude of estimated $\beta_{i,j,t}$. On the first day (January 31), there were influential inter-province transmissions from Hubei to Jiangxi, Heilongjiang, Zhejiang, Henan, Shandong, Jiangsu and Shaanxi, sorted by the magnitudes of strength

(big to small). There were a few additional detected such transmissions on the second day, including those from Hubei to Guangxi, from Zhejiang to Fujian, and from Guangdong to Yunnan, Hunan and Fujian. The number of such identified inter-province routes, however, reduced rapidly over the next three days. On the fifth day (February 4), no influential transmission routes were found from Hubei to directly affect other provinces, and there were only three influential routes identified nationally, including Zhejiang–Shaanxi, Zhejiang–Jiangxi and Jiangxi–Shanghai. The number of those detected inter-province routes declined again in the next few days, and on day 13, only Henan–Heilongjiang was found influential. On days 19 and 20 (February 18 and 19), there were no influential inter-province transmissions identified. The above findings suggest that the number of influential inter-province transmissions overall dropped quickly in the first five days and then reduced steadily for the rest fifteen days. This is consistent with the observations of Figure 3 (a), where the time-varying estimates of the Granger causality of Hubei on other provinces are plotted. On each day, we report the mean, standard deviation (Std. Dev.), the 25% quantile ($Q_1$) and 75% quantile ($Q_3$) of those estimates in Table 1, which also leads to consistent findings.

As for the self-transmission, we firstly examine Figure 2 (b). It can bee seen that there were quite a few detected influential self-transmissions on the first two days. However, this number dropped quickly over days 3–5, and only self-transmissions of Heilongjiang, Guangdong and Zhejiang were found influential on day 5. Since then, the number of influential self-transmissions increased quickly with growing magnitudes (influence). On the sixteenth day (February 15), 16 out of the 20 examined provinces had an estimated $\beta_{i,i,20}$ of at least 0.2. Those large self-transmissions, however, disappeared rapidly again in the next three days. On February 18 and 19, there were no influential self-transmissions identified. This is consistent with our findings on Figure 3 (b), where time-varying estimated $\beta_{i,i,t}$ are plotted for each province. We report daily descriptive statistics of those estimates in Table 1, which also results in consistent conclusions.

# 4 Discussions

Since 23 January 2020, many cities on mainland China started to introduce travel restrictions, including five cities (Wuhan, Huanggang, Ezhou, Chibi and Zhijiang) of Hubei

province [14]. According to [43], the average incubation period of 2019-nCoV is up to 10 days. Thus, our estimated dynamic transmission routes supports the significant effectiveness of the interventions taken by the Chinese authorities [14–24]. This is evidenced by Figure 2 (a)–(e), where the number of influential inter-province transmissions from Hubei to other provinces reduced very quickly. Compared to multiple influential routes originating in Hubei detected on the first two days (January 31 and February 1), by February 4 (around 10 days after the travel restrictions), there were already no such transmissions identified. On the other hand, from February 5 to 16, Table 1 suggests that the averaged magnitudes of self-transmission on each day were strengthening steadily. This may also be explained by the interventions, which have effectively blocked inter-province transmissions, such that the growth rate of each province could only be caused by its internal transmissions.

We now focus on the inter-province transmission routes. Since influential routes from Hubei were no longer detected since day 5, we calculate the average $\beta_{i,j,t}$ of the 19 provinces affected by Hubei over the first four days (January 31 – February 3). The top five destinations are presented in Panel A of Table 2. Apart from its geographic neighbors Jiangxi and Henan, Hubei has lead to influential routes to Heilongjiang, Zhejiang and Shandong directly. This may be explained by the substantial floating population working and living in Hubei from those provinces. Excluding routes originating in Hubei, the Panel B of Table 2 suggests that the top destinations of transmission routes have not changed much over the first four days and the rest sixteen days. Despite minor differences in ranking, Shaanxi, Heilongjiang, Jiangxi, Anhui, Henan and Jiangsu appear to be the destinations suffered most from inter-province transmissions from origins other than Hubei. Similarly, the top five sources (excluding Hubei) of those transmissions are basically identical over the days 1–4 and days 5–20, as shown in Panel C of Table 2. This is consistent with the fact that travel restrictions in Hubei should not affect the connections among other provinces. In all cases, Jiangxi, Henan, Guangdong, Zhejiang and Anhui are the most influential origins other than Hubei.

It is worth noting that Jiangxi, Henan and Anhui belong to both the top origins and destinations of the inter-province transmissions, excluding Hubei. Since the impact of Hubei is not considered, this cannot be explained by the two influential transmission routes

of Hubei–Jiangxi and Hubei–Henan listed in Panel A of Table 2. To see this, over days 5–20, the transmissions out of Hubei are no longer significant and thus should not affect routes from Jiangxi and Henan to another province. In contrast, one explanation is the large migrant workers from Jiangxi, Henan and Anhui to other provinces (excluding Hubei). According to the Report on China's migrant population development of 2017 [44], Jiangxi (7.25%), Henan (6.30%) and Anhui (6.27%) are among the top five provinces in mainland China, ranked by the percentages of migrant workers in 2017.

# 5 Conclusions

Coronaviruses have lead to three major outbreaks ever since the SARS occurred in 2003. Although the exact origin is still debatable, the current shock, namely 2019-nCoV, has taken place in Wuhan, the capital city of Hubei province in mainland China. As the fourth large-scale outbreak of coronaviruses, 2019-nCoV is spreading quickly to all provinces in China and has recently become a world-wide epidemic. As a significant complement to existing research, this study employs a tvSVAR model and retrospectively investigates and visualizes the transmission routes in mainland China.

Demonstrated in Figure 2, our baseline results review both the dynamic inter-province and self-transmission routes. Since February 4, the spread out of Hubei was largely reduced, leading to no identifiable routes to other provinces. Simultaneously, the self-transmissions started to accelerate and peaked on around February 15 for most provinces. Given an average incubation period of 10 days, those results support the argued effectiveness of the travel restrictions to control the spread of 2019-nCoV, which took place in multiple cities of Hubei on January 23. On February 18–19, there existed no influential inter-province or self-transmission routes. Thus, the growth rates of confirmed cases are of a more random nature in all provinces thereafter, implying that the spread of 2019-nCoV has been under control. For the detected inter-province transmissions, our findings demonstrate that Jiangxi, Heilongjiang, Zhejiang, Henan and Shandong are the top 5 provinces affected mostly via routes directly from Hubei. When the influence of Hubei is excluded, Jiangxi, Henan and Anhui are among both the top origins and destinations of transmission routes.

Our results have major practical implications for public health decision- and policy-

makers. For one thing, the implemented timely ad-hoc public health interventions are proven effective, including contact tracing, quarantine and travel restrictions. For another, apart from the origin of the virus, as provinces with largest migrant workers percentages, virus preventions are also of crucial importance in Jiangxi, Henan and Anhui to controlling the epidemics like the outbreak of 2019-nCoV in the future. With limited resources, taking ad-hoc interventions in such provinces may most effectively help stop the spread of a new virus, from an economic perspective.

# Author Contributions

X.J. collected data and designed the research. L.C. and Y.S. performed the research and analyzed the data. X.J., L.C., and Y.S. wrote the paper.

# References

1. Chen, Y., Liu, Q. & Guo, D. Emerging coronaviruses: genome structure, replication, and pathogenesis. *Journal of Medical Virology* (2020).
2. Kahn, J. S. & McIntosh, K. History and recent advances in coronavirus discovery. *The Pediatric Infectious Disease Journal* **24,** S223–S227 (2005).
3. Bogoch, I. I. *et al.* Pneumonia of Unknown Etiology in Wuhan, China: Potential for International Spread Via Commercial Air Travel. *Journal of Travel Medicine* (2020).
4. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature,* 1–4 (2020).
5. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature,* 1–5 (2020).
6. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet* **395,** 497–506 (2020).
7. Cohen, J. & Normile, D. *New SARS-like virus in China triggers alarm* 2020.
8. Lu, H., Stratton, C. W. & Tang, Y.-W. Outbreak of Pneumonia of Unknown Etiology in Wuhan China: the Mystery and the Miracle. *Journal of Medical Virology* **92,** 401–402 (2020).
9. Parry, J. *China coronavirus: cases surge as official admits human to human transmission* 2020.
10. Li, G. & De Clercq, E. *Therapeutic options for the 2019 novel coronavirus (2019-nCoV)* 2020.

11. World Health Organization (WHO). `https://www.who.int/healthtopics/coronavirus` (2020).

12. National Health Commission of the People's Republic of China. `http://www.nhc.gov.cn/xcs/xxgzbd/gzbd_index.shtml` (2020).

13. Peng, L., Yang, W., Zhang, D., Zhuge, C. & Hong, L. Epidemic analysis of COVID-19 in China by dynamical modeling. *arXiv preprint arXiv:2002.06563* (2020).

14. Tang, B. *et al.* Estimation of the Transmission Risk of the 2019-nCoV and Its Implication for Public Health Interventions. *Journal of Clinical Medicine* **9,** 462 (2020).

15. Tang, B. *et al.* An updated estimation of the risk of transmission of the novel coronavirus (2019-nCov). *Infectious Disease Modelling* (2020).

16. Shen, M., Peng, Z., Guo, Y., Xiao, Y. & Zhang, L. Lockdown may partially halt the spread of 2019 novel coronavirus in Hubei province, China. *medRxiv* (2020).

17. Li, X., Zhao, X. & Sun, Y. The lockdown of Hubei Province causing different transmission dynamics of the novel coronavirus (2019-nCoV) in Wuhan and Beijing. *medRxiv* (2020).

18. Chinazzi, M. *et al.* The effect of travel restrictions on the spread of the 2019 novel coronavirus (2019-nCoV) outbreak. *medRxiv* (2020).

19. Backer, J. A., Klinkenberg, D. & Wallinga, J. Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28 January 2020. *Eurosurveillance* **25** (2020).

20. Lai, S. *et al.* Assessing spread risk of Wuhan novel coronavirus within and beyond China, January-April 2020: a travel network-based modelling study. *medRxiv* (2020).

21. Quilty, B. J., Clifford, S., *et al.* Effectiveness of airport screening at detecting travellers infected with novel coronavirus (2019-nCoV). *Eurosurveillance* **25** (2020).

22. Clifford, S. J. *et al.* Interventions targeting air travellers early in the pandemic may delay local outbreaks of SARS-CoV-2. *medRxiv* (2020).

23. Hellewell, J. *et al.* Feasibility of controlling 2019-nCoV outbreaks by isolation of cases and contacts. *medRxiv* (2020).

24. Jin, G., Yu, J., Han, L. & Duan, S. The impact of traffic isolation in Wuhan on the spread of 2019-nCov. *medRxiv* (2020).

25. Chen, T. *et al.* A mathematical model for simulating the transmission of Wuhan novel Coronavirus. *bioRxiv* (2020).

26. Chen, Y., Cheng, J., Jiang, Y. & Liu, K. A time delay dynamical model for outbreak of 2019-nCoV and the parameter identification. *arXiv preprint arXiv:2002.00418* (2020).

27. Huang, N. E. & Qiao, F. A data driven time-dependent transmission rate for tracking an epidemic: a case study of 2019-nCoV. *Science Bulletin* (2020).

28. Jung, S.-m. *et al.* Epidemiological identification of a novel infectious disease in real time: Analysis of the atypical pneumonia outbreak in Wuhan, China, 2019-20. *medRxiv* (2020).

29. Lin, Q., Hu, T. & Zhou, X.-H. Estimating the daily trend in the size of COVID-19 infected population in Wuhan. *medRxiv* (2020).

30. Nishiura, H. *et al.* Estimation of the asymptomatic ratio of novel coronavirus infections (COVID-19). *medRxiv* (2020).

31. Nishiura, H *et al.* The Extent of Transmission of Novel Coronavirus in Wuhan, China, 2020. *Journal of clinical medicine* **9** (2020).

32. Read, J. M., Bridgen, J. R., Cummings, D. A., Ho, A. & Jewell, C. P. Novel coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions. *medRxiv* (2020).

33. Sanche, S. *et al.* The Novel Coronavirus, 2019-nCoV, is Highly Contagious and More Infectious Than Initially Estimated. *arXiv preprint arXiv:2002.03268* (2020).

34. Wu, J. T., Leung, K. & Leung, G. M. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *The Lancet* (2020).

35. Xiong, H. & Yan, H. Simulating the infected population and spread trend of 2019-nCov under different policy by EIR model. *Available at SSRN 3537083* (2020).

36. Yang, Y. *et al.* Epidemiological and clinical features of the 2019 novel coronavirus outbreak in China. *medRxiv* (2020).

37. Zeng, T., Zhang, Y., Li, Z., Liu, X. & Qiu, B. Predictions of 2019-nCoV Transmission Ending via Comprehensive Methods. *arXiv preprint arXiv:2002.04945* (2020).

38. Zhao, S. *et al.* Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. *International Journal of Infectious Diseases* (2020).

39. Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58,** 267–288 (1996).

40. Zou, H. The adaptive lasso and its oracle properties. *Journal of the American statistical association* **101,** 1418–1429 (2006).

41. Fan, J. & Gijbels, I. *Local polynomial regression* 1996.

42. Beck, A. & Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences* **2,** 183–202 (2009).

43. World Health Organization (WHO). `https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200127-sitrep-7-2019--ncov.pdf` (2020).

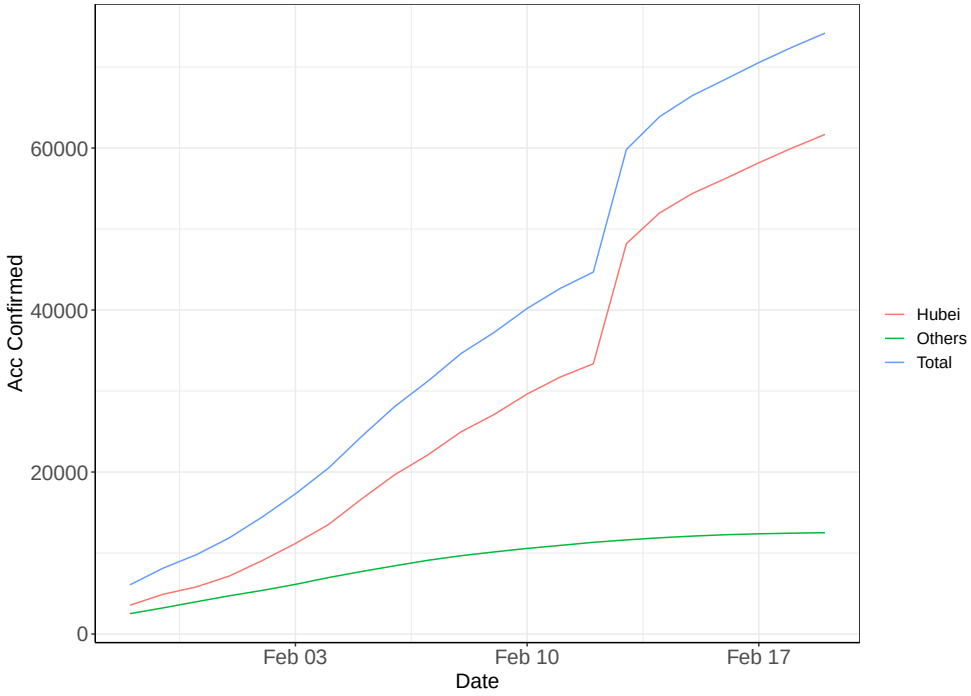
44. National Health and Family Planning Commission of China. *Report on China's migrant population development* (China Population Publishing House, Beijing, 2017).

Table 1: Summary of daily estimated transmission routes

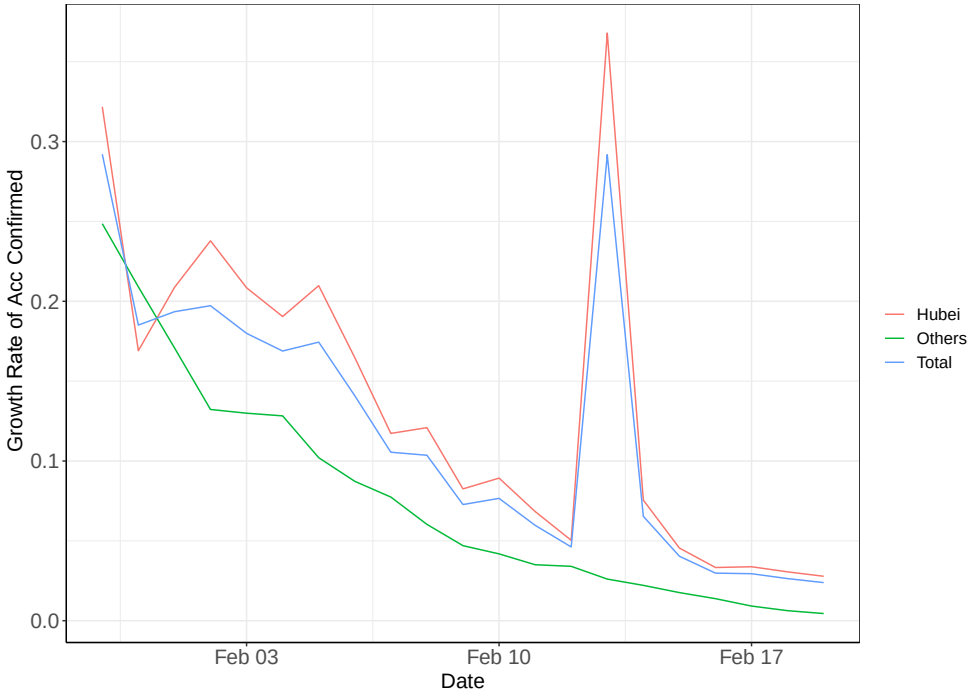
| | From Hubei to others | | | | Self-transmissions | | | |
|---|---|---|---|---|---|---|---|---|
| Day | Mean | Std. Dev. | $Q_1$ | $Q_3$ | Mean | Std. Dev. | $Q_1$ | $Q_3$ |
| 1 | 0.1602 | 0.1366 | 0.0802 | 0.2306 | 0.1430 | 0.1651 | 0.0154 | 0.2451 |
| 2 | 0.1906 | 0.1593 | 0.1019 | 0.2770 | 0.1684 | 0.2081 | 0.0160 | 0.2847 |
| 3 | 0.1405 | 0.0961 | 0.0825 | 0.1961 | 0.1273 | 0.1508 | 0.0357 | 0.1864 |
| 4 | 0.1469 | 0.0924 | 0.0781 | 0.1938 | 0.1119 | 0.1143 | 0.0487 | 0.1837 |
| 5 | 0.0467 | 0.0548 | 0.0042 | 0.0662 | 0.1124 | 0.1008 | 0.0437 | 0.1706 |
| 6 | 0.0483 | 0.0539 | 0.0071 | 0.0695 | 0.1177 | 0.0967 | 0.0881 | 0.1687 |
| 7 | 0.0496 | 0.0509 | 0.0130 | 0.0725 | 0.1242 | 0.0982 | 0.0969 | 0.1751 |
| 8 | 0.0518 | 0.0507 | 0.0236 | 0.0766 | 0.1328 | 0.1013 | 0.1103 | 0.1873 |
| 9 | 0.0523 | 0.0514 | 0.0179 | 0.0788 | 0.1417 | 0.1064 | 0.1165 | 0.1933 |
| 10 | 0.0516 | 0.0515 | 0.0086 | 0.0794 | 0.1528 | 0.1147 | 0.1276 | 0.2152 |
| 11 | 0.0507 | 0.0508 | 0.0066 | 0.0770 | 0.1546 | 0.1129 | 0.1311 | 0.2172 |
| 12 | 0.0490 | 0.0490 | 0.0035 | 0.0698 | 0.1844 | 0.0807 | 0.1280 | 0.2304 |
| 13 | 0.0543 | 0.0457 | 0.0231 | 0.0756 | 0.2164 | 0.0936 | 0.1557 | 0.2950 |
| 14 | 0.0518 | 0.0442 | 0.0162 | 0.0697 | 0.2465 | 0.0946 | 0.1623 | 0.3387 |
| 15 | 0.0468 | 0.0444 | 0.0110 | 0.0636 | 0.2789 | 0.1245 | 0.1742 | 0.3763 |
| 16 | 0.0388 | 0.0439 | -0.0010 | 0.0556 | 0.3646 | 0.1614 | 0.2151 | 0.4746 |
| 17 | 0.0447 | 0.0408 | 0.0053 | 0.0600 | 0.3297 | 0.1477 | 0.2241 | 0.4420 |
| 18 | 0.0431 | 0.0360 | 0.0123 | 0.0646 | 0.1919 | 0.0997 | 0.1294 | 0.2467 |
| 19 | 0.0032 | 0.0017 | 0.0022 | 0.0041 | 0.0021 | 0.0016 | 0.0012 | 0.0027 |
| 20 | 0.0028 | 0.0016 | 0.0017 | 0.0036 | 0.0016 | 0.0015 | 0.0009 | 0.0018 |

Table 2: Top five provinces of the inter-province transmissions

| Days | High – Low | | | | |
|---|---|---|---|---|---|
| *Panel A: top destinations (affected by Hubei)* | | | | | |
| 1–4 | Jiangxi | Heilongjiang | Zhejiang | Henan | Shandong |
| *Panel B: top destinations (affected by provinces excluding Hubei)* | | | | | |
| All | Shaanxi | Heilongjiang | Jiangxi | Anhui | Henan |
| 1-4 | Shaanxi | Jiangxi | Heilongjiang | Henan | Jiangsu |
| 5-20 | Heilongjiang | Shaanxi | Jiangxi | Anhui | Henan |
| *Panel C: top origins (excluding Hubei)* | | | | | |
| All | Jiangxi | Henan | Guangdong | Zhejiang | Anhui |
| 1-4 | Jiangxi | Guangdong | Zhejiang | Henan | Anhui |
| 5-20 | Henan | Jiangxi | Guangdong | Anhui | Zhejiang |

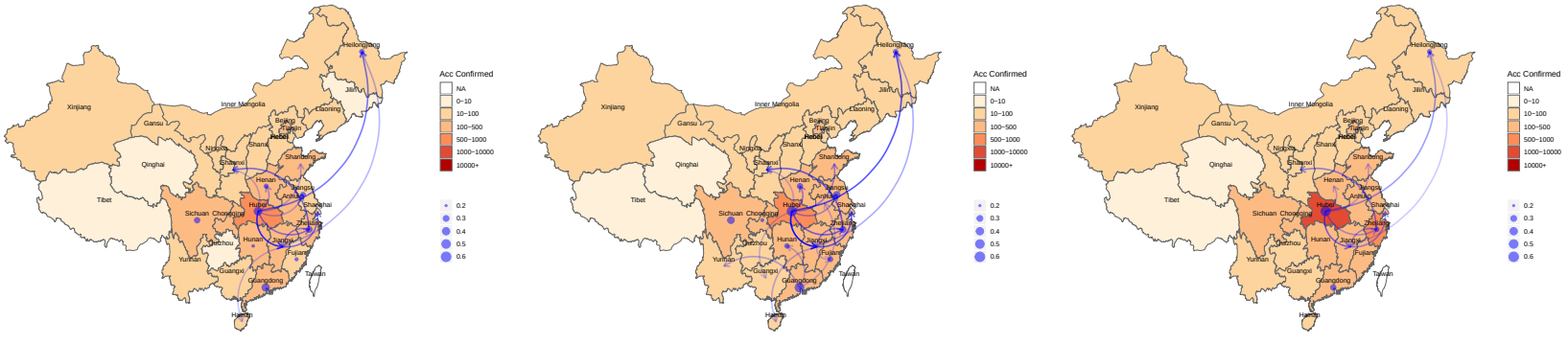



(a) accumulated cases




(b) Growth rate

Figure 1: Accumulated confirmed cases and growth rate: 31/1/2020–19/2/2020



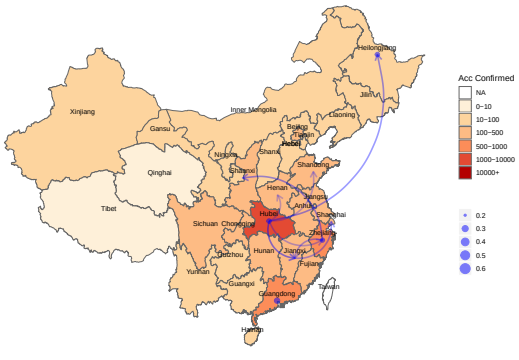

(a) January 31 (Day 1)



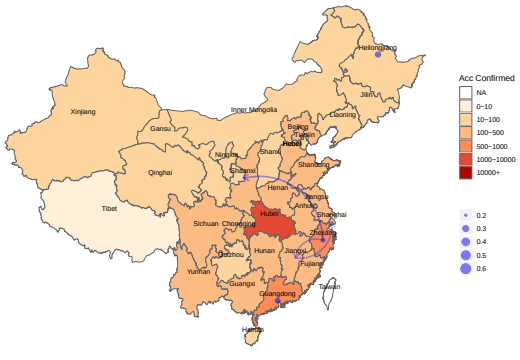

(b) February 1 (Day 2)



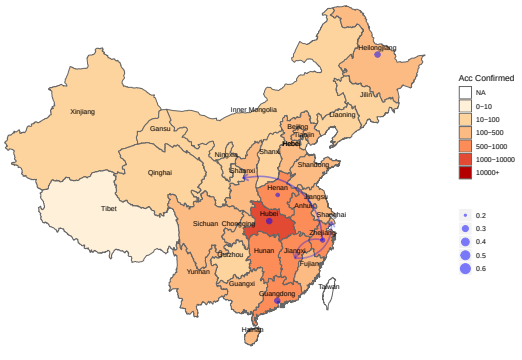

(c) February 2 (Day 3)



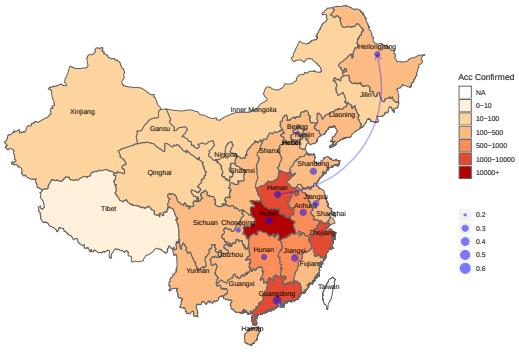

(d) February 3 (Day 4)



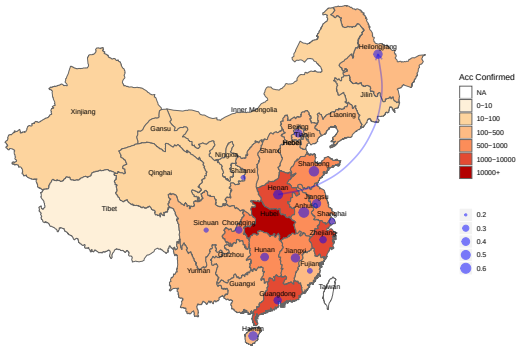

(e) February 4 (Day 5)



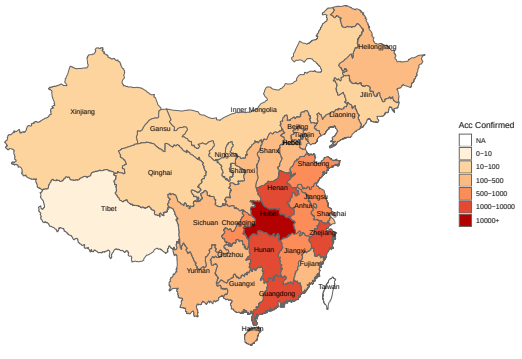

(f) February 8 (Day 9)




(g) February 12 (Day 13)

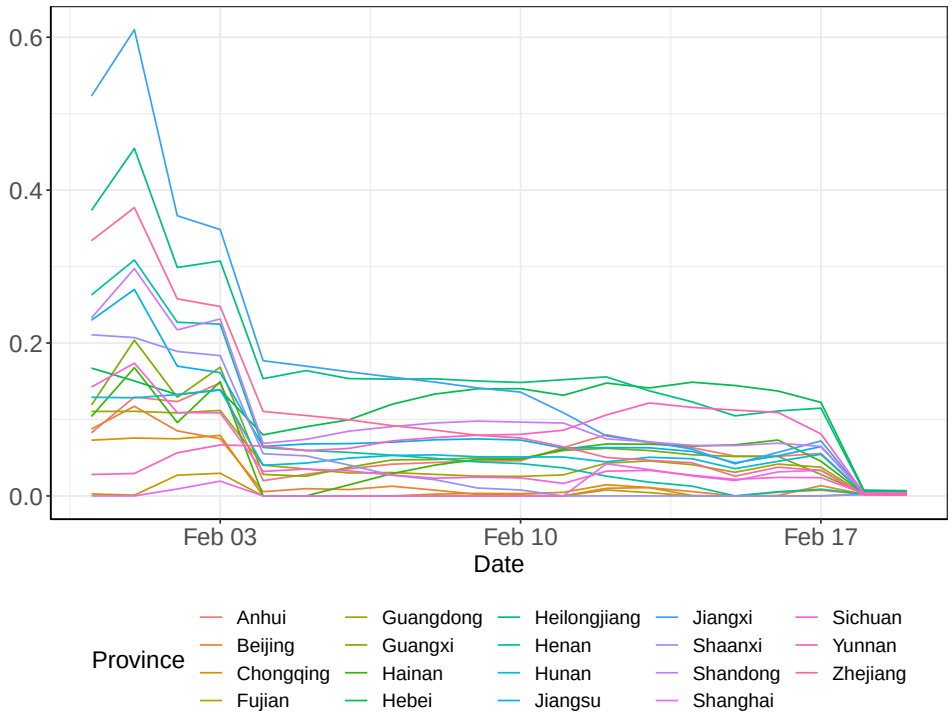



(h) February 16 (Day 17)

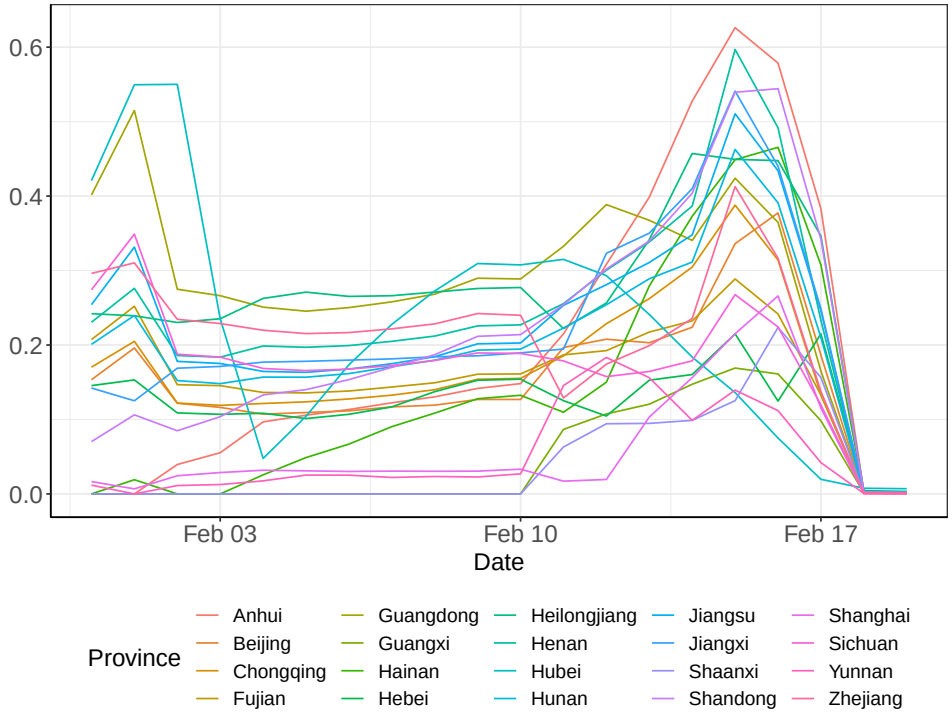



(i) February 19 (Day 20)

Figure 2: Estimated routes of transmission among provinces of China: 31/1/2020–19/2/2020




(a) Inter-province transmissions: from Hubei to others




(b) Self-transmissions

Figure 3: Estimated time-varying coefficients: 31/1/2020–19/2/2020