Population Flow Drives Spatio-Temporal Distribution of COVID-19 in China

Jayson S. Jia[1], Xin Lu[2, 3], Yun Yuan[4], Ge Xu[5], Jianmin Jia[6,7*], Nicholas A. Christakis[8]

[1]Faculty of Business and Economics, The University of Hong Kong, Hong Kong SAR, China.
[2]College of Systems Engineering, National University of Defense Technology, Changsha, China
[3]Department of Global Public Health, Karolinska Institutet, Stockholm, Sweden
[4]School of Economics and Management, Southwest Jiaotong University, Chengdu, China.
[5]School of Management, Hunan University of Technology and Business, Changsha, China
[6]Shenzhen Finance Institute, School of Management and Economics, The Chinese University of Hong Kong, Shenzhen, China.
[7]Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen, China.
[8]Yale Institute for Network Science, Yale University, New Haven, CT, U.S.A.

[*]Corresponding author. Email: jmjia@cuhk.edu.cn (JMJ)

Sudden, large-scale and diffuse human migration can amplify localized outbreaks into widespread epidemics. Rapid and accurate tracking of aggregate population flows may therefore be epidemiologically informative. Here we use 11,478,484 mobile-phone-data-based counts of individuals leaving or transiting through the prefecture of Wuhan between 1 January and 24 January 2020 as they moved to 296 prefectures throughout mainland China. First, we document the efficacy of quarantine in ceasing movement. Second, we show that the distribution of population outflow from Wuhan accurately predicts the relative frequency and geographical distribution of infections with SARS-CoV-2 until 19 February 2020, across mainland China. Third, we develop a spatio-temporal 'risk source' model that leverages population flow data (which operationalizes the risk that emanates from epidemic epicentres) not only to forecast the distribution of confirmed cases, but also to identify regions that have a high risk of transmission at an early stage. Fourth, we use this risk source model to statistically derive the geographical spread of COVID-19 and the growth pattern based on the population outflow from Wuhan; the model yields a benchmark trend and an index for assessing the risk of community transmission of COVID-19 over time for different locations. This approach can be used by policy-makers in any nation with available data to make rapid and accurate risk assessments and to plan the allocation of limited resources ahead of ongoing outbreaks.

Tracking population flows is especially exigent in the context of China's COVID-19 outbreak, which emerged in Wuhan (a prefecture-city in the province of Hubei) in the run-up to Chinese Lunar New Year eve on January 24, 2020 with its annual *chunyun* mass migration (which can involve as many as 3 billion trips). The potential scale and range of the outbreak's diffusion was particularly alarming given Wuhan's position as a central hub in China's rail and aviation networks and given the severity of COVID-19.

We used nationwide mobile phone data to track population outflow from Wuhan and linked this to COVID-19 infection counts by location – at the prefecture level. Our data include 296 prefectures in 31 provinces and regions in China (average population 4.40 million, 94.07% of China's population). Mobile phone geolocation data, which can reliably quantify human movement, provide precise, verifiable, and real-time information.[5-11] We conceptualize epidemiological morbidity as a function of human population movement from a disease epicenter. We thus normalize disease risk by population inflow from Wuhan rather than the size of local population.

Our approach differs from prior work linking individual mobility and disease spread[1-4, 12] in terms of: our use of real-time data about actual movement; our focus on aggregate population flows rather than individual tracking; and our particular modeling approach. That is, other recent research on COVID-19 has used *historical* population flow data (e.g., previous years' *chunyun* migrations) to estimate case exportation during the current outbreak.[14-18] But the benefits of observing rather than estimating population movements are substantial since inaccurate predictions can have important consequences for policy-making: under-reaction can result in disease spread, and over-reaction can lead to medically, socially, and economically inefficient policies. Moreover, distinct from prior approaches to epidemiological modelling,[12-18] we take advantage of detailed data about population flow emanating at the source of the outbreak to develop a population-flow-based "risk source" model to test the extent to which population flow data can capture the spatio-temporal dynamics of the spread of the SARS-CoV-2 virus.

To measure total aggregate population outflow from Wuhan prior to its quarantine on January 23, 2020, we used country-wide data, provided by a major national carrier, tracking all movement out of Wuhan between January 1 and January 24, 2020. The symptom onset of the first recorded case in Wuhan was December 1, 2019; by February 19, the end of our study period, 74,576 infection cases had been verified in mainland China.[19-22] Our time period includes the time that news about the outbreak initially appeared (on December 31, 2019 and January 9, 2020) and the annual Lunar New Year migration (which culminated on January 24, 2020). The dataset included any mobile phone user who had spent at least 2 hours in Wuhan during this period, and it tracked the total daily flow of such individuals to all other prefectures throughout mainland China. Locations were detected when users simply had their phones on. The dataset includes two measures of population outflow: the carrier's own customer count and their extrapolated count of total population movement. We use the latter in our primary analyses and the former as a robustness check (see Supplementary Information).

We defined population flow as the total aggregate count of people entering any given prefecture from Wuhan during the whole observation period (January 1 to 24). Since Wuhan (population 11.08 million in 2018) is a major transportation hub, many of these people were through travelers rather than residents. The definition is also weighted by number of transits through Wuhan since some people may have entered and exited Wuhan on several occasions in January (especially if they lived in neighboring prefectures). This can be thought of as a linear weighting of additional infection and transmission risk from repeated transits. There were 11,478,484 counts of movements from Wuhan: 8,685,007 to other prefectures within Hubei and 2,793,477 to prefectures in other provinces.

Key dates during this period were January 24, Lunar New Year's Eve (outbound holiday travel is typically completed before this evening), and January 23, when Wuhan was quarantined. We observed the efficacy of the quarantine (Fig. 1b, c), which was manifested in a 52% (38%) drop of inter- (intra-) provincial population outflow on January 23 compared to January 22 (when there were 546,324 and 141,208 counts of intra- and extra- provincial travel, respectively), and a further of 94% (84%) drop on January 24 compared to January 23. With the imposition of the quarantine – first with respect to Wuhan (and two neighboring prefectures) at 10 a.m. on January 23, and then with respect to 12 other prefectures in Hubei by the end of the day on January 24 – population outflow from Wuhan almost completely stopped (the average daily outflow thereafter was just 1,087 people to all prefectures outside of Hubei, probably government workers).

We combined the population flow dataset with the count and geographical location of COVID-19 confirmed cases nationwide (Fig. 1), which used consistent and stringently enforced case ascertainment during this period. As of February 19, 2020, there were 74,576 infection cases in mainland China; 29,549 cases occurred outside of Wuhan; and there were 2,118 fatalities.[22]

Population flow from Wuhan may be hypothesized to export the virus to other locations, where it causes local outbreaks (i.e., either by importation or "community transmission"[19-22]). And indeed, we find a strong correlation between total population flow and number of infections in each prefecture (Fig. 2a, b). Consistent with our hypothesis, the cumulative number of infections is highly correlated with aggregate population outflow from Wuhan from January 1 to 24, and the correlation increases over time from $r = 0.522$ on January 24 to 0.919 on February 5, and further to 0.952 on February 19 ($p < .001$ for all) (Fig. 2a, b, c). Since there is little travel throughout the country during this period, the population outflow variable is comparable to a lagged variable in a time series. The correlation exhibited the same robust pattern even when using different time windows of population outflow (Extended Fig. 1). The rates of confirmed infection cases based on population outflow from Wuhan remained uniform across time (Extended Fig. 9). The correlation between population outflow from Hubei province (excluding Wuhan itself) and number of infections in each prefecture (Fig. 2c) followed a similar pattern but was substantially weaker, $r = 0.365$ on January 24 to 0.583 on February 19.

For completeness, we compared the predictive strength of aggregate population outflow to certain other factors – such as the relative frequency of Baidu search engine queries for virus-related terms in each prefecture (e.g., novel coronavirus, flu, SARS, atypical pneumonia, surgical mask),[23-25] each prefecture's GDP and population, and also other movement variables. Each of these factors became *less* predictive of local outbreak size over time, either for cumulative or daily reported cases (Fig. 2c, d, Extended Fig.2-3).

We also evaluated a gravity model.[4,13] Gravity models were originally developed to model flow volumes or other interactions between geographical areas based simply on distance between two locales and their populations. Here, we use a special case of the gravity model with only the "recipient" prefecture's population variable since Wuhan is always the "donor" and thus a constant value (Supplementary Information 4.1). This model (with a significantly negative parameter for distance) predicts the high quantity of travel from Wuhan to other prefectures in

Hubei and to geographically proximate provinces (Fig. 1). But it does not explain the high traffic of population outflow to more distant coastal cities. That outflow does not strictly follow a gravity model is not surprising given the rationales for *chunyun* migration patterns, which are primarily based on social connections.[8,26]

We also tested a gravity model to predict the infection count. Although "recipient" population size and distance were significant predictors ($p < .001$), a mediation analysis shows that population flow from Wuhan mediates the effect of distance. Fig. 2c and 2d intuitively illustrate why this is the case. Aggregate population flow from Wuhan exhibits a high and progressively stronger correlation with infection prevalence in destination locations over time. In contrast, the predictive strength of prefecture's distance from Wuhan, population size, and GDP (an alternative source of "gravity") declines over time. There is no advantage to estimating population flow and to estimating infection spread using estimated population flow when actual population flow is observable, as in our case.

Next, we use two sets of models – one cross-sectional and the other dynamic – to statistically model and benchmark the extent to which aggregate population outflow from Wuhan predicts the spread and distribution of COVID-19 infections across mainland China. We develop what we call a "risk source" model that leverages observed population flow data to operationalize the risk emanating from the epidemic source.

We first modeled the effect of outflow on infection by using the following multiplicative exponential model:

$$y_i = c \prod_{j=1}^{m} e^{\beta_j x_{ji}} \, e^{\sum_{k=1}^{n} \lambda_k I_{ik}} \tag{1}$$

where y_i is the number of the cumulative (or daily) confirmed cases in prefecture i (depending on the model); x_{1i} is cumulative population outflow from Wuhan to prefecture i from January 1 to 24; x_{2i} is the GDP of prefecture i; x_{3i} is the population size of prefecture i; m is the number of variables included; and c and β_j are parameters to estimate. And λ_k is the fixed effect for province k; n is the number of prefectures considered in the analysis; I_{ik} is a dummy for prefecture i and $I_{ik} = 1$, if $i \in k$ (prefecture i belongs to province k), otherwise $I_{ik} = 0$. (See Supplementary Information for more details.)

We applied a nonlinear least squares method (Levenberg-Marquardt Algorithm) to estimate the parameters of a model with confirmed cases as the dependent variable and

We applied a nonlinear least squares method (Levenberg-Marquardt Algorithm) to estimate the parameters of a model with confirmed cases as the dependent variable and aggregate Wuhan population outflow from 1–24 January 2020 as the sole predictor variable ($R^2 = 0.772$ on 24 January to $R^2 = 0.946$ on 19 February) and a model with population size and GDP as additional co-variates ($R^2 = 0.809$ on 24 January 24 to $R^2 = 0.967$ on 19 February) (Supplementary Tables 1, 2). Although these additional co-variates improve the fit, the parameter for population flow from Wuhan becomes increasingly dominant, whereas the GDP and population of a prefecture become increasingly less predictive over time. Overall, the performance of the models continuously improved as more infected cases were confirmed, suggesting that the spreading pattern of the virus gradually converged to the distribution of the population outflow from Wuhan to other prefectures in China. As a robustness check, we evaluate a model using daily confirmed cases and find consistent results (Supplementary Tables 3, 4).

The logic behind this convergence over time, as well as the model's predictive strength, is that population flow from Wuhan to other prefectures fundamentally determines the eventual distribution of total infections in mainland China. During the earliest phase of the outbreak, before the quarantine of Wuhan, there was a relative lack of awareness of the virus and few countermeasures preventing its spread. SARS-CoV-2 should thus have spread relatively randomly across the entire prefecture of Wuhan; that is, our results imply that the number of infected people was uniformly distributed (statistically speaking) in the population outflowing from Wuhan into different prefectures across the country.

Using the daily predicted cases in model (1), we are also able to calculate a daily risk score for prefectures based on the difference between their predicted and confirmed cases on any given date (see Supplementary Information). A higher-than-expected level of infection suggests more community transmission (i.e., "underperforming" compared to the benchmark derived from the outflow population from Wuhan). On the other hand, "over-performing" prefectures, with fewer cases than expected are also noteworthy, since they could have implemented highly successful public health measures (or be prone to inaccurate data reporting). Extended Fig. 4 identifies prefectures with transmission risk index values over the upper bound of the 90% confidence interval on January 29, for example, and this was indeed associated with imminent quarantine. The predictive strength of aggregate population flow from Wuhan and the overall fit of model (1) over time can also act as an early warning index of an epidemiological transition; they reflect the degree to which imported infections are dominant at any point in time. If model strength declines significantly at any location, this may indicate that community transmission may be overtaking imported cases.

We next developed a spatio-temporal model to explore changes in distribution and growth of COVID-19 across all prefectures over time (rather than on individual dates) (Supplementary Information 3.2). We use a Cox proportional hazards framework and replace the constant scaling parameter of model (1) with a time-varying hazard rate function $\lambda_0(t)$,which typically has an S-shaped property (e.g., logistic, generalized logistic, or Gompertz functions[27-28]) that epidemics typically follow:

$$\lambda(t|x_i) = \lambda_0(t)\left(\prod_{j=1}^{m} e^{\beta_j x_{ji}}\right)e^{\sum_{k=1}^{n} \lambda_k I_{ik}} \qquad (2)$$

where $\lambda(t|x_i)$ is the hazard function describing the number of cumulative confirmed cases at time *t* given population outflow from Wuhan to prefecture *i,* and other variables $x_i = \{x_{1i}, x_{2i}, \dots x_{mi}\}$ are the realized values of the covariates for prefecture i; and the other notation is the same as model (1).

This model extends our risk source model to a dynamic context; it incorporates all infection cases across all locales and dates to statistically derive the COVID-19 epidemic curve and growth pattern across China. We used the same method as before to estimate the parameters (see Supplementary Information). When using only the single variable of total population outflow from Wuhan (from January 1-24) to each other prefecture, we observe $R^2 = 0.927$ for the exponential-logistic model (Fig. 3a); and the inclusion of local population and GDP increases R^2 to 0.957 (alternate models are in Supplementary Table 5).

We use a similar logic as before in contrasting expected and observed outcomes to gauge epidemiological risk. Here, model predictions serve as reference patterns across time (Extended Fig. 5, 6). The differences in the growth trends between predicted and confirmed cases can signal higher levels of SARS-CoV-2 community transmission. We use the integral of the differences over time to create a total transmission risk index (normalized by subtracting the mean and dividing by the standard deviation) and identify a list of prefectures above and below the 90% confidence interval (Extended Fig. 7, Supplementary Table 11). Indeed, our model identifies a list of statistically significant "underperformers"; in most of these cases, we observed the subsequent imposition of quarantine (see the Supplementary Information, including Supplementary Table 12 and Extended Fig. 5-6). On the other hand, prefectures with lower trends than expected might have had more successful public health measures. Fig. 3b depicts the dynamic shifts in risk index score for selected prefectures, which allows monitoring which prefectures performed better in controlling transmission risk over time.

In sum, using detailed mobile phone geolocation data to compute aggregate population movements, we track the transit of people from Wuhan to the rest of mainland China through January 24, 2020. The geographic flow of people anticipates the subsequent location, intensity, and timing of outbreaks in in the rest of mainland China through February 19, 2020. These data outperform other measures, such as population size, wealth, or distance from the risk source. We modeled the epidemic curves of COVID-19 across different locales using population flows and showed that deviations from model predictions served as tools to detect the burden of community transmission.

The logic of our population-flow-based "risk source" model differs from classic epidemiological models that rely on assumptions regarding population mixing, population compartment sizes, and viral properties. By assuming that risk arises from human population movements, our "risk source" model is able to parsimoniously capture the distribution of the epidemic. The model has several advantages: it makes no assumptions regarding travel patterns or effective distance effects; allows for non-linear estimations; generates a non-arbitrary, source-linked risk score; and is easily adapted to other empirical contexts. Importantly, the multiplicative functional form can also accommodate multiple risk sources – for example in countries where there are multiple disease epicenters. As an example, we evaluated the distinct impact of population flow from Hubei (excluding Wuhan) as an alternative risk source in our models, and indeed find that it had little impact on COVID-19 spread and growth in the country (Supplementary Tables 6 and 10).

We have focused on the relative strength of the outbreak in each area, rather than the absolute number of cases, though one can predict the number of cases by using reported data to calibrate the parameters of the model. A key contribution of our approach is to robustly characterize the structure or relative distribution of cases across different geographic areas and over time, which is driven fundamentally by the cumulative outflow from Wuhan. Moreover, another benefit is that non-systematic inaccuracy of COVID-19 case-finding is relatively unimportant as long as we capture the *distribution* of population flow accurately over time, which we do.

Our approach is generalizable to any dataset that captures population movements (e.g., train ticketing or car tolling data). This method can also be implemented in a live fashion (if suitable data are available) to facilitate policy decisions – for example the allocation of resources and manpower across specific geographic locales based on the predicted strength of the epidemic. This could also yield a dynamic performance metric when contrasted against real-time reports of infections, and, as we show, identify which areas have higher virus transmission risk or more effective measures.

Other techniques to forecast the levels of an epidemic in defined populations in advance have, of course, been proposed – whether the use of online search behaviour[23-25] or the use of network sensors (i.e., the monitoring of people who are at heightened risk for falling ill given their network position).[29] Our approach relies on data regarding population flow. Indeed, historical (i.e., baseline) information about population flows – undisturbed by the imposition of quarantines or by publicity regarding outbreaks, both of which happened here – could also be valuable to public health experts and government officials when new outbreaks occur.

When people move, they take contagious diseases with them. Their movements are thus a harbinger of the future status of an epidemic, and this offers the prospect of using data-analytic techniques to control an epidemic before it strikes too hard.

REFERENCES

1. Colizza, V., Barrat, A., Barthélemy, M. & Vespignani, A. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proc. Natl. Acad. Sci. USA* **103**, 2015-2020 (2006).
2. Halloran, M. E., Vespignani, A., Bharti, N., Feldstein, L. R., Alexander, K. A., Ferrari, M., ... & Del Valle, S. Y. Ebola: mobility data. *Science* **346**, 433-433 (2014).
3. Brockmann, D. & Helbing, D. The Hidden Geometry of Complex, Network-Driven Contagion Phenomena. *Science* **342**, 1337-1342 (2013)
4. Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J. J. & Vespignani, A. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proc. Natl. Acad. Sci. USA* **106**, 21484-21489 (2009).
5. Brockmann, D., Hufnagel, L. & Geisel, T. The scaling laws of human travel. *Nature* **439**, 462–465 (2006).
6. González, M. C., Hidalgo, C. A. & Barabási, A. L. Understanding individual human mobility patterns. *Nature* **453**, 779–782 (2008).
7. Onnela, J. P., Arbesman, S., Gonzalez, M. C., Barabási, A. L. & Christakis, N.A. Geographic Constraints on Social Network Groups. *Plos One* **6**, e16939 (2011)
8. Lu, X., Bengtsson, L. & Holme, P. Predictability of population displacement after the 2010 Haiti earthquake. *Proc. Natl. Acad. Sci. USA* **109**, 11576-11581 (2012).
9. Yan, X., Wang, W., Gao, Z. & Lai, Y. Universal model of individual and population mobility on diverse spatial scales, *Nat. Commun.* **8**, 1639 (2017).
10. Csáji, B. C., Browet, A., Traag, V. A., Delvenne, J. C., Huens, E., Van Dooren, P., ... & Blondel, V. D. Exploring the mobility of mobile phone users. *Physica A* **392**, 1459-1473 (2013).
11. Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. L., Noor, A. M., Snow, R. W., & Buckee, C. O. Quantifying the impact of human mobility on malaria. *Science* **338,** 267-70 (2012).
12. Adda, J. Economic activity and the spread of viral diseases: Evidence from high frequency data. *Q. J. Econ.* **131**, 891-941 (2016).

13. Viboud, C., Bjørnstad, O. N., Smith, D. L., Simonsen, L., Miller, M. A., & Grenfell, B. T. Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science* **312**, 447-451 (2006).

14. 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. *Lancet*. **395**, 689-697 (2020).

15. Wu, J.T., Leung, K., Bushman, M. et al. Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. *Nat. Med.* (2020).

16. Chinazzi, M., Davis, J. T., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S., ... & Viboud, C. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* (2020).

17. Du, Z., Wang, L., Cauchemez, S., Xu, X., Wang, X., Cowling, B. J. & Meyers, L. A. Risk of 2019 novel coronavirus importations throughout China prior to the Wuhan. *Lancet* **361**, 1761-6 (2020).

18. Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., & Shaman, J. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2). *Science* (2020).

19. Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., ... & Yuan, M. L. A new coronavirus associated with human respiratory disease in China. *Nature*, 1-8 (2020).

20. Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., ... & Niu, P. A novel coronavirus from patients with pneumonia in China, 2019. *New Engl. J. Med*. (2020).

21. Chan, J. F. W., Yuan, S., Kok, K. H., To, K. K. W., Chu, H., Yang, J., ... & Tsoi, H. W. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet*. (2020).

22. China Center for Disease Control and Prevention. (2020).

23. Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S. & Brilliant, L. Detecting influenza epidemics using search engine query data. *Nature* **457**, 1012-1014 (2009).

24. Lazer, D., Kennedy, R., King, G. & Vespignani, A. The parable of Google Flu: traps in big data analysis. *Science* **343**, 1203-1205 (2014).

25. Viboud, C. & Vespignani, A. The future of influenza forecasts. *Proc. Natl. Acad. Sci. USA*, **116**, 2802-2804 (2019).

26. Massey, D. S. & García España, F. The Social Process of International Migration, *Science* **237**, 733-738 (1987).

27. Bürger, R., Chowell, G., & Lara-Dıíaz, L. Comparative analysis of phenomenological growth models applied to epidemic outbreaks. *Math Biosci Eng*. **6**, 4250–4273 (2019).

28. Roosa, K., Lee, Y., Luo, R., Kirpich, A., Rothenberg, R., Hyman, J. M., Yan, P., & Chowell, G. Short-term Forecasts of the COVID-19 Epidemic in Guangdong and Zhejiang, China: February 13-23, 2020. *J Clin Med* **9**, 596 (2020).

29. Christakis, N. A. & Fowler, J. H. Social network sensors for early detection of contagious outbreaks. *PloS One* **5,** e12948 (2010).

Figures

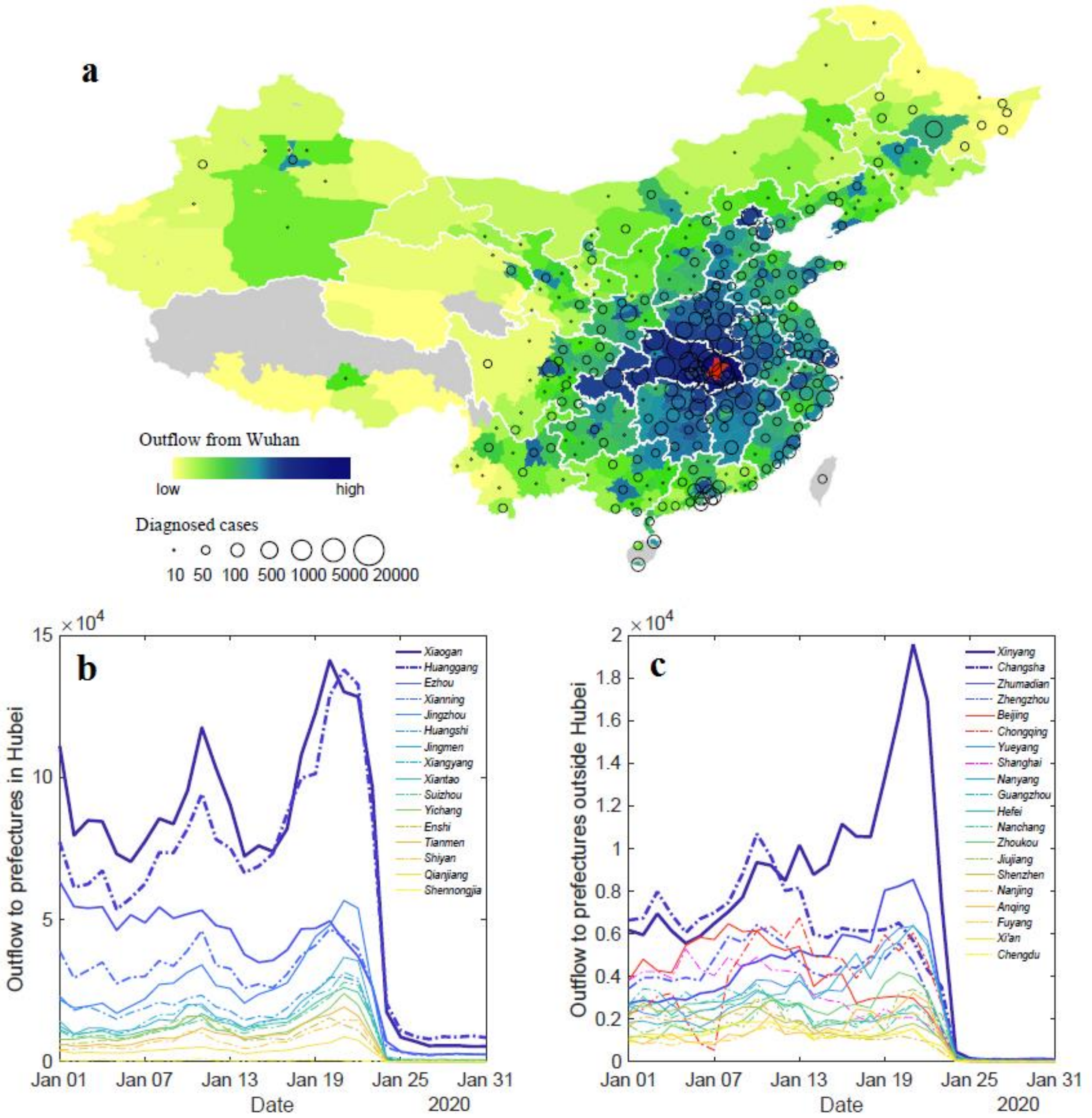


Figure 1 | Geographical distribution of population outflow and confirmed COVID-19 cases as of February 19, 2020. **a**, There is a high overlap between the geographical distribution of aggregate population outflow from Wuhan until 24 January 2020 (in red) and the number of confirmed cases of COVID-19 in other Chinese prefectures (*n* = 296 prefectures). Map source: National Catalogue Service For Geographic Information. Grey areas lack population outflow data. **b**, **c**, During the time that is historically the peak period for outbound Lunar New Year holiday travel, total population outflow from Wuhan to other parts of Hubei (**b**) is over three times higher than the population outflow to outside provinces (**c**). After the implementation of the quarantine at 10:00 on 23 January 2020, population outflow from Wuhan became minimal, except to the adjacent prefectures (**b**). In **b**, the first peak possibly corresponds to the start of the winter break of (roughly one million) college students in Wuhan and the second peak is associated with outbound Chunyun travel.

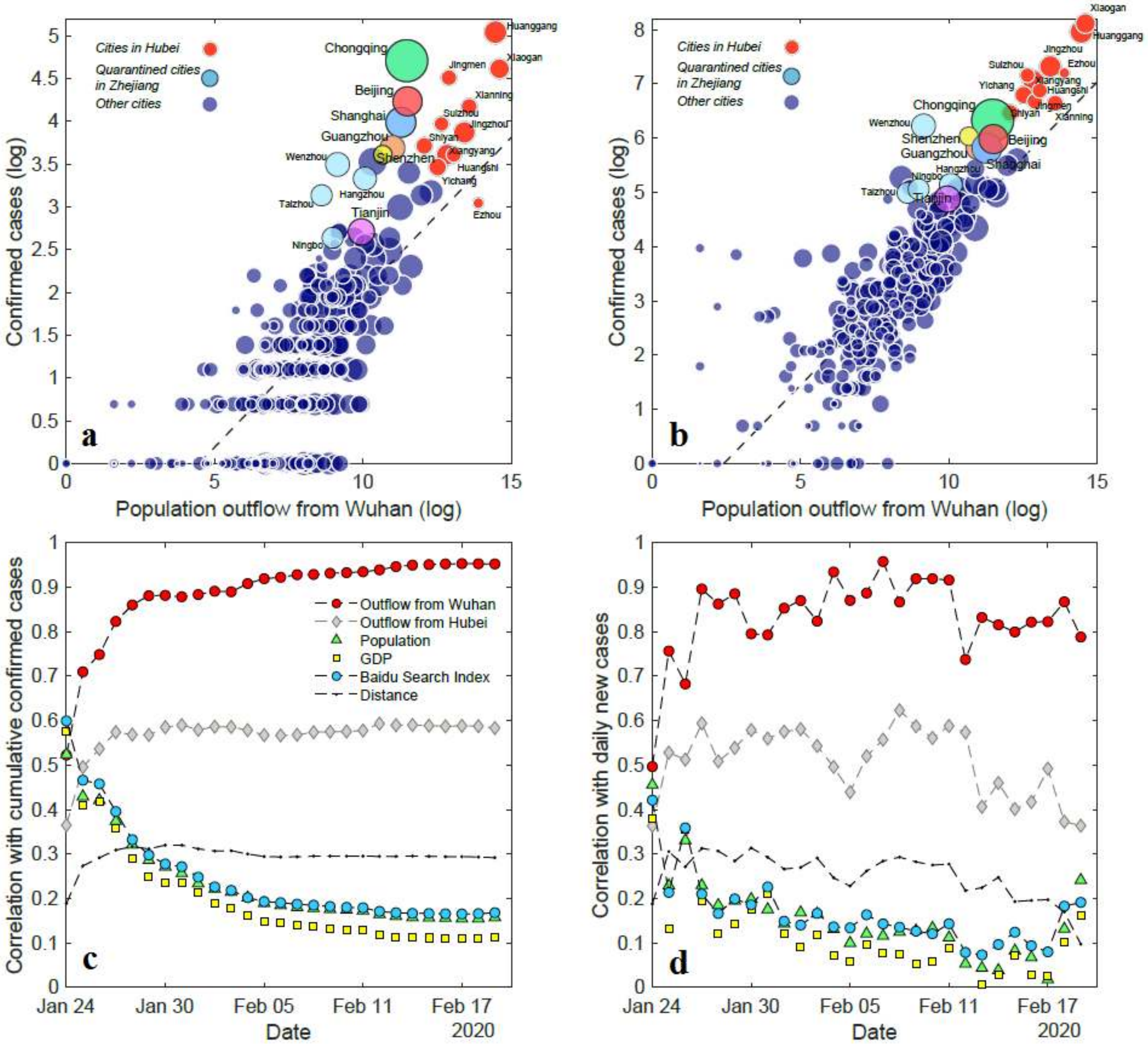


Figure 2 | Factors correlated with confirmed COVID-19 cases. a, b, The relationship between aggregate population outflow from Wuhan (up to January 24) and confirmed cases by prefecture on January 26 (**a**) and February 19 (**b**). Red circles are prefectures in Hubei; light blue circles are four quarantined prefectures in Zhejiang (including Wenzhou); and the six largest prefectures in China are indicated with unique colors. **c**, Relationship over time between number of confirmed cases (**c**, cumulative through February 19 on x-axis) and prefectures' (i) cumulative population inflow (up to Jan. 24) from Wuhan, (ii) cumulative inflow from Hubei province excluding Wuhan, (iii) frequency of Baidu search terms related to the virus, (iv) GDP, (v) population, and (vi) distance from Wuhan. Over time, the correlation between population outflow from Wuhan and the number of infection cases increases from Pearson's $r = 0.522$ on January 24 to $r = 0.952$

(N=296). The decline in the predictive strength of online search behavior might reflect information saturation, while the decline in predictive strength of GDP, population size, and distance suggests that late-stage *chunyun* migration from Wuhan was to a more diverse set of prefectures (and not merely to the closet, largest, and most developed prefectures) and/or that community transmissions began to predominate. The correlation with daily infections (**d**) is consistent, with Pearson's *r* ranging from 0.496 on January 24 to a peak of 0.926 on February 4 (N=296). Fluctuations are likely lags in case reporting (that are smoothed in **c**); weaker correlations on the last few days reflect that >90% of prefectures outside of Hubei reported no new cases.

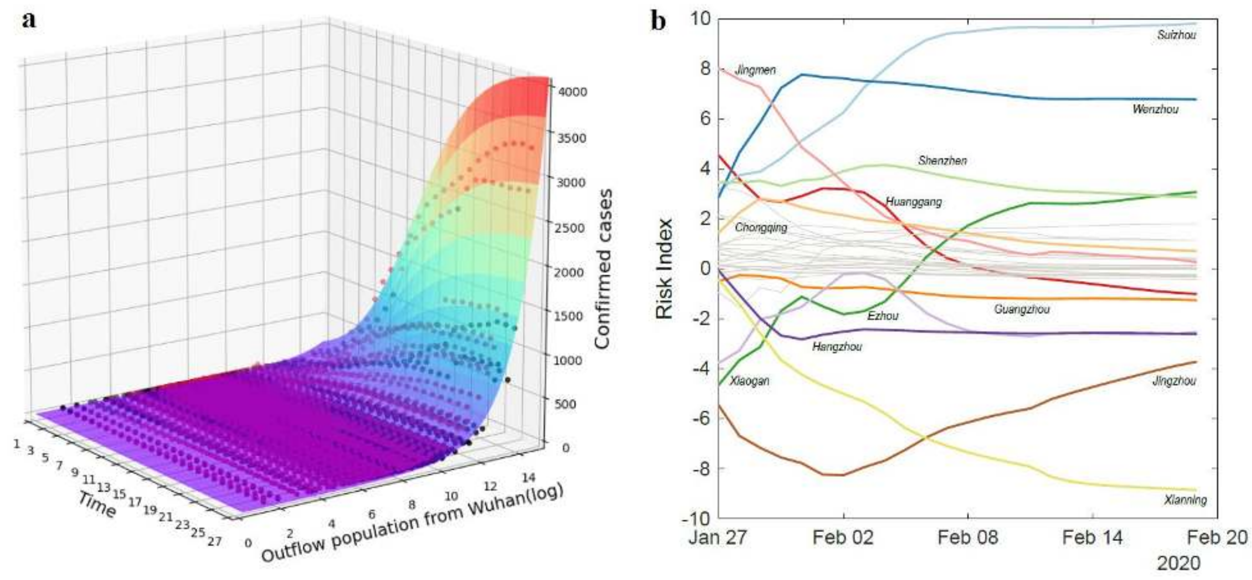


Figure 3 | Predictive model based on population outflow. **a,** The surface indicates the fitted performance of our epidemiological model (see Supplementary Information, model (3)) with just a single variable x_{1i} indicates outflow population from Wuhan to prefecture *i* (*log* transformed), for all prefectures, with *t* as the number of days after *chunyun* is over (i.e., *t* = 1 is January 24). The dots represents the actual number of comfired cases under a given x_{1i} and *t*. Red dots represent prefectures where the reported number of confirmed cases is greater than the model's predicted values; black dots are all other cases, R^2 = 0.930 (N=7,992). **b**, Risk scores over time provide a dynamic picture of shifting transmission risks in different prefectures.

Supplementary Information. Attached.

Acknowledgements. We thank a major unnamed national carrier for providing the anonymised and aggregated data enabling the computation of population movements. J.J. is supported by the National Natural Science Foundation of China (72042009 and 71490722) and Shenzhen Institute of Artificial Intelligence and Robotics for Society (2020-INT001). J.S.J. is supported by the Research Grants Council of Hong Kong (14505217). X.L. is supported by the National Natural Science Foundation of China (82041020, 91846301, 71771213, 71901067 and 61773120) and the Science and Technology Department of Sichuan Province (2020YFS0007). G.X. is supported by the National Natural Science Foundation of China (71704052, 91846301) and Hunan Provincial Key Laboratory of New Retailing Virtual Reality Technology (2015TP). We thank staff at the telecom carrier for their assistance in data preparation. This work was deemed exempt from IRB review.

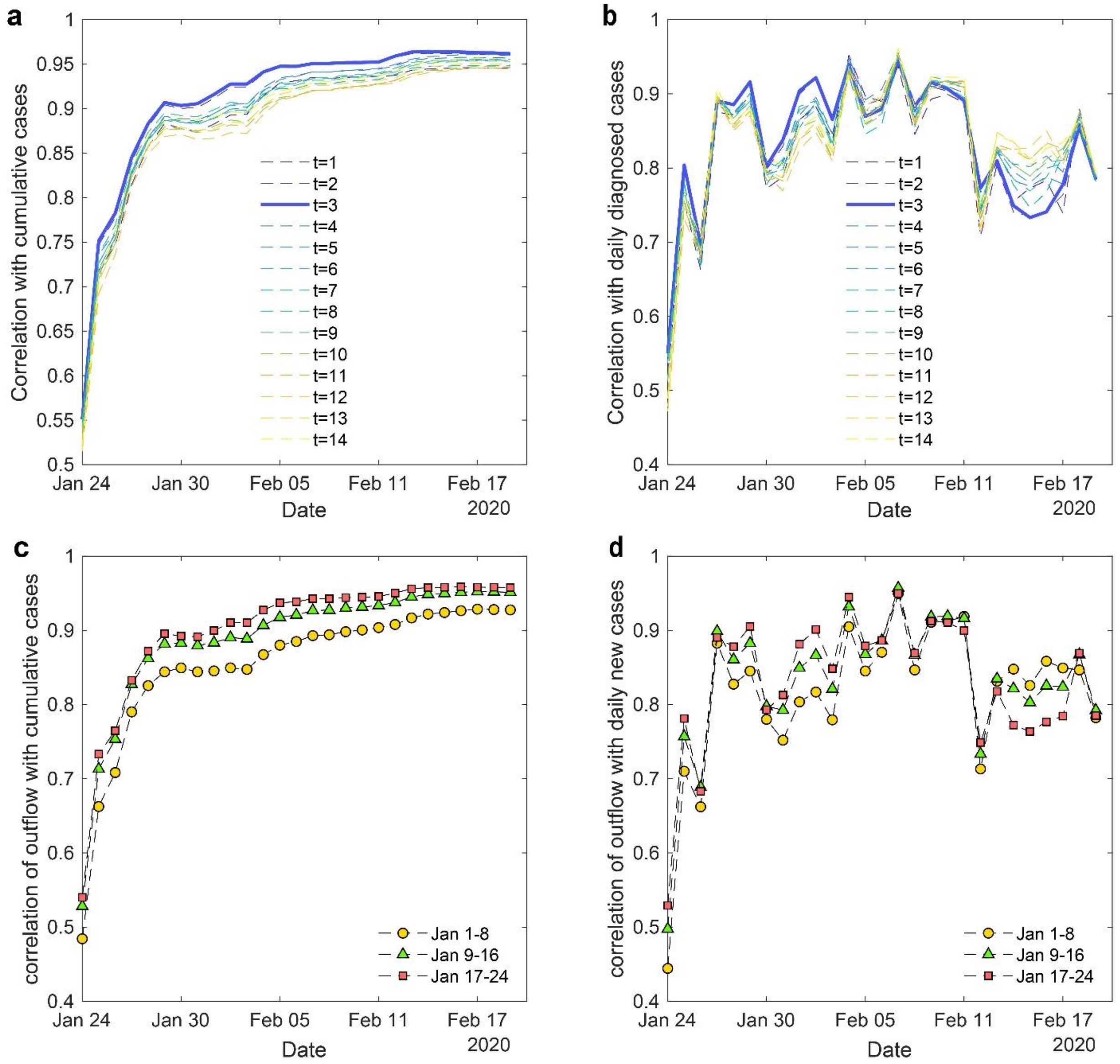
Author Contributions. All authors made equal contributions to the paper. JSJ, JMJ, and NAC conceived the research. JMJ, YY, XL, JSJ, and GX analysed the data. JSJ and NAC wrote the paper. JMJ, JSJ, and XL obtained funding. All authors contributed to research design, analytic development, and critical revisions.

Competing Interests. The authors declare no competing interests.

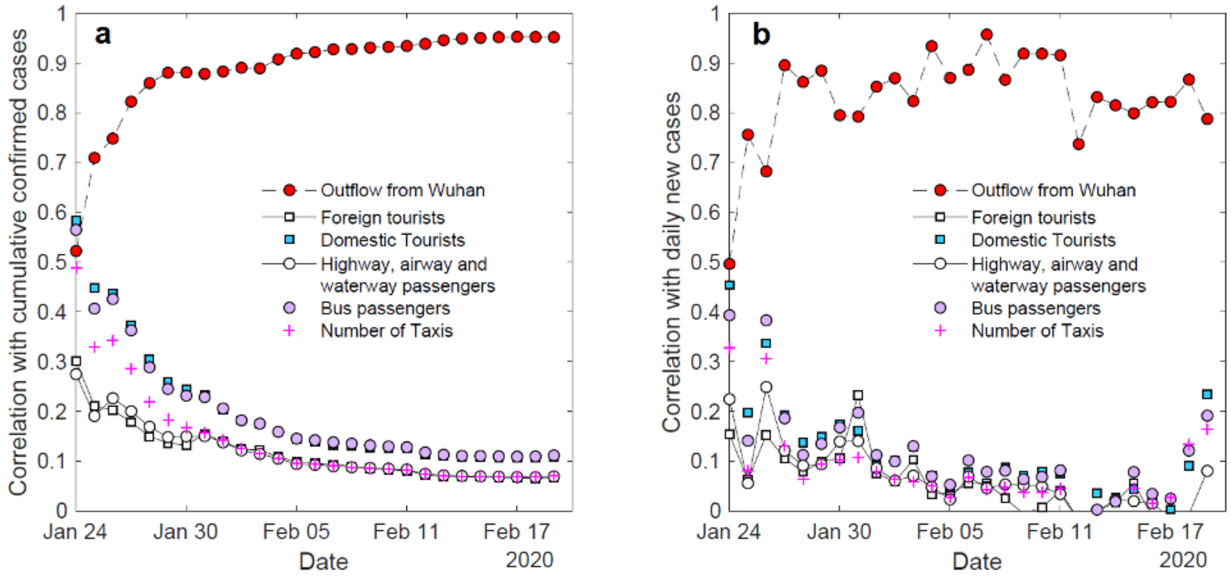
Correspondence. Correspondence regarding this article should be addressed to Jianmin Jia (jmjia@cuhk.edu.sz)

Data and Code Availability Statement. Data and code necessary to reproduce the primary results of this study are included in this published article.

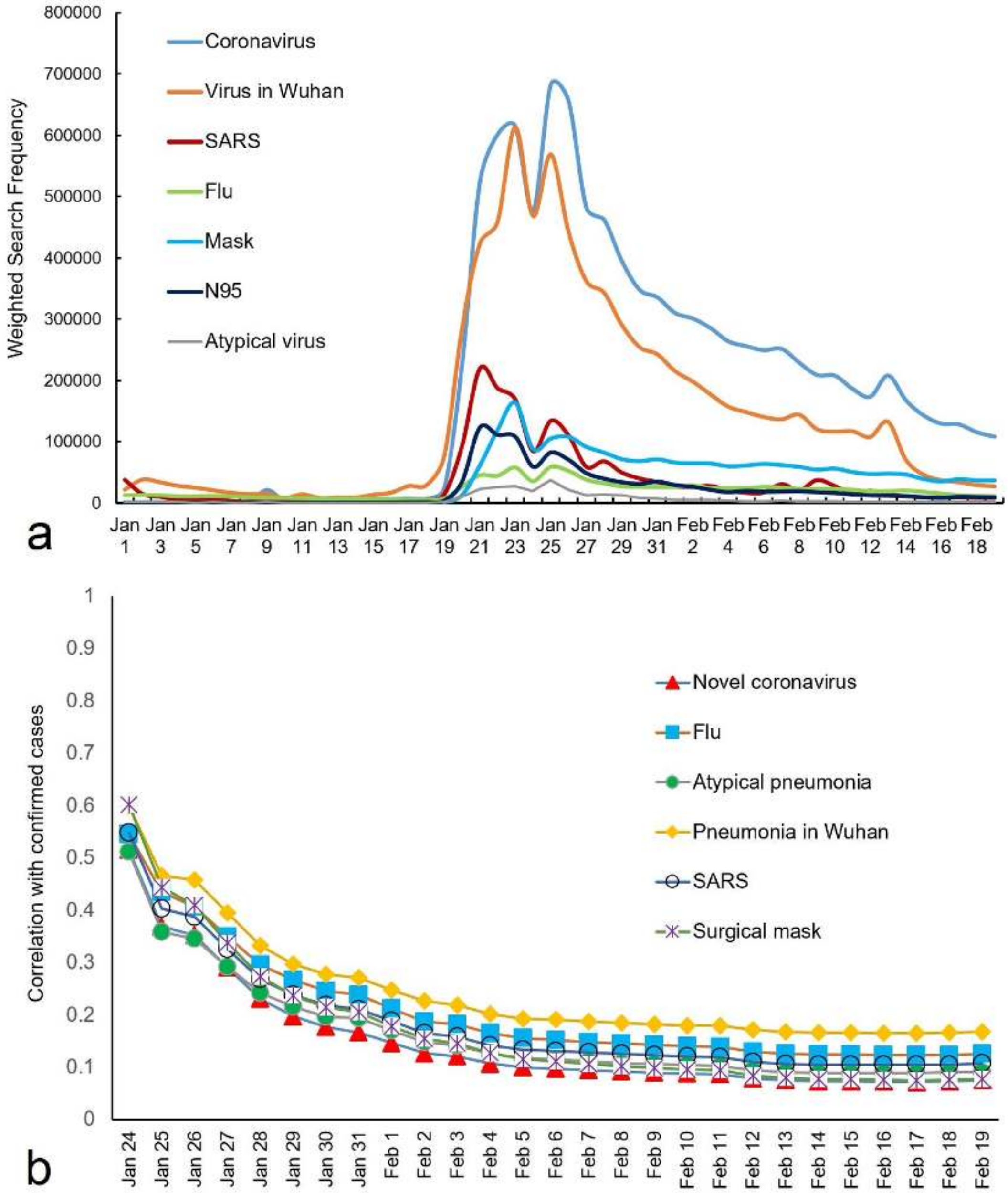
Extended Data Figures



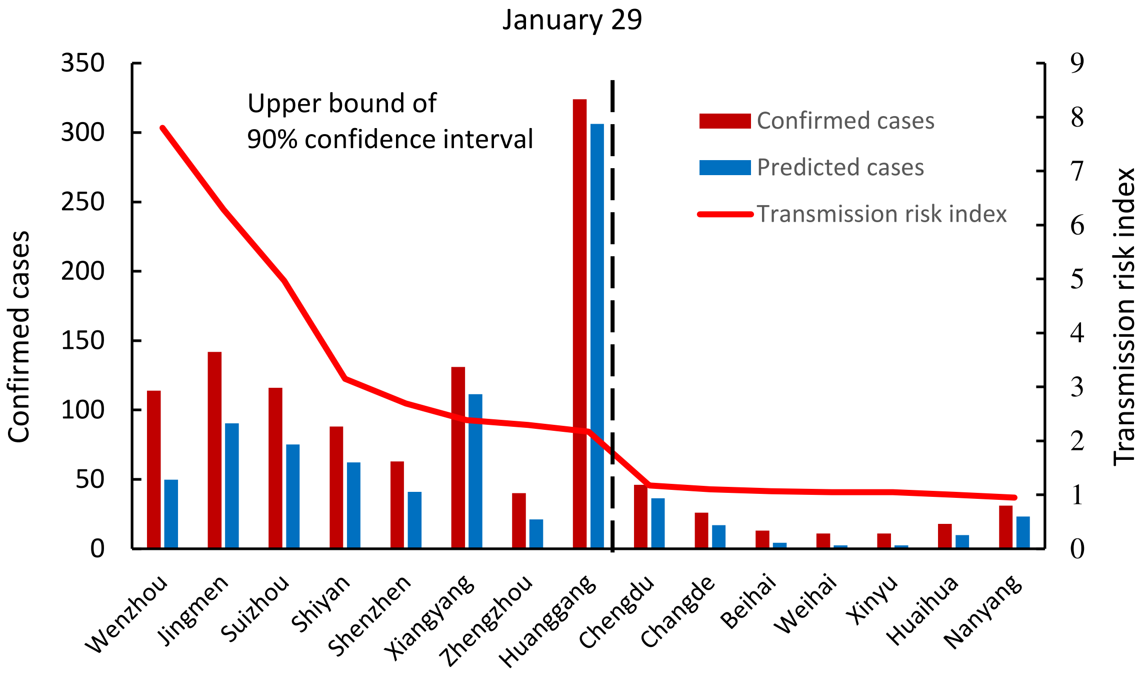
Extended Data Figure 1 | Time window sensitivity test for correlational analysis. **a, b,** Pearson's correlation (N = 296 prefectures) between *daily* population outflow from Wuhan on different days ranging from 1 to 14 days before January 24, with (**a**) the cumulative number of diagnosed cases over time, and (**b**) the number of newly diagnosed (daily) cases over time; e.g., *t* = 3 indicates that the correlation is between daily outflow from Wuhan on January 21 with (**a**) cumulative or (**b**) daily confirmed cases from January 24 onwards. **c**, **d**, Pearson's correlation (N = 296 prefectures) between population outflow during three different (8-day) time periods from January 1 to 24, 2020 and (**c**) the cumulative number of diagnosed cases over time, and (**d**) the number of newly diagnosed (daily) cases over time.



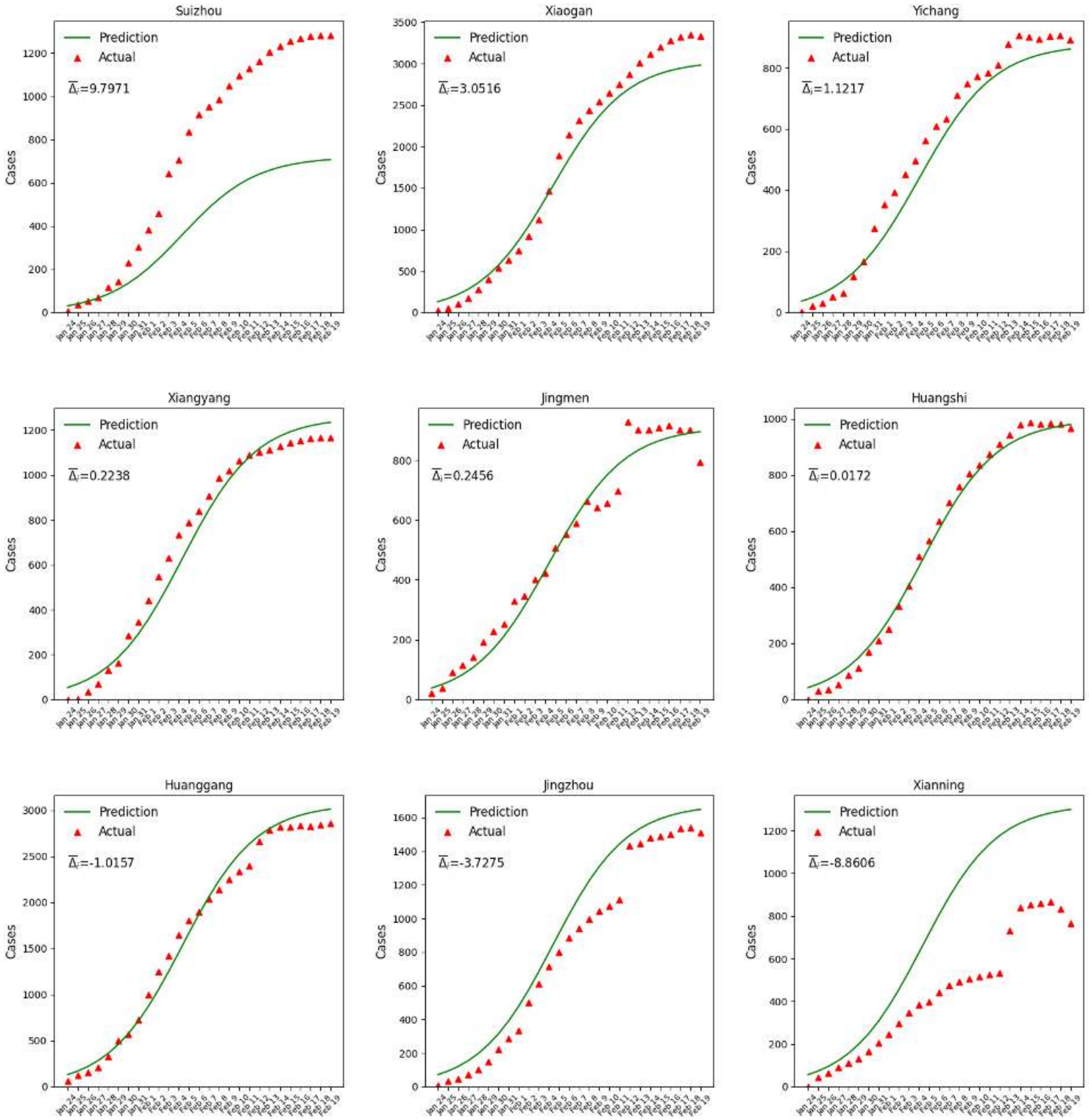
Extended Data Figure 2 | Correlation with alternative population movement measures. Pearson's correlation (N = 296 prefectures) between alternative publicly available movement measurements from the 2018 City/Prefectures Statistical Year Book of China (with aggregate population outflow from Wuhan from January 1-24, 2020 as a reference) and COVID-19 count using (**a**) cumulative confirmed cases over time, and (**b**) for daily confirmed cases over time. Foreign tourist, domestic tourist, and "highway, airway, and waterway passenger" numbers reflect inter-prefecture travel, while bus passengers and number of taxis reflect local travel.



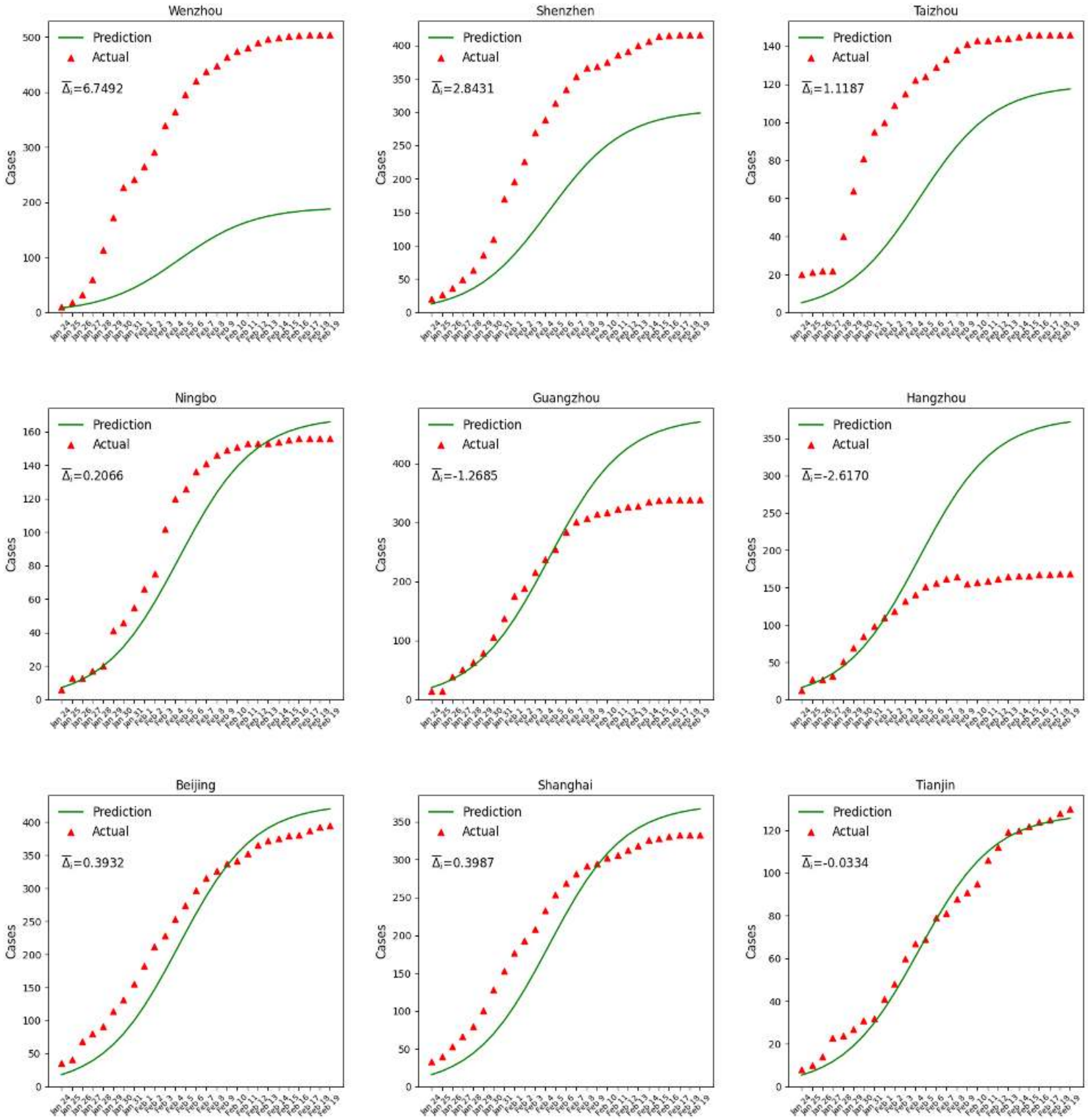
Extended Data Fig. 3 | Search terms and correlation with confirmed cases. a, Search frequency of Baidu search terms related to the COVID-19 outbreak: the search terms are direct translations of the Chinese keywords Baidu users used during the study period (note the official WHO name "COVID-19" was only announced on February 11). **b**, Pearson's correlation (*n* = 296 prefectures) between Baidu search terms and the (cumulative) number of confirmed cases of COVID-19 over time. The initially high and then decreasing predictive strength of search may reflect the fact that, initially, high volumes of information search about the virus signalled stronger risk perception in any given prefecture (for example, because of early reported cases, having more relatives in Wuhan, and so on), but that—over time—information saturation reduced the impetus for specific searches.



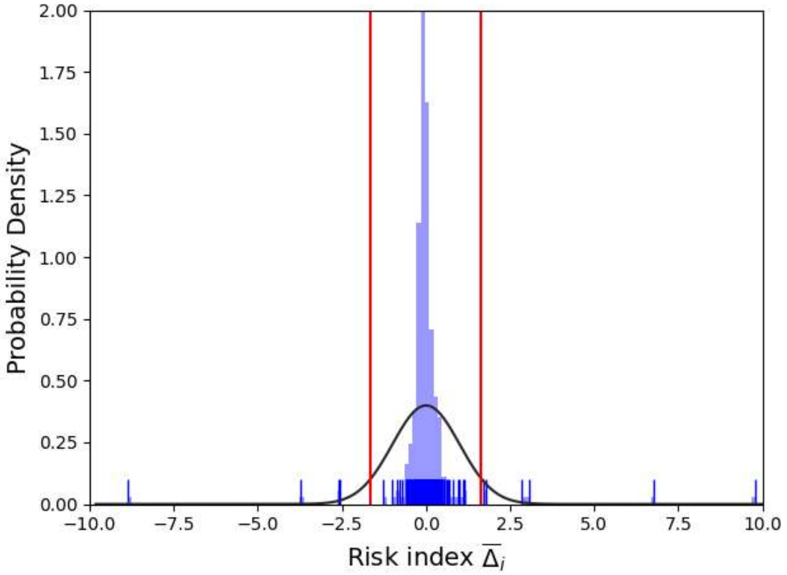
Supplementary Figure 4 | Prefectures with high transmission risk index on January 29, 2020. The predicted structure of the spread of the SARS-CoV-2 virus can be used as a benchmark to identify which locales deviate significantly. Since model (1) predicts the number of cases in a prefecture based on the population outflow from Wuhan (i.e., imported cases and the initial local community transmission of the virus), a greater difference between predicted and confirmed cases suggests a higher level of community transmission. Prefectures to the left of the dashed line have community transmission risk index values over the upper bound of the 90% confidence interval. Our model identified Wenzhou as having the most severe community transmission risk on January 29, 2020. The government announced a full quarantine of the prefecture on February 2, 2020.



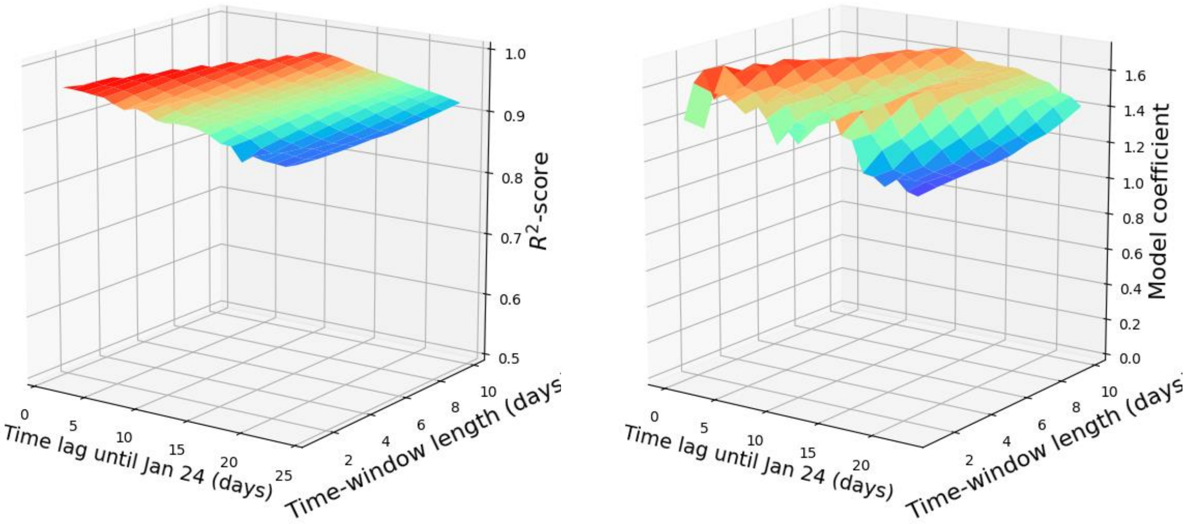
Extended Data Figure 5 | Benchmark (predicted) versus actual virus growth in Hubei's prefectures. Model (2) used aggregate population outflow from Wuhan from January 1-24, 2020 to provide a reference growth *pattern* (i.e., epidemic curves) for COVID-19's spread across time and space, without making *a priori* assumptions of growth pattern or mechanism. Differences in the growth trends between predicted and confirmed cases can signal higher levels of COVID-19 community transmission (Supplementary Table 11). The discrete jumps in confirmed cases in some prefectures after Feb 13 reflected a change in the local governments' infection count criteria; clinically diagnosed cases came to be included in total confirmed case counts in those prefectures (within Hubei province).



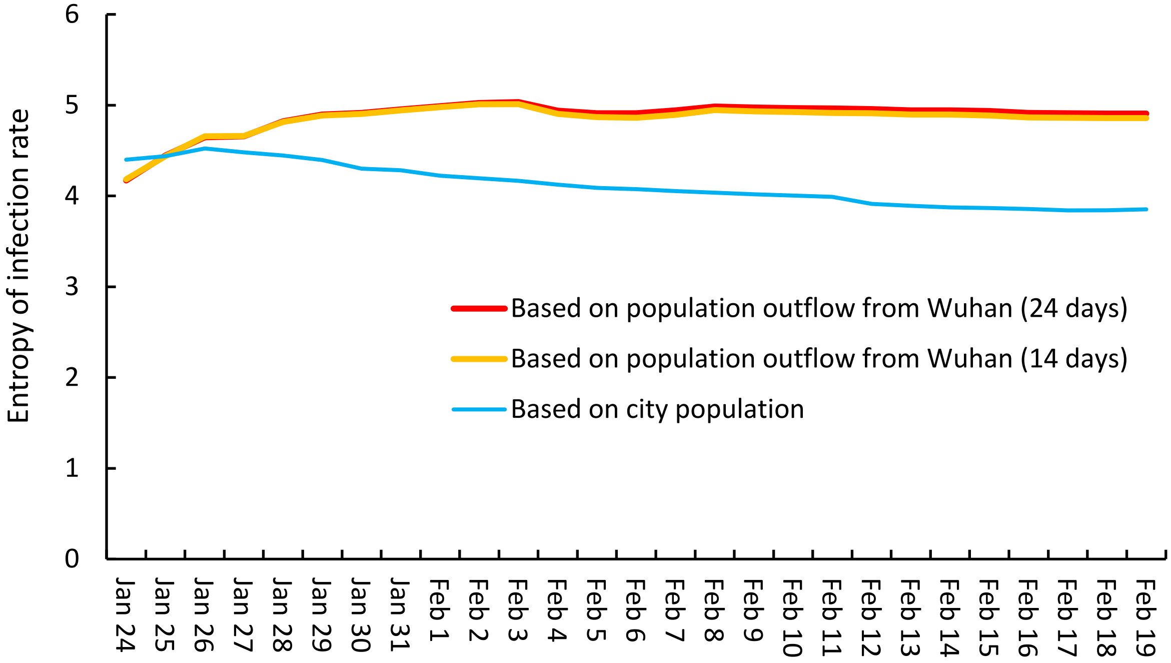
Extended Data Figure 6 | Benchmark (predicted) versus actual virus growth in selected prefectures outside of Hubei. Model (2) used aggregate population outflow from Wuhan from January 1-24, 2020 to provide a reference growth *pattern* (i.e., epidemic curves) for COVID-19's spread across time and space, without making *a priori* assumptions of growth pattern or mechanism. Differences in the growth trends between predicted and confirmed cases can signal higher levels of COVID-19 community transmission (Supplementary Table 11).



Extended Data Figure 7 | The distribution of transmission risk index $\overline{\Delta}_i$. The transmission risk index is the normalized score of the integral of the differences between actual confirmed infection cases and predicted numbers in our model. Prefectures above the 90% confidence interval of the index are likely experiencing more local community transmission than imported cases, and prefectures below the 90% confidence interval may have a better performance in the control of the virus (see Supplementary Table 11).



Extended Data Figure 8 | Robustness check of model (2) with different time lags and time window lengths. We explore which time window and time lags of aggregate population outflow best explain the spread and intensity of COVID-19. "Time window" refers to how many days of outflow data were used; "time lag" (0 to 23) is how many days before January 24 the time window starts. For example, time lag = 1 and time window = 2 is using outflow data between January 23-24. The surfaces show that a more recent time lag improves (**a**) the R^2 as well as (**b**) the parameter value of the population outflow coefficient in model (2).



Extended Data Figure 9 | Entropy of three forms of incidence rate over time. The entropy of incidence rates based on the aggregate population outflow from Wuhan from January 1 to 24, 2020 increased over time, i.e., became increasingly uniform among different prefectures in China, especially during the first week. The entropy curves based on population outflow for 24 days or 14 days before Jan 24 have an almost identical pattern. The entropy of incidence rates based on each prefecture's total population declined over time (except over the first few days); that is population-based incidence rates exhibit no such uniformity.

Supplementary Information

Population Flow Drives Spatio-Temporal Distribution of COVID-19 in China

Jayson S. Jia[1], Xin Lu[2, 3], Yun Yuan[4], Ge Xu[5], Jianmin Jia[6,7*], Nicholas A. Christakis[8]

[1]Faculty of Business and Economics, The University of Hong Kong, Hong Kong SAR, China.

[2]College of Systems Engineering, National University of Defense Technology, Changsha, China

[3]Department of Public Health Sciences, Karolinska Institutet, Stockholm, Sweden

[4]School of Economics and Management, Southwest Jiaotong University, Chengdu, China.

[5]School of Management, Hunan University of Technology and Business, Changsha, China

[6]Shenzhen Finance Institute, School of Management and Economics, The Chinese University of Hong Kong, Shenzhen, China.

[7]Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen, China.

[8]Yale Institute for Network Science, Yale University, New Haven, CT, U.S.A.

[*]Corresponding author. Email: jmjia@cuhk.edu.cn (JJ)

Table of Contents

1. Data Description

1.1 Mobility outflow data

The penetration rate of mobile phone usage among the population in China between the ages of 15 and 65 is almost 100%, and prior estimates suggest such mobile phone users are representative of the whole population in this age range.[30-31]

The population flow data we use, provided by one of the three national mobile carriers in China, was aggregated from the records of mobile phone activities (including geolocation) of all their users, nationwide. A user only had to have their phone on (and not necessarily use it) for their locations to be noted. The number of trips made by users who moved from Wuhan to other prefectures (including commuters) during the period of January 1-24, 2020, was aggregated to produce a national-level population outflow matrix (of total movements from Wuhan in the whole period to each other prefecture). To exclude the large number of users who only briefly transited through Wuhan, users who stayed in Wuhan less than two hours were filtered out of the data. The study period coincided with the run up to the annual *chunyun* mass migration (which can involve as many as 3 billion trips[32]) which culminated on Chinese Lunar New Year eve on January 24, 2020.

Daily number of trips from Wuhan to other prefectures is the final form of the primary dataset provided by the operator. This was supplemented by counts of daily number of trips from Hubei province as a whole (excluding Wuhan) to other prefectures, as well as counts of movements of different types of Wuhan populations (see Section 1.2). We use these additional data as robustness checks for our risk source model (see Section 5). All data were anonymized and aggregated. No personally identifying information was processed in our analyses here. Our aggregate population flow data is made publicly available with this paper, as is our analytic code (in Section 5 below), as part of a replication package.

The carrier provided two separate measurements of population outflow. One was based on the observed movements of the carrier's own customers. The other was an extrapolated measure of movements of the whole population.

To produce a representative estimate for the total number of trips made by entire population, the number of trips made by the carrier's own users was extrapolated to the whole network, with a variety of factors being considered. Specifically, information about user coverage, ratio of calls and messages with other operators, and information about users' age and gender, which was combined and modeled using a machine learning approach by the operator, generating estimates of the total number of users leaving Wuhan and going to each destination prefecture, not just the count of the people moving who were customers of this carrier. This extrapolation of counts was then *validated with real coverage ratios (i.e., of the two other carriers) in several prefectures, as well as all provinces in China*, documenting high accuracy in predicting the number of users from the whole network with data from this one operator.

Our evaluations (below) show both measurements produce nearly identical results. And the observed correlation (in our analysis) between both measurements of population outflow with the independently observed COVID-19 cases serves as a further check of validity.

We also note that our primary analyses rely on *relative* population flows, not actual numbers of people moving from Wuhan to each destination. In other words, our methodology does not require observing the movements of the entire population. The reason for this is that our primary analyses use normalized variables – that is, expresses them as a percentage of total population outflow. This methodological feature is advantageous since researchers or policy makers may use our approach even if they only have access to a dataset that is a subsample of the population, as long as that dataset is representative of the *relative* distribution and quantity of population flow.

While we use phone data to compute total aggregate *flows* between prefectures, there is also a rich tradition of using individual-level data to track human mobility and population movements in humans[33-38], including to study disease transmission.[39-41]

1.2 Different types of population outflow data

1.2.1 Own customer versus extrapolated data

We use the measurement of population outflow based on the extrapolated measure in our primary analyses so that we are able to discuss *total population* movements (particularly with respect to discussing the impact of the quarantine), and use the carrier's own customer count measure for robustness checks. Since our primary analyses uses the *relative distribution* (rather than absolute quantity) of population outflow, both measurements typically yield the same results (see Section 5).

1.2.2 Resident versus non-resident inflow

In addition, the carrier provided separate counts of non-Wuhan-residents who were *returning* to their home prefectures (as opposed to being Wuhan residents who were visiting other prefectures). These 'returnees' (which we refer to as "returning resident outflow") were people who were returning to the prefecture where they receive local telecommunications service (which is charged at a lower preferential rate, and also where their registered home address is) after visiting Wuhan (for at least 2 hours) between January 1 to 24. Returning residents likely have a different impact on local disease transmissions dynamics compared to non-residents. Since public health policies effectively quarantined (or encouraged self-quarantine of) individuals within their own households, infected returning residents (who by definition are returning to their own homes) may be more likely contribute to family and community infection clusters compared to infected visitors (since not all of them are staying with relatives). Indeed, 78-85% of infection clusters occurred within families. [42-46] On the other hand, visitors may be more likely to visit public areas (e.g., hotels and tourist destinations). We evaluated this measure in Section 5.

1.2.3 Population outflow from Hubei excluding Wuhan

Finally, the carrier provided counts of population outflow from other prefectures in Hubei (excluding Wuhan). Hubei had the largest share of COVID-19 infections in China and prefectures in Hubei received the highest total population inflow from Wuhan in our study period. Indeed, 75.66% of Wuhan population outflow went to other prefectures within Hubei province, which had 56.43% of total COVID-19 cases in China outside of Wuhan. Consequently, we also later evaluated whether population outflow from these prefectures brought COVID-19 export risk to prefectures outside of Hubei province (see Section 5). Previewing briefly, we found that the effect of population outflow from non-Wuhan Hubei prefectures on COVID-19 spread is much more limited compared to the impact of population outflow from Wuhan prefecture itself (see Section 5).

1.3 Merge with data regarding 296 prefectures

In later modeling analyses, we merge outflow data with demographic and economic variables from the *2018 City/Prefectures Statistical Year Book of China*, which serve as co-variates and control variables. This yields a final sample of 296 prefectures included in our analysis (Wuhan prefecture is not included; the relatively new Sansha prefecture is excluded since it is a set of islands with a population of only 448). These prefectures have an average population of 4.40 million and a total population of 1.308 billion, representing 94.07% of the country's total population.

In our analysis, we combined the merged dataset of human mobility and economic variables with the daily count and geographical location of confirmed cases of COVID-19 nationwide, provided by the Chinese Center of Disease Control and Prevention (CCDC), which used a consistent (and stringently enforced) procedure for case ascertainment and reporting. We report data from January 24 to February 19. After February 19, more than 90% of prefectures outside of Hubei province often reported zero infection cases daily.

Since our analyses focus on the effect of population outflow from Wuhan on COVID-19 outbreaks elsewhere, we do not use or require infection counts from Wuhan. Thus, any inaccurate counts (e.g., due to Wuhan's hospitals being overwhelmed) or changes in reporting methodology in Wuhan do not affect our analysis.

2. Correlational Analysis

2.1 Sensitivity tests for the correlational analysis

In the primary analysis, we used the aggregate population outflow data of the 24 days between January 1 to January 24 to document how the correlation between total confirmed cases and the population outflow from Wuhan to different prefectures increased over time (see Fig. 2c and 2d in main text). Here, we document the correlation between population outflow from Wuhan on different days ranging from 1 to 14 days before January 24, with the number of diagnosed cases over time (Extended Data Figure 1a, b). The Wuhan population outflow on the four most recent days prior to January 24 (i.e., January 21 through 24) is the most highly correlated, over time, with the number of local confirmed cases in different areas. This may relate to the large movement of people out of Wuhan right before the quarantine was imposed and the latency of the virus (which has an observed incubation period of 2-14 days, with a mean = 4-5 days[45-47]).

As another robustness check of potential bias generated by selection of periods for calculating total population outflow and confirmed cases, we repeat the correlational analysis after dividing the observation period into periods. Specifically, we first equally divide the 24 days (between January 1 and 24) into three 8-day periods; then, with the outflows from these different periods, we calculate correlation between outflow, and the cumulative confirmed cases (Extended Data Figure 1c) or newly diagnosed cases (Extended Data Figure 1d) over the period of January 24 to February 19.

We can see that, first of all, there is limited difference in the correlation coefficients when outflows are calculated from different periods, implying that the general pattern of movement did not change significantly in January, i.e., during the Spring Festival Travel period. Secondly, more recent outflow had higher correlation with the total number of confirmed cases after January 24, 2020, indicating that the spatial-temporal pattern of population flow close to Chinese New Year Eve (January 24) is more relevant for the prediction of new infections outside Wuhan. Lastly, for all three periods, the correlation between outflow and daily number of new cases increases rapidly from roughly 0.5 for the first day to about 0.9 four days after.

2.2 Correlations with alternative population movement measures

For comparison purposes, we also examine the correlation between each prefecture's infection count and the prefecture's intra- and inter- prefectural movement data from the 2018 City/Prefectures Statistical Year Book of China (Extended Data Figure 2). All correlations decline over time, particularly towards the end of the study period when >90% of prefectures had no new daily cases. Only the outflow from Wuhan is highly correlated with cumulative and daily confirmed cases.

3. Risk Model Based on Observed Population Flows From Outbreak Epicenter

3.1 Analysis based on daily infection data

3.1.1 Model specification

The correlational analysis in Fig. 2c suggested that other factors, such as GDP and local population, are correlated with the number of confirmed cases (although they have declining predictive power over time). To test these and other factors as control variables and as alternative variables against the aggregate Wuhan population outflow, we consider two functional forms: an exponential model (1) as our basic model, and a power model (2) as a robustness check. As illustrated in Figures 2a, 2b (log scale), these functional forms are well-suited for our data. We also include fixed effects to control for provincial differences.

$$y_i = c \prod_{j=1}^{m} e^{\beta_j x_{ji}} \, e^{\sum_{k=1}^{n} \lambda_k I_{ik}} \tag{1}$$

$$y_i = c \prod_{j=1}^{m} x_{ji}^{\beta_j} \, e^{\sum_{k=1}^{n} \lambda_k I_{ik}} \tag{2}$$

where y_i is the number of cumulative (or daily) confirmed cases in prefecture i, x_{1i} is cumulative aggregate population outflow from Wuhan (i.e., total outflow between January 1-24) to prefecture i, $x_{2\mathrm{i}}$ is the GDP of prefecture i, $x_{3\mathrm{i}}$ is the population size of prefecture i, m is the number of variables included, and c and β_j are parameter estimates. Of course, other variables could be included to test other factors of interest. In the above models, λ_k is the fixed effect for province k, n is the number of prefectures considered in the analysis, I_{ik} is a dummy for prefecture i and $I_{ik} = 1$, if $i \in k$ (prefecture i belongs to province k), otherwise $I_{ik} = 0$.

Compared with gravity models, we use outflow to replace distance, which provides a better measure of interactions between Wuhan and other prefectures (see Section 4 for more detailed discussion). Since our models capture flow of human population from the epidemic risk source explicitly, yielding risk for the destination location, we name our models “risk source” models. When taking the log on the both sides of the models, we have the following generalized linear models:

$$\log(y_i) = \log(c) + \sum_{j=1}^{m} \beta_j x_{ji} + \sum_{k=1}^{n} \lambda_k I_{ik} \qquad (3)$$

$$\log(y_i) = \log(c) + \sum_{j=1}^{m} \beta_j \log(x_{ji}) + \sum_{k=1}^{n} \lambda_k I_{ik} \qquad (4)$$

Thus, model (3) becomes a Poisson regression model, implying that y_i is assumed to follow a Poisson distribution. Model (4) is another type of Poisson regression model with the logarithm transformation of the independent variables x_{ji}. Poisson regression models are typically used for modeling count data – here, in this context, the number of confirmed infection cases.

To aid comparability and eliminate differences in units, we normalize the data in two ways: for the exponential (1), we normalize data first by taking the log, $x'_{ji} = \log(x_{ji})$ (the of the log-log plot in Figure 2a and b suggests a logarithmic transformation is appropriate), and then by standardizing $x''_{ji} = (x'_{ji} - Mean) / Std$; for the power model (2), we normalize the data by $x'_{ji} = x_{ji} / \sum_{i=1}^{n} x_{ji}$, where n is the number of prefectures, which basically uses the fraction of population outflow from Wuhan to each prefecture and fractions of GDP and populations in the estimation.

Importantly, we can add more risk source flows (from other candidate epicenters of the epidemic) into the models easily, if we want, such as outflow from Hubei (excluding Wuhan).

3.1.2 Model estimation and results

To conduct statistical analyses for our proposed models, we merged the population outflow data with population and GDP data for each prefecture from the *2018 China City (Prefecture) Statistical Year Book* (produced by the National Bureau of Statistics of China). We excluded the epicenter, Wuhan, from this analysis and also the island prefecture of Sansha (with a population of only 448 people and no GDP data). Hence, 296 prefectures are included in our analyses.

Although the logarithmic transformations of models (1) and (2) could be estimated using a linear regression approach, we need a consistent estimation method that can also be applied to the nonlinear models used here, and for estimating the multiplicative forms of models (1) and (2) directly. In machine learning, the steepest gradient algorithm, the Newton algorithm and the Levenberg–Marquardt algorithm, are commonly used in solving nonlinear least squares problems. In particular, the Levenberg–Marquardt (LM) algorithm is a combination of the steepest gradient algorithm and the Newton algorithm, which has the advantages of both[48-49]. This method has been widely applied for solving various nonlinear problems, and it is used to estimate parameters for our models. The results are provided in Supplementary Table 1.

We first fit the models (1) and (2) by using the single variable of aggregate population outflow from Wuhan (from January 1 to 24) to other prefectures. The model fit gradually improved from $R^2 = 0.772$ for both models (1) and (2) on January 24 to $R^2 = 0.946$ for both models on February 19. As shown in Supplementary Table 1, when we include prefecture population and GDP into the models, the fit improves a certain degree ($R^2 = 0.864$ on January 24 and $R^2 = 0.965$ on February 19 for both models). Over time, the key parameter β_1 for the outflow from Wuhan becomes an increasingly dominant predictor, while parameters β_2 and β_3 decline over time (i.e., prefecture GDP and local population become increasingly less predictive of number of confirmed cases over time). But, in general, these parameters are quite stable over time, especially after the first few days as the number of observed cases increases. The estimated values for these parameters suggest that population outflow from Wuhan to other prefectures is much more important than prefecture GDP and local population size in predicting confirmed cases (see Supplementary Table 1). The scaling factor a increases as the number of confirmed cases increases over time.

The possibly less accurate and timely infection count during the earlier dates (as case reporting systems were still being set up) may possibly have contributed to the lower initial fit of our model. Several studies show that COVID-19 has a median incubation time period of about 4.75 days (interquartile range: 3.0-7.2) that can extend up to 14 days[1-3]. Thus, the parameter estimates from the first few days of our model may be less interpretable. Indeed, our model fit improves even more after 14 days, by which time the incubation period was over for most latent infection cases.

Date	R^2	c	β_1	β_2	β_3
24-Jan	0.809	0.425	1.609	-0.008	0.159
25-Jan	0.858	0.161	1.392	-0.289	1.356
26-Jan	0.871	0.631	1.131	-0.144	0.608
27-Jan	0.902	1.355	1.035	-0.030	0.380
28-Jan	0.933	1.974	1.070	-0.012	0.289
29-Jan	0.933	2.502	1.089	-0.025	0.336
30-Jan	0.938	2.244	1.156	-0.053	0.398

31-Jan	0.939	3.375	1.051	0.062	0.338
1-Feb	0.943	3.923	0.995	0.100	0.359
2-Feb	0.937	2.324	0.991	0.054	0.406
3-Feb	0.953	4.355	0.949	0.041	0.479
4-Feb	0.948	4.926	0.932	0.032	0.457
5-Feb	0.954	4.913	1.056	0.062	0.441
6-Feb	0.947	4.921	1.197	0.103	0.397
7-Feb	0.944	4.577	1.243	0.137	0.369
8-Feb	0.948	4.842	1.285	0.166	0.348
9-Feb	0.949	4.727	1.283	0.215	0.341
10-Feb	0.949	4.886	1.309	0.226	0.327
11-Feb	0.948	4.506	1.330	0.245	0.316
12-Feb	0.948	4.540	1.350	0.255	0.300
13-Feb	0.948	4.422	1.402	0.291	0.260
14-Feb	0.966	3.885	1.294	0.169	0.268
15-Feb	0.970	4.766	1.300	0.159	0.244
16-Feb	0.969	4.985	1.331	0.179	0.234
17-Feb	0.968	5.349	1.364	0.199	0.224
18-Feb	0.968	4.425	1.388	0.230	0.214
19-Feb	0.967	4.420	1.406	0.249	0.218

Supplementary Table 1 | Results for exponential model (1) using cumulative case count on each date.

Overall, our models' performance continuously improved as more infection cases were confirmed. It is noteworthy and significant that population flow up to January 24 can predict the final distribution pattern of virus spread over different geographical locales two weeks later (and with increasing accuracy). This suggests that the spread did not grow exponentially and that the overall spread pattern is basically governed – at least in early stages of the epidemic – by the structure of population outflow from Wuhan to other prefectures. We make use of this logic to infer the incidence of community transmission in various areas.

Supplementary Table 2 provides the estimation results for the power model (2), which exhibits the same pattern of results as model (1); data fit gradually improved from R^2 = 0.623 on January 24 to R^2 = 0.956 on February 19. These parameters are quite stable over time, especially after the first few days, as the number of observed cases increase. The key parameter β_1 for the outflow from Wuhan is quite stable over time, while parameters β_2 and β_3 have a declining pattern over time.

Date	R^2	c	β_1	β_2	β_3
24-Jan	0.864	1.636	0.777	-0.385	1.372
25-Jan	0.869	1.682	0.524	-0.206	0.707
26-Jan	0.905	2.914	0.473	-0.073	0.472
27-Jan	0.935	3.023	0.490	-0.046	0.367
28-Jan	0.937	3.182	0.498	-0.046	0.432
29-Jan	0.939	3.196	0.522	-0.067	0.508
30-Jan	0.940	3.191	0.495	0.032	0.434

31-Jan	0.944	4.845	0.470	0.049	0.462
1-Feb	0.938	5.275	0.465	0.025	0.526
2-Feb	0.954	5.322	0.446	0.021	0.618
3-Feb	0.949	7.142	0.439	0.017	0.589
4-Feb	0.955	6.065	0.499	0.034	0.567
5-Feb	0.948	6.717	0.567	0.053	0.510
6-Feb	0.945	7.145	0.588	0.067	0.473
7-Feb	0.949	6.754	0.607	0.078	0.446
8-Feb	0.950	8.166	0.605	0.098	0.436
9-Feb	0.950	6.012	0.617	0.102	0.417
10-Feb	0.949	6.438	0.627	0.110	0.404
11-Feb	0.949	7.696	0.636	0.114	0.382
12-Feb	0.959	8.215	0.617	0.107	0.373
13-Feb	0.967	7.814	0.611	0.079	0.342
14-Feb	0.970	7.459	0.614	0.076	0.311
15-Feb	0.969	8.653	0.628	0.083	0.297
16-Feb	0.969	8.509	0.643	0.092	0.285
17-Feb	0.969	8.041	0.654	0.104	0.270
18-Feb	0.968	8.171	0.662	0.111	0.275
19-Feb	0.965	8.661	0.684	0.126	0.287

Supplementary Table 2 | Results power model (2) using cumulative case count on each date.

3.1.3 Daily model estimation and results

As a robustness check, we consider *daily* cases (i.e., new cases) on each date as the dependent variable in models (1) and (2). Tables 3 and 4 provide estimation results for the exponential model (1) and power model (2), respectively, for *daily* cases. As can be seen, the outflow variable continues to play a dominating role (compared with population and GDP) in predicting daily new cases for most of the days. Although the overall model fitting is still good, the results for daily cases are more variable than the cumulative COVID19 cases. The differences and fluctuations in R^2 are akin to as the fluctuations in Fig. 2b (compared to Fig. 2a); there were unsmooth fluctuations in case count which may be caused by natural fluctuations or lags in reporting (e.g., from case overload). A cumulative dependent variable smooths these fluctuations. Nonetheless, this model has high explanatory power until the last two days of the study, when there were very few new infections in China outside of Wuhan (the same reason correlation declines precipitously in Fig. 2b).

Date	R^2	c	β_1	β_2	β_3
24-Jan	0.833	0.012	2.407	-0.880	3.645
25-Jan	0.824	1.822	-0.605	-1.179	0.321
26-Jan	0.781	1.223	0.918	0.245	0.200
27-Jan	0.898	1.235	1.335	0.110	0.126
28-Jan	0.872	1.162	1.156	-0.035	0.455
29-Jan	0.907	1.183	1.263	-0.153	0.587

30-Jan	0.868	1.965	1.144	0.495	0.243
31-Jan	0.838	1.737	0.845	0.153	0.426
1-Feb	0.867	2.060	0.882	-0.159	0.725
2-Feb	0.939	2.605	0.753	0.010	0.876
3-Feb	0.828	2.079	0.933	0.038	0.375
4-Feb	0.922	0.146	3.094	0.839	0.503
5-Feb	0.833	8.779	6.960	2.406	0.830
6-Feb	0.824	0.600	1.984	0.354	0.195
7-Feb	0.952	0.605	2.000	0.350	0.100
8-Feb	0.922	2.145	1.341	0.506	0.199
9-Feb	0.831	0.414	2.083	0.334	0.009
10-Feb	0.904	0.949	2.145	0.519	0.089
11-Feb	0.900	1.023	1.936	0.276	-0.071
12-Feb	0.626	2.536	1.077	0.067	0.154
13-Feb	0.895	0.205	2.108	-3.599	-1.631
14-Feb	0.805	1.817	1.359	-1.652	-0.988
15-Feb	0.880	0.000	13.940	6.048	0.803
16-Feb	0.924	0.000	12.609	5.230	0.535
17-Feb	0.771	0.004	4.631	1.879	-0.220
18-Feb	0.366	0.000	9.265	3.138	2.359
19-Feb	0.350	-1.761	-0.774	-1.008	-0.369

Supplementary Table 3 | Results for exponential model (1) using daily new case count on each date.

Date	R^2	c	β_1	β_2	β_3
24-Jan	0.833	0.142	1.092	-0.733	4.735
25-Jan	0.814	5.507	0.000	-0.634	0.331
26-Jan	0.781	4.284	0.417	0.204	0.259
27-Jan	0.898	3.166	0.606	0.091	0.164
28-Jan	0.872	6.359	0.524	-0.029	0.591
29-Jan	0.907	4.686	0.573	-0.128	0.762
30-Jan	0.868	4.780	0.520	0.412	0.315

31-Jan	0.838	7.525	0.383	0.127	0.553
1-Feb	0.867	6.686	0.400	-0.132	0.942
2-Feb	0.939	6.944	0.342	0.008	1.137
3-Feb	0.828	6.451	0.423	0.032	0.487
4-Feb	0.922	2.233	1.404	0.698	0.654
5-Feb	0.833	0.219	3.160	2.003	1.078
6-Feb	0.824	4.824	0.900	0.294	0.253
7-Feb	0.952	6.400	0.908	0.292	0.129
8-Feb	0.922	12.387	0.609	0.421	0.259
9-Feb	0.831	2.507	0.946	0.278	0.012
10-Feb	0.904	1.844	0.974	0.432	0.115
11-Feb	0.900	2.978	0.879	0.230	-0.092
12-Feb	0.626	5.101	0.489	0.056	0.200
13-Feb	0.895	0.172	0.957	-2.995	-2.119
14-Feb	0.805	1.637	0.617	-1.374	-1.283
15-Feb	0.880	0.000	6.327	5.033	1.043
16-Feb	0.924	0.001	5.723	4.352	0.695
17-Feb	0.771	0.148	2.102	1.564	-0.286
18-Feb	0.366	0.001	4.206	2.613	3.066
19-Feb	0.323	-1.035	0.000	-0.598	-0.529

Supplementary Table 4 | Results power model (2) using daily new case count on each date.

3.1.4 Estimation for assessing community transmission risk

Since our model predicts the distribution and geographical structure of COVID-19 cases based on the population outflow from Wuhan into different prefectures, the predicted structure of the virus spread can be used as a benchmark to identify which locales deviate significantly. We use the normalized difference of confirmed and predicted cases (subtracting mean and dividing by standard deviation) to create a community transmission risk index.

Since the number of predicted cases in a prefecture is driven almost completely by population inflow from Wuhan in our model, the predicted value is by definition a function of the number of imported cases and the initial local transmissions of the virus. Any difference (delta) between predicted and confirmed cases would suggest a higher level of COVID-19 community transmissions (i.e., spread from infected individuals not from Wuhan). Thus, we regard places that have more confirmed infections than predicted cases as having higher community transmission risks (compared to the benchmark). On the other hand, prefectures with fewer cases than expected by our model are also noteworthy, since they could either have had more successful public health measures or be at higher risk of inaccurate data reporting (which might risk giving local officials a false sense of security).

Since numerous restrictive public health measures (e.g., travel restrictions, social distancing, emphasis on wearing surgical masks outdoors, banning social gatherings, widespread public service announcements, as well as a general wave of fear and paranoia that motivated individuals to avoid going outside) were implemented locally throughout China in the aftermath of the quarantine of Wuhan and Hubei province, the high predictive power of the model for most cases suggests that these measures and factors were generally successful in limiting growth in community transmissions in most locales. Travel limitations were particularly severe: in many prefectures there was literal ‘rationing’ of families’ trips outside of their homes to once per day by one or two family members (usually not outside the neighborhood/district) (this was often enforced by neighborhood watches, using mobile phone scanning, or printed ‘trip passes’). In addition, many transmissions were within family clusters.[42-45] According to the Chinese CDC[47], 83% of transmission so far have happened in family clusters. Since most Chinese families live together in households of 2-3 people, this suggests that number of cases of transmission could be scaled up proportionally, especially to the extent that all prefectures adopted similar infection control measures.

Extended Data Figure 4 identifies prefectures with community transmission risk index values over the upper bound of the 90% confidence interval. Our model identified Wenzhou as having the most severe community transmission risk on Jan. 29 (Wenzhou’s total confirmed cases at that point was far beyond that expected by the model). The government announced a full quarantine of the prefecture on Feb. 2.

3.2 Dynamic model for virus spread and growth

3.2.1 Dynamic model development

Models (1) and (2) provide cross sectional analyses at a daily level – in other words, they provide snapshots by day (using infection counts of a particular day) and do not account for trends across time (though the trends can be described by the constant *c* discretely). We next adapt these models into dynamic models to explore changes in distribution and growth of COVID-19 across all prefectures, over time. By including all reported cases in a single model, we are also able to document the growth of the epidemic over time. A typical growth pattern for epidemiological events follows a sigmoidal pattern[50-53] Indeed, we document such a pattern for infection spread of COVID-19 in the various Chinese prefectures.

To combine our previous daily analysis with time as well as fixed effects to control for provincial differences, we consider a Cox proportional hazards model to integrate our risk source models (1) and (2) with a growth function, which allows the effect of a unit increase in a covariate to be multiplicative with respect to the underlying baseline hazard function as follows:

$$\lambda(t|x_i) = \lambda_0(t)\left(\prod_{j=1}^{m} e^{\beta_j x_{ji}}\right) e^{\sum_{k=1}^{n} \lambda_k I_{ik}} \tag{5}$$

$$\lambda(t|x_i) = \lambda_0(t)\left(\prod_{j=1}^{m} x_{ji}^{\beta_j}\right) e^{\sum_{k=1}^{n} \lambda_k I_{ik}} \tag{6}$$

where $\lambda(t|x_i)$ is the hazard function describing the number of cumulative confirmed cases at time *t* given an population outflow from Wuhan to prefecture *i* and other variables; $\lambda_0(t)$ is the underlying baseline hazard function with *t* = 1 starting from January 24; $x_i = \{x_{1i}, x_{2i}, \dots x_{mi}\}$ are the realized values of the covariates for prefecture *i*; and the other notation is the same as for models (1) and (2). We arrive at these models by using a time-varying function $\lambda_0(t)$ to replace the constant scaling parameter *c* in models (1) and (2) to obtain models (5) and (6). These dynamic models imply that the hazard responds exponentially: each unit increase in x_i results in a proportional scaling of the hazard. In particular, Models (5) and (6) capture the effect of risk source outflow from Wuhan in a spatio-temporal manner.

We consider the three most popular sigmoidal functions: logistic, generalized logistic (also called a Richards model), and Gompertz functions for $\lambda_0(t)$ in the hazard model (5). These curves grow exponentially initially, and then saturates at the later stage. These functions have been used to forecast outbreaks for different infectious diseases[50-53] and yield the following three models with provincial fixed effects:

a) Exponential-Logistic (EL) model: $\lambda(t|x_i) = \frac{\alpha}{1+e^{-\gamma t+\omega}}\left(\prod_{j=1}^{m} e^{\beta_j x_{ji}}\right)e^{\sum_{k=1}^{n}\lambda_k I_{ik}}$ (7)

b) Exponential-Generelized-Logistic (EGL) model:

$$\lambda(t|x_i) = \frac{\alpha}{[1+\mathrm{g}e^{-r(t-\mathrm{t}_i)}]^{\frac{1}{g}}}\left(\prod_{j=1}^{m} e^{\beta_j x_{ji}}\right)e^{\sum_{k=1}^{n}\lambda_k I_{ik}} \quad (8)$$

c) Exponential-Gompertz (EG) model: $\lambda(t|x_i) = \alpha a^{b^t}\left(\prod_{j=1}^{m} e^{\beta_j x_{ji}}\right)e^{\sum_{k=1}^{n}\lambda_k I_{ik}}$ (9)

where α, *γ, ω, a, b,* and *g* are parameters to estimate; and the other notation is the same as models (1) and (2). In a similar fashion, using the three functions for $\lambda_0(t)$ in model (6) leads to the following spatio-temporal models:

d) Power-Logistic (PL) model: $\lambda(t|x_i) = \frac{\alpha}{1+e^{-\gamma t+\omega}}\left(\prod_{j=1}^{m} x_{ji}^{\beta_j}\right) e^{\sum_{k=1}^{n}\lambda_k I_{ik}}$ (10)

e) Power-Generelized-Logistic (EGL) model:

$$\lambda(t|x_i) = \frac{\alpha}{[1+\mathrm{g}e^{-r(t-\mathrm{t}_i)}]^{\frac{1}{g}}}\left(\prod_{j=1}^{m} x_{ji}^{\beta_j}\right) e^{\sum_{k=1}^{n}\lambda_k I_{ik}} \quad (11)$$

f) Power-Gompertz (PG) model: $\lambda(t|x_i) = \alpha a^{b^t}\left(\prod_{j=1}^{m} x_{ji}^{\beta_j}\right) e^{\sum_{k=1}^{n}\lambda_k I_{ik}}$ (12)

3.2.2 Dynamic model estimation and results

As in our previous analysis, 296 prefectures are included here. Since we have $t = 27$ days of observations until February 19, the total sample size for the final analysis is $n = 7{,}992$. We use the same machine learning method as before to estimate the parameters in models (7) through (12). We first fit these models using only the single variable of population outflow from Wuhan to other prefectures, and observe $R^2 = 0.930$. Figure 3 in the main text illustrates the basic features of our models just using the single variable of aggregate population outflow from Wuhan (from January 1 to 24).

In order to enhance the model's predictive strength, we added prefecture GDP and local population to models (7) through (12) (Supplementary Table 5). All six models have a good fit, $R^2 = 0.957\text{-}0.958$. For the both exponential and power types of models, whatever baseline hazard functional forms $\lambda_0(t)$ used, the key paramater β_i values are unchanged. The value of paramater β_1 is much larger than β_1 and β_3, implying that the outflow population from Wuhan plays a dominating role in predicting infections over time and space.

	Exponential-Logistic model (EL)	Exponential-Generalized-Logistic model (EGL)	Exponential-Gompertz model (EG)	Power-Logistic model (PL)	Power-Generalized-Logistic model (PGL)	Power-Gompertz model (PG)
R^2	0.957	0.958	0.958	0.957	0.958	0.958
g		0.170			0.170	
$\gamma/a/r$	0.274	0.181	0.004	0.274	0.181	0.004
$\omega/b/t_i$	3.386	10.910	0.850	3.386	10.910	0.850
α	0.930	0.900	0.559	5.042	5.603	4.231
β_1	1.360	1.360	1.360	0.617	0.617	0.617
β_2	0.106	0.106	0.106	0.088	0.088	0.088
β_3	0.273	0.273	0.273	0.355	0.355	0.355
Fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
N	7,992	7,992	7,992	7,992	7,992	7,992

Supplementary Table 5 | Results for dynamic models (7) through (12).

Since prefectures in Hubei are closely linked with Wuhan, they suffered more infections compared with prefectures in other provinces, and are thus themselves perceived to be a bigger source of risk. Indeed, 75.66% of Wuhan population outflow went to other prefectures within Hubei province, which had 56.43% of the total confirmed cases. During January 1- 24, there was an average daily population outflow of 478,270 from Wuhan, and an average daily population outflow of 916,090 from Hubei (excl. Wuhan) to prefectures in other provinces/regions.

To investigate this issue, we include secondary risk resource variable x_{4i}, outflow from Hubei (excluding Wuhan), into our dynamic models. Results are provided in the following Table 6. The estimated value for B_4 is essentially zero, i.e., the outflow population from Hubei (excluding Wuhan) had no significant effect on confirmed cases overall. More specifically, the addition of outflow from Hubei has no effect on R^2, compared with our previous analysis without this variable (see Supplementary Table 5).

	Exponential-Logistic model (EL)	Exponential-Generalized-Logistic model (EGL)	Exponential-Gompertz model (EG)	Power-Logistic model (PL)	Power-Generalized-Logistic model (PGL)	Power-Gompertz model (PG)
R^2	0.957	0.958	0.958	0.957	0.958	0.958
g		0.170			0.205	
$\gamma/a/r$	0.274	0.181	0.004	0.274	0.185	0.004
$\omega/b/t_i$	3.386	10.910	0.850	3.389	10.986	0.850
α	1.103	1.259	1.165	6.065	4.245	3.369
β_1	1.360	1.360	1.360	0.619	0.618	0.623
β_2	0.106	0.106	0.106	0.091	0.088	0.093
β_3	0.273	0.273	0.273	0.357	0.355	0.355
β_4	-0.001	-0.001	-0.001	0.000	0.000	0.000
Fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
N	7,992	7,992	7,992	7,992	7,992	7,992

Supplementary Table 6 | Effect of outflow from Hubei in dynamic models (7) - (12).

From the analysis above, we find that estimated results are identical between different models. There is no advantage to use the generalized logistic hazard function with an additional parameter g, compared with the original logistic function. In a prediction analysis, we also find that the generalized logistic models perform the worst in early days of the virus spread compared with logistic models (EL and PL) as well as Gompertz models (EG and PG).

3.2.3 Time period and time lag sensitivity analysis

We also conduct a sensitivity analysis by using different time periods for aggregate population outflow from Wuhan based on logistic and Gompertz models. We compare using population outflow from Jan 1-12 to outflow from Jan 13-24, and find that the latter (the more recent dates of travel) are stronger predictors in term of R^2 (0.962 – 0.963) and the outflow parameter value of β_1 (0.688 vs. 0.557 for power models; 1.530 vs. 1.224 for exponential models, Supplementary Table 7 and 8). This implies that *more recent outflow from Wuhan played a more important role in determining the spread of virus*, which is consistent with our correlational analysis (Extended Data Figure 1) and with the outflow-only model robustness check (Extended Data Figure 8).

	Exponential-Logistic model (EL)	Exponential-Gompertz model (EG)	Power-Logistic model (PL)	Power-Gompertz model (PG)
R^2	0.952	0.953	0.952	0.953
γ/a	0.274	0.004	0.274	0.004
ω/b	3.386	0.850	3.386	0.850
α	1.517	0.995	5.245	4.628
β_1	1.224	1.224	0.557	0.557
β_2	0.074	0.074	0.061	0.061
β_3	0.348	0.348	0.452	0.452
Fixed effects	Yes	Yes	Yes	Yes
N	0.952	0.953	0.952	0.953

Supplementary Table 7 | Results for dynamic models using outflow data from Jan 1 to 12.

	Exponential-Logistic model (EL)	Exponential-Gompertz model (EG)	Power-Logistic model (PL)	Power-Gompertz model (PG)
R^2	0.962	0.963	0.962	0.963
γ/a	0.274	0.004	0.274	0.004
ω/b	3.387	0.850	3.387	0.850
α	0.637	0.751	3.101	1.747
β_1	1.530	1.530	0.688	0.688
β_2	0.149	0.149	0.124	0.124
β_3	0.186	0.186	0.242	0.242
Fixed effects	Yes	Yes	Yes	Yes
N	0.962	0.963	0.962	0.963

Supplementary Table 8 | Results for dynamic models using outflow data from Jan 13 to 24.

3.3 Modelling daily confirmed cases

Our spatio-temporal risk source models can be used for modeling daily reported new cases. In order to do so, we need to have the first order derivative forms of our original models. We choose Logistic and Gompertz models for the baseline hazard function $\lambda_0(t)$ for this purpose, and the relevant models are as follows:

a) ΔEL model:

$$\frac{d\lambda(t|x_i)}{dt} = \frac{\alpha e^{-\gamma t+\omega}\gamma}{[1+e^{-\gamma t+\omega}]^2}\left(\prod_{j=1}^{m} e^{\beta_j x_{ji}}\right)e^{\sum_{k=1}^{n}\lambda_k I_{ik}} \quad (13)$$

b) ΔEG model:

$$\frac{d\lambda(t|x_i)}{dt} = \alpha a^{b^t} b^t \ln(a)\ln(b)\left(\prod_{j=1}^{m} e^{\beta_j x_{ji}}\right)e^{\sum_{k=1}^{n}\lambda_k I_{ik}} \quad (14)$$

c) ΔPL model:

$$\frac{d\lambda(t|x_i)}{dt} = \frac{\alpha e^{-\gamma t+\omega}\gamma}{[1+e^{-\gamma t+\omega}]^2}\left(\prod_{j=1}^{m} x_{ji}^{\beta_j}\right)e^{\sum_{k=1}^{n}\lambda_k I_{ik}} \quad (15)$$

d) ΔPG model:

$$\frac{d\lambda(t|x_i)}{dt} = \alpha a^{b^t} b^t \ln(a)\ln(b)\left(\prod_{j=1}^{m} x_{ji}^{\beta_j}\right)e^{\sum_{k=1}^{n}\lambda_k I_{ik}} \quad (16)$$

We use these derivative forms to model daily reported new cases up to February 19, and provide the results in the following table. As can be seen, the parameters remain stable though the model fitting is not as good as the original models for cumulative confirmed cases.

We also include the secondary risk resource variable x_{4i}, outflow from Hubei (excluding Wuhan), into the derivative models. The estimated value for B_4 is nearly zero as well (see Supplementary Table 10). Thus, we confirm again that outflow population from Hubei (excluding Wuhan) had no additional significant effect on daily confirmed cases over time across locales.

	ΔExponential-Logistic model (ΔEL)	ΔExponential-Gompertz model (ΔEG)	ΔPower-Logistic model (ΔPL)	ΔPower-Gompertz model (ΔPG)
R^2	0.713	0.716	0.713	0.716
γ/a	0.241	0.002	0.241	0.002
ω/b	2.987	0.847	2.986	0.847
α	1.047	1.359	0.862	1.760
β_1	1.453	1.444	0.657	0.652
β_2	0.150	0.150	0.123	0.122
β_3	0.264	0.265	0.343	0.345
Fixed effects	Yes	Yes	Yes	Yes
N	7,992	7,992	7,992	7,992

Supplementary Table 9 | Modeling daily confirmed cases in dynamic models (13) - (16).

	ΔExponential-Logistic model (ΔEL)	ΔExponential-Gompertz model (ΔEG)	ΔPower-Logistic model (ΔPL)	ΔPower-Gompertz model (ΔPG)
R^2	0.713	0.716	0.713	0.716
α	0.650	1.288	2.145	2.694
γ/a	0.241	0.002	0.241	0.002
ω/b	2.987	0.847	2.986	0.847
β_1	1.452	1.443	0.656	0.653
β_2	0.150	0.149	0.121	0.122
β_3	0.264	0.265	0.341	0.343
β_4	-0.001	0.004	0.010	0.008
Fixed effects	Yes	Yes	Yes	Yes
N	7,992	7,992	0.713	0.716

Supplementary Table 10 | Modeling daily confirmed cases in dynamic models (13) - (16).

3.4 From reference points to reference patterns

The risk model in Supplementary Information section 3.1.4 used the aggregate Wuhan population outflow to generate a single reference point (on a single date) of expected infections for each prefecture. Here, we used Wuhan population outflow to provide a reference growth *pattern* (i.e., across time) for COVID-19's spread (Fig. 3 in the main text) across 296 prefectures in China (the population of these prefectures comprise 94% of China's total population; some prefectures, mainly in sparsely populated autonomous regions, were not included in the model analysis due to lack of recent GDP data).

In contrast to SIR models, which estimates epidemiological spread in a mechanistic model, our model does not make *a priori* assumptions of any growth pattern or mechanism (beyond there being some relationship between local infection count and Wuhan's population outflow to that prefecture). Rather, the model leverages machine learning to statistically derive COVID-19's epidemic curve and growth pattern across China from all data points regarding confirmed cases across time and space. With more data, the hazard model fit becomes better, and a clearer picture of the epidemic curve emerges.

Similar to the static model, we can again use the predicted growth pattern of the virus as a benchmark to identify which locales deviate significantly based on our hazard models. Analogous to the logic of our earlier analysis using deviation between predicted and actual values of infection numbers to infer risk, here the differences in the growth trends between predicted and confirmed cases can signal higher levels of COVID-19 transmission. Once again, prefectures with lower trends than expected by our model might have had more successful public health measures in controlling the spread of the virus, while prefectures with higher-than-expected trends likely have more community transmission.

Most of the prefectures in Hubei province (excluding Wuhan) fit our dynamic models of virus spread very well, as illustrated in Extended Data Figure 5. Since the virus likely spread before the quarantine was imposed in Hubei and Wuhan, and before other strict public health measures were introduced, these epidemic curves likely represent something close to the virus's 'natural' growth pattern. Indeed, they closely resemble a classic S-curve. The machine-learning-based hazard model fitting process naturally chooses some of these prefectures as the reference growth trends as the basis for the comparison with other prefectures. Prefectures that have actual growth trends that are much higher than the predicted spread patterns are 'underperforming,' for example Suizhou and Xiaogan (the first two graphs in Extended Data Figure 5). Prefectures that have exhibited better-than-expected performance in controlling the spread of the virus include Jingzhou and Xianning (the last two graphs in Extended Data Figure 5).

3.5 A transmission risk index

It should be noted that our model does not merely predict expected levels of infection, but also creates benchmark 'growth patterns' for epidemiological growth. Hence, performance, as evaluated by our model, has less to do with absolute number of infections, and more to do with growth pattern of infections compared to what the model predicts (which is derived from observing the growth pattern in different locales across China).

In order to assess the risk of COVID-19 spread for different prefectures, we develop a measure of total transmission risk by exploiting the integral of the differences between actual confirmed infection cases and predicted numbers in our model.

$$\Delta_i = \sum_{t=1}^{T}\left[\lambda(t|i) - \hat{\lambda}(t|x_i)\right] \qquad (17)$$

where $\lambda(t|i)$ is the cumulative number of confirmed cases at time *t* for prefecture i, $\hat{\lambda}(t|x_i)$ is the estimated number of cases by our hazards models at time *t* for prefecture *i,* and *T* is the total time period (days) considered. We normalize the measure Δ_i by subtracting the mean and dividing by the standard deviation to form the final transmission risk index $\bar{\Delta}_i$. Extended Data Figure 7 is the distribution of the index based on our modeling results up to February 19. Prefectures above the 90% confidence interval of the index are likely experiencing more local transmissions than imported cases, and prefectures below the 90% confidence interval have a better performance in the virus spread control.

Supplementary Table 11 shows the top and bottom 10 prefectures on the transmission risk index up to February 19. When the value of risk index $\bar{\Delta}_i$ is larger than 1.645, 1.960, or 2.576, the corresponding prefectures have a statistically significant transmission risk at the 90%, 95%, or 99% confidence interval, respectively. Suizhou, Wenzhou, Xiaogan, and Shenzhen have a highly significant transmission risk index score at above the 99% confidence interval. We discuss these cases in the following section.

High transmission risk prefectures		Low transmission risk prefectures	
Prefecture	**Risk index $\bar{\Delta}_i$**	**Prefecture**	**Risk index $\bar{\Delta}_i$**
Suizhou	9.797	Xianning	-8.861
Wenzhou	6.749	Jingzhou	-3.727
Xiaogan	3.052	Hangzhou	-2.617
Shenzhen	2.843	Ezhou	-2.545
Xinyu	1.786	Guangzhou	-1.269
Bengbu	1.718	Huanggang	-1.016
Yichang	1.122	Huzhou	-0.847
Taizhou	1.119	Jiujiang	-0.791
Bozhou	1.005	Ji'an	-0.694
Shaoyang	0.950	Chuzhou	-0.596

Note: The threshold for the 90% and 95% confidence intervals are 1.644 and 1.960, respectively.

Supplementary Table 11 | Transmission risk index $\bar{\Delta}_i$ for top 10 and bottom 10 prefectures on February 19.

3.6 Comparison between incidence rate and risk index

In epidemiology, incidence rate and incidence proportion are commonly used measures for epidemic risk, which are typically based on the number of persons in a vulnerable population that are infected at a given time: Local population is the most commonly used denominator for incidence rate. Since our analyses and risk index uses population outflow from Wuhan (the epicenter of the outbreak) as the benchmark for risk assessment, we may re-conceptualize our risk index as a population-outflow based incidence rate measurement.

In other words, we re-conceive incidence rate as the ratio between infection count and risk importation count, i.e., the ratio of confirmed cases and population outflow. Supplementary Table 12 provides a comparison of the three different measures of risk on February 12.

City name	Region	Confirmed cases	Population based incidence rate	Outflow based incidence rate	Risk index

Suizhou	Hubei	1160	0.046%	0.372%	9.641
Xiaogan	Hubei	2874	0.055%	0.130%	2.600
Yichang	Hubei	810	0.021%	0.296%	1.270
Xiangyang	Hubei	1101	0.019%	0.293%	0.816
Jingmen	Hubei	927	0.032%	0.231%	0.659
Shiyan	Hubei	562	0.016%	0.322%	0.535
Huangshi	Hubei	911	0.034%	0.193%	0.471
Huanggang	Hubei	2662	0.036%	0.139%	-0.442
Ezhou	Hubei	1065	0.096%	0.098%	-2.602
Jingzhou	Hubei	1431	0.022%	0.212%	-5.230
Xianning	Hubei	534	0.018%	0.067%	-8.334
Wenzhou	Zhejiang	490	0.006%	5.222%	6.784
Taizhou	Zhejiang	144	0.002%	2.631%	1.204
Ningbo	Zhejiang	153	0.003%	1.926%	0.094
Zhoushan	Zhejiang	10	0.001%	0.914%	-0.205
Lishui	Zhejiang	17	0.001%	1.014%	-0.323
Jiaxing	Zhejiang	42	0.001%	1.257%	-0.438
Quzhou	Zhejiang	21	0.001%	0.686%	-0.595
Shaoxing	Zhejiang	41	0.001%	1.184%	-0.627
Jinhua	Zhejiang	55	0.001%	1.052%	-0.633
Huzhou	Zhejiang	10	0.000%	0.319%	-0.870
Hangzhou	Zhejiang	162	0.002%	0.688%	-2.597
Shuangyashan	Heilongjiang	39	0.004%	975.000%	0.493
Jixi	Heilongjiang	44	0.003%	275.000%	0.440
Suihua	Heilongjiang	45	0.001%	10.843%	0.335
Qiqihar	Heilongjiang	33	0.001%	20.625%	0.183
Qitaihe	Heilongjiang	16	0.002%	200.000%	0.163
Jiamusi	Heilongjiang	15	0.001%	25.000%	0.155
Mudanjiang	Heilongjiang	12	<0.001%	25.000%	0.084
Hegang	Heilongjiang	5	<0.001%	125.000%	-0.026
Heihe	Heilongjiang	10	0.001%	27.778%	-0.071
Yichun	Heilongjiang	0	<0.001%	0.000%	-0.129
Daqing	Heilongjiang	15	0.001%	2.404%	-0.182
Harbin	Heilongjiang	159	0.002%	3.586%	-0.290
Shenzhen	Guangdong	391	0.009%	0.894%	2.213
Chongqing	Chongqing	84	0.001%	0.339%	0.017
Beijing	Beijing	28	0.000%	0.253%	-0.019
Shanghai	Shanghai	42	0.001%	0.486%	-0.061
Chengdu	Sichuan	110	0.001%	0.359%	-0.062
Suzhou	Jiangsu	505	0.001%	0.523%	-0.076
Fuzhou	Fujian	35	0.001%	0.312%	-0.076

Lanzhou	Gansu	352	0.003%	0.355%	-0.077
Shenyang	Liaoning	311	0.002%	0.392%	-0.077
Tianjin	Tianjin	107	0.001%	0.511%	-0.077
Kunming	Yunnan	31	0.002%	0.792%	-0.105
Yinchuan	Ningxia	46	0.001%	0.289%	-0.117
Xi'an	Shanxi	150	0.002%	3.383%	-0.149
Jinan	Shandong	47	0.001%	0.389%	-0.153
Changchun	Jilin	64	0.001%	0.389%	-0.163
Zhengzhou	Henan	125	0.001%	0.410%	-0.220
Changsha	Hunan	27	0.000%	0.160%	-0.370
Hefei	Anhui	53	0.001%	0.314%	-0.390
Shijiazhuang	Hebei	87	0.001%	0.229%	-0.409
Qingdao	Shandong	157	0.002%	0.289%	-0.451
Harbin	Heilongjiang	137	0.002%	0.132%	-0.541
Nanjing	Jiangsu	223	0.003%	0.142%	-0.557
Guangzhou	Guangdong	323	0.004%	0.541%	-1.477

Supplementary Table 12 | Comparison between incidence rate and risk index.

We first discuss the prefectures in Supplementary Table 12 that are in Hubei province. Our transmission risk index identifies Suizhou and Xiaogan (the first two graphs in Extended Data Figure 6) in Hubei as the most statistically significant “underperformers,” with both at the 99% confidence interval. Consistent with this prediction, which we made on February 12, 2020 based on data through that date, the Hubei provincial government implemented an even more restrictive quarantine for Wuhan and Xiaogan on February 16, whereby people were strictly prohibited from leaving their homes. Although Xiaogan had the highest number of confirmed cases in Hubei province (except Wuhan), its population-based incidence rate was not the highest and the outflow-based incidence rate was also relatively small.

Suizhou (170km from Wuhan, population 2.21M), despite being the second smallest prefecture in Hubei, had the highest outflow based incidence rate in the province and the highest value on the transmission risk index: Even our analyses at earlier dates suggested it should deserve greater government scrutiny. Although the city was not subject to stricter quarantine policies like Xiaogan, media reports have suggested the city was seriously stricken and have reported that dozens of local government officials were called to account for inadequate control of the outbreak.[54-55]

On the other hand, Xianning and Jingzhou (the last two graphs in Extended Data Figure 5) as well as Ezhou exhibited better-than-expected performance in controlling the spread of the virus. The three prefectures had the lowest outflow-based incidence rates, though Ezhou had the highest population-based incidence rate.

In Zhejiang province, a cluster of four highly interconnected prefecture-cities (Wenzhou, Taizhou, Ningbo, and Hangzhou; all four were in our top-10 list) were the first prefectures to be quarantined outside of Hubei. Beginning February 2, the Zhejiang provincial government adopted measures including allowing only one person per household to leave their home every two days to buy basic necessities. And, indeed, Wenzhou and Taizhou were above the 95% confidence interval in our transmission risk index; Ningbo and Hangzhou were also among the top-10 list in the risk index though they did not reach the 90% significance level (Supplementary Table 12, Extended Data Figure 6). The decision to quarantine these two prefectures may have also been influenced by their relative proximity and interconnectivity with Wenzhou in particular. Wenzhou has a reputation of being a city of geographically well-connected entrepreneurs, and highlights the downside risk of greater socio-economic connectivity; 1.6 million people originally from Wenzhou are estimated to run businesses in other parts of China (and another 600,000 overseas).

Early in the outbreak, Shenzhen and Guangzhou had a relatively high absolute number of confirmed cases; local media in both prefectures also reported on numerous primary or community transmission cases. Guangzhou exhibited an improving (i.e., declining) trend in the risk index over time, and had the lowest risk index outside of Hubei, Zhejiang, and Heilongjiang (i.e., it had fewer infections over time than expected given population inflow from Wuhan). However, Shenzhen had a persistently high risk index score, albeit with a slightly decreasing trend. As an economic boomtown and migrant hub, it is possible that Shenzhen may have faced additional virus import risk from many other regions in China. Finally, as shown in the last three graphs in Extended Data Figure 6, Beijing, Shanghai, and Nanjing have improving trends on the transmission risk index, although they are not yet significantly over-performing.

Overall and in general, our transmission risk index provides a simple and useful method to identify situations of transmission risk in different locales.

3.7 Entropy of population versus outflow based incidence rates

We use the following entropy measure to capture the uniformity of incidence rates of COVID-19 based on the population outflow from Wuhan:

$$E = -\sum_{i=1}^{n} p_i \ln(p_i) \qquad (18)$$

where p_i is the population-outflow normalized incidence rate of prefecture i, i.e., confirmed infections in a prefecture divided by the aggregate population outflow from Wuhan to prefecture i between January 1 and 24, with rates normalized to sum to 1. The larger the value of entropy, the more uniformity of the infection rates among prefectures. We also create an entropy measure for the uniformity of incidence rates based on population of that prefecture (which is the classic definition of incidence rate).

Entropy increased in the first week of the study period and remained flat thereafter, which suggests that the rates of confirmed infection cases based on population outflow from Wuhan remained uniform across Chinese prefectures. One possible reason for our model's robustness beyond January 24 (in the still-early stages of this epidemic) may relate to the fact that, as recent research has shown, most early transmissions occurred in family clusters[21], which would explain why infection growth remains proportional to population outflow from Wuhan (with average household size as a possible scaling factor).

4. Comparative study regarding gravity models

4.1 Static model comparison

Gravity models, inspired by Newton's gravity law, were originally developed to model mobility flows between two populations considering their relative distance.[56] They have been broadly used to study spatial interactions between different places, including traffic behavior, economic exchange, migration, and disease spread (see Barbosa et al. (2018)[33] for a recent review). A typical gravity model for mobility flow T_{ij} between two populations $P_i$and P_j in areas *i* and *j* has a multiplicative form: $T_{ij} = cP_i^{\beta_1}P_j^{\beta_2}d_{ij}^{-\beta_3}$, where d_{ij} is the distance between areas *i* and *j*, and *c* and βi are parameters to estimate. The distance function is typically modeled by a power law $d_{ij}^{-\beta_3}$ or an exponential form $e^{-\beta_3 d_{ij}}$.[33]

In the area of epidemic research, gravity models have been widely used to study the spatial spread of virus and its relationship with mobility.[38, 57-61] When using the gravity model to estimate the epicenter Wuhan's effect on population outflow as well as infections in other prefectures T_{ij}, we let *j* = 1 for Wuhan; the Wuhan population variable can then be dropped in the model (since it is constant across all prefectures in the model). In order to be comparable with our risk source models (1) and (2), we also control for prefecture GDP and provincial fixed effects. Then we can have the following gravity models for our purpose of comparison.

$$T_{i1} = ce^{\beta_1 P_i}e^{\beta_2 G_i}e^{-\beta_3 d_{i1}}e^{\sum_{k=1}^{n}\lambda_k I_{ik}} \qquad (19)$$

$$T_{i1} = cP_i^{\beta_1}G_i^{\beta_2}d_{i1}^{-\beta_3}e^{\sum_{k=1}^{n}\lambda_k I_{ik}} \qquad (20)$$

where notations are the same as for models (1) and (2), and all the variables are also normalized in the same way as before. We use the gravity models to predict both outflow from Wuhan to prefecture *i* and infections in prefecture *i* with and without fixed effects. Provincial fixed effects allow the unobserved location specific effects to be correlated with the explanatory variables. Since these effects may have a similar role as distance, we also conduct analyses without fixed effects in order to have a typical gravity model for comparison. We also replace the distance variable with the outflow variable $\mathrm{O}_\mathrm{i}^{-\beta_0}$to jointly model it with distance in our analysis.

Results are provided in Supplementary Table 13 for the exponential model (19) and Supplementary Table 14 for the power model (20) at the date of February 19 for cumulative confirmed cases. Since the two types of models have almost the same goodness of fit, we just discuss results for model (19). As we can see, when provincial fixed effects are in place, the parameter value and effect of distance are reduced in the model. In general, the gravity models can predict outflow from Wuhan to different prefectures very well, R^2 = 0.902, but are less predictive of number of confirmed cases, R^2 = 0.758, when fixed effects are not included. If we use outflow instead of distance in the model, then $R^2 = 0.941$, which is much higher than the original gravity model without fixed effects.

When we include both outflow and distance into the models, the parameter β_0 for outflow has a much larger value than others in the model, the parameter β_3 for distance becomes nearly zero, and the rest of parameters remain stable. Thus, adding the distance variable into the model does not contribute to model fit (i.e., no change for R^2) in the presence of the population outflow variable. Although "recipient" population size and distance were significant predictors for infections ($p < .001$) in the absence of outflow, a mediation analysis shows that population outflow from Wuhan fully mediates the effect of distance on infections (i.e., the parameter of distance is no longer statistically significant when considered jointly with outflow). In fact, there is no advantage to estimating population flow and estimating infection spread by using estimated population flow, when population flow is actually observable, as in our case. Thus, it becomes possible to jointly evaluate the effects of population outflow and disease spread in the same model as our risk source models; whereas gravity models typically evaluate each relationship separately.

Dependent Variable	Outflow from Wuhan	Outflow from Wuhan	Confirmed Cases	Confirmed Cases	Confirmed Cases	Confirmed Cases	Confirmed Cases
Constant (*c*)	9623.48	15.919	21.934	2.278	13.669	4.222	3.935
Outflow (β_0)					1.821	1.507	1.565
GDP (β_1)	-0.408	0.154	0.124	0.286	0.136	0.152	0.144
Population (β_2)	0.505	0.656	0.751	0.683	0.135	0.221	0.202
Distance (β_3)	-1.250	-0.844	-1.196	-0.579			0.025
Fixed effects	No	Yes	No	Yes	No	Yes	Yes
R^2	0.902	0.966	0.758	0.941	0.941	0.965	0.965
N	296	296	296	296	296	296	296

Supplementary Table 13 | Comparison regarding exponential gravity model (19). Empty cells denote when the variable listed in the dependent variable column is not present in the model. Presence of fixed effects in the model is denoted by 'Yes' and 'No.'

Dependent Variable	Outflow from Wuhan	Outflow from Wuhan	Confirmed Cases	Confirmed Cases	Confirmed Cases	Confirmed Cases	Confirmed Cases
Constant (*c*)	6016.43	38.371	22.168	2.278	133.400	8.661	3.935

Outflow (*β_0*)					0.827	0.684	1.565
GDP (*β_1*)	-0.343	0.124	0.102	0.286	0.113	0.126	0.144
Population(*β_2*)	0.659	0.853	0.981	0.683	0.176	0.287	0.202
Distance (*β_3*)	-1.861	-1.253	-1.783	-0.579			0.025
Fixed effects	No	Yes	No	Yes	No	Yes	Yes
R^2	0.901	0.966	0.757	0.941	0.941	0.965	0.965
N	296	296	296	296	296	296	296

Supplementary Table 14 | Comparison regarding power gravity model (20).

Previous research also shows that gravity models are only suitable within a certain distance[57-58]. Viboud et al. (2006)[57] found that workflows capture the spread of influenza better than simple Euclidean distance or other movement metrics. Balcan et al. (2009)[58] used a gravity model to investigate the different contributions of the long- and short-range mobility flows on infectious diseases. Brockmann and Helbing (2013)[62] proposed a concept of effective distance, reflecting that a small fraction of traffic flow is effectively equivalent to a large distance, and vice versa. Our outflow population from Wuhan can be regarded as an “effective distance” measure between Wuhan and other prefectures that reflects effective interactions in social and economic perspectives. Our risk source models can be viewed as a further development by replacing the distance variable in gravity models with population outflow from risk source. With these empirical relationships in mind, we develop a “risk source” model that focuses on leveraging observed population flow data to operationalize the risk emanating from the epidemic source or epicenter.

4.2 Spatio-temporal model comparison

Gravity models may be used to study the geographic properties of the spread of something like a virus. Even though some prior work has also studied temporal features of transmission using various approaches[57-58], this work has not offered an integrated spatio-temporal model. In order to study the combined properties of the spread of a virus over time and space jointly in a unified model, researchers have tried to extend gravity models by incorporating some time-varying components into their models.[63-66] In particular, several recent papers have used the framework of Cox proportional hazard model to develop spatio-temporal epidemiological models. [64-67]

We consider replacing the constant *c* with a baseline hazard function $\lambda_0(t)$ based on the Cox hazard model framework, which leads to the following spatio-temporal gravity models:

$$\lambda(t|P_i, G_i, d_{i1}) = \lambda_0(t) e^{\beta_1 P_i} e^{\beta_2 G_i} e^{-\beta_3 d_{i1}} e^{\sum_{k=1}^{n} \lambda_k I_{ik}} \qquad (21)$$

$$\lambda(t|P_i, G_i, d_{i1}) = \lambda_0(t) P_i^{\beta_1} G_i^{\beta_2} d_{i1}^{-\beta_3} e^{\sum_{k=1}^{n} \lambda_k I_{ik}} \qquad (22)$$

These models are in fact special cases of our risk source models (5) and (6) where the outflow variable is replaced by distance. Alternative approaches beyond gravity models can also be used to create sophisticated dynamic spatio-temporal models (e.g., Tang et al. (2019) for a Bayesian framework considering time and geographic spread[68]); our spatio-temporal risk source models offer a dynamic framework that is parsimonious, descriptive, and insightful.

As with our earlier models (5) and (6), we use three most popular sigmoidal functions: logistic, generalized logistic, and Gompertz functions for $\lambda_0(t)$ in this comparison analysis. These models have been often used in the epidemic literature to model infectious disease spread.[51-53]

Results based on models (21) and (22) are provided in Supplementary Table 15. When we include both outflow and distance into our spatio-temporal risk source models, the parameter for distance is much smaller than other parameters in the model, i.e., the outflow variable play the most important role in driving COVID-19 dissemination over time and space in the country. Compared with our early results in Supplementary Table 5, adding the distance variable into our models does not improve model fit (i.e., no change for R^2) in the presence of the outflow variable.

	Exponential-Logistic model (EL)	Exponential-Generalized-Logistic model (EGL)	Exponential-Gompertz model (EG)	Power-Logistic model (PL)	Power-Generalized-Logistic model (PGL)	Power-Gompertz model (PG)
g		0.171			0.171	
$\gamma/a/r$	0.274	0.181	0.004	0.274	0.181	0.004
$\omega/b/t_i$	3.386	10.913	0.85	3.386	10.913	0.85
α	1.097	1.196	1.218	4.97	4.896	4.231
Outflow (β_0)	1.422	1.421	1.421	0.645	0.645	0.644
GDP (β_1)	0.095	0.095	0.095	0.08	0.08	0.08
Population (β_2)	0.253	0.253	0.253	0.328	0.329	0.329
Distance (β_3)	0.029	0.028	0.028	0.042	0.041	0.041
Fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.957	0.958	0.958	0.957	0.958	0.958
N	7,992	7,992	7,992	7,992	7,992	7,992

Supplementary Table 15 | Comparison between outflow and distance in dynamic models.

5. Assessment of alternative population outflow measures

In our main analysis, we use the total population outflow from Wuhan as the primary risk resource and outflow from Hubei as the secondary risk resource. Here we evaluate alternative measures of population outflow (Section 1.2), using different approaches, as robustness checks.

First, we distinguish returning residents versus 'migrants' in our analyses. The carrier also provided us a count of returning residents who had spent at least 2 hours in Wuhan from January 1-24. Residence is defined by customer's registered home address, which is also used for billing purposes (local calls receive a preferential rate). Thus, we can separate total population outflow from Wuhan into two types of travelers: 1) returning residents of each prefecture (who visited Wuhan and then traveled back to their home prefectures); 2) non-returning-residents ('migrants') who travelled from Wuhan before entering the prefecture. In vast majority of the latter cases were likely Wuhan or Hubei residents. In our dataset, 38.78% of population outflow from Wuhan were returning residents of other prefectures in Hubei; only 9.10% were returning residents from prefectures outside of Hubei.

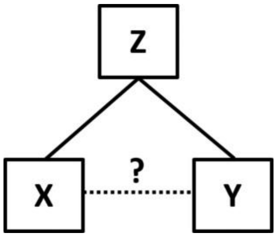
Second, we investigate the effect of population outflow from Hubei (excluding Wuhan) on infection count in prefectures across China. Besides being the most stricken province in China, Hubei residents were subject to particularly stringent travel restrictions in China and internationally.

Third, we evaluate how population outflow that occurred after the quarantine was imposed affected COVID-19 spread. We note in the main text that a very small stream of population outflow continued after the quarantine of Wuhan and prefectures in Hubei (January 24). We note that some movement of government, medical, rescue, and logistical service providers across prefecture borders is to be expected. By testing the impact of post-quarantine population flow, we evaluate the effectiveness of the quarantine of Wuhan city and Hubei province.

Since these alternative measures of population outflow are highly inter-correlated (e.g., $r = 0.876$ for returning resident outflow and non-returning-resident outflow), simply including them into existing models as additional variables would introduce collinearity problems. Here we use nonparametric statistics and machine learning approaches to address this.

5.1 Kernel-based conditional independence test

Let X, Y and Z denote sets of random variables. The Conditional Independence (CI) between X and Y given Z, is denoted by $(X \perp Y)|Z$, meaning: if and only if, given the values of Z, further knowing the values of X (or Y) does not provide any additional information on the likelihood of Y (or X) occurring. In the dependence graph, this corresponds to whether the link between X and Y exists conditional on that the other two links exist.



Zhang et al. (2012)[69] developed a Kernel-based Conditional Independence test (KCI-test), by constructing an appropriate test statistic and deriving its asymptotic distribution under the null hypothesis of conditional independence. This approach can test CI for continuous variables without assuming a functional form between the variables as well as the data distributions.

In this test, X represents an outflow variable, Y is cumulative confirmed cases, and Z denotes a set of variables including local population size, GDP, and other outflow variables not included in X. Specifically, we let x_1 = Non-returning-resident outflow from Wuhan during January 1-24, x_2 = Returning resident outflow from Wuhan during January 1-24, x_3 = Population outflow from Hubei (excluding Wuhan) during January 1-24, x_4 = Additional population outflow from Wuhan during January 25 – February 6, and x_5 = Additional population outflow from Hubei (excluding Wuhan) during January 25 – February 6; y = cumulative confirmed cases by February 19; and z_1 = local population, and z_2 = GDP. We use the full sample of 296 prefectures (excluding Wuhan) in this test. The results are as follows:

X	*Y*	*Z*	*p-value*
x_1	y	$z_1, z_2, x_2, x_3, x_4, x_5$	0.033
x_2	y	$z_1, z_2, x_1, x_3, x_4, x_5$	0.476
x_3	y	$z_1, z_2, x_1, x_2, x_4, x_5$	0.043
x_4	y	$z_1, z_2, x_1, x_2, x_3, x_5$	0.226
x_5	y	$z_1, z_2, x_1, x_2, x_3, x_4$	0.810

Supplementary Table 16 | Results of kernel-based conditional independence test for confirmed cases.

In order to be consistent with our Poisson model (3), we use the logarithmic transformation of confirmed cases for an alternative test. The overall test results become more significant as follows:

X	Y	Z	*p-value*
x_1	log(y+1)	$z_1, z_2, x_2, x_3, x_4, x_5$	0.000
x_2	log(y+1)	$z_1, z_2, x_1, x_3, x_4, x_5$	0.020
x_3	log(y+1)	$z_1, z_2, x_1, x_2, x_4, x_5$	0.012
x_4	log(y+1)	$z_1, z_2, x_1, x_2, x_3, x_5$	0.022
x_5	log(y+1)	$z_1, z_2, x_1, x_2, x_3, x_4$	0.259

Supplementary Table 17 | Results of kernel-based conditional independence test for the logarithm transformation of confirmed cases.

Results show that non-returning-resident outflow from Wuhan (i.e., x_1, most of whom are likely Wuhan residents) contributed significantly more than returning resident outflow from Wuhan (x_2) and population outflow from Hubei (x_3) to local confirmed cases (though both x_2 and x_3 are significant at 95% confidence level for the logarithmic transformed cases). *In other words, population outflow from Wuhan in particular, rather than even neighboring locales, was the primary predictor of COVID-19 infections.* Indeed, Wuhan residents should have had a longer potential exposure to the virus and were thus relatively more likely to be infected than mere visitors to Wuhan.

Compared with the total population outflow from Wuhan during January 1-24 (x_1+x_2), the additional outflow from Wuhan from January 25 – February 6 (x_4) had either no significant effect or a limited effect on confirmed cases (though it was significant at 95% for the latter case of logarithmic transformation). The additional population outflow from Hubei (excluding Wuhan) during January 25 – February 6 (x_5) had no significant effect on confirmed cases when other outflow variables were considered. *These results suggest that the quarantine of Hubei province was effective in controlling and preventing further spread of COVID-19 (i.e., SARS-CoV-2 infections) to other parts of the country.*

5.2 Random forest tree analysis

We used random forest (RF) models to assess the relative contribution of the outflow variables to number of confirmed cases.[70-71] RF models use bagging (bootstrap aggregating) of decision trees in order to reduce variance of single trees, and thus improve prediction accuracy. A random forest consists of a large number of regression trees; the overall prediction of the forest is the average of predictions from the individual trees. Since individual trees produce multidimensional step functions, their average is a multidimensional step function that can nevertheless predict smooth functions.

Our RF model consists of 300 trees, and we use 296 prefecture samples to fit our model, the vector X (= $[x_1, x_2, x_3, x_4, x_5]$) is regarded as the inputs, and vector Y (=[log(*confirmed_cases y* + 1)] is regarded as the target. The test error of RF models is estimated on the out-of-bag (OOB) data, and the function to measure the quality of a tree split is *MSE* (mean squared error), which is equal to using variance reduction as the feature selection criterion. After multiple iterations of training, the optimal value of *MSE* is 0.0861 (R^2=0.9639) calculated by predicted Y and actual Y.

The variable importance metric in CART trees and random forests is called Gini importance for classification problems, but the MSE reduction is used as the regression random forest importance criterion. The MSE reduction (as factor importance indicator) according to regressor X_j for the complete forest is obtained as the average over all N (=300) trees of these differences. And for comparison purposes, all variable importance metrics have been normalized to sum to 1. Results are provided in the following table.

Variables	**Importance (sum to 1)**
Non-returning-resident outflow from Wuhan – Jan 1-24 (x_2)	0.564
Returning resident outflow from Wuhan – Jan 1-24 (x_1)	0.159
Local population (z_1)	0.116
Local GDP (z_2)	0.052
Population outflow from Hubei – Jan 1-24 (x_3)	0.049
Additional outflow from Hubei – Jan 25–Feb 6 (x_5)	0.031
Hubei dummy	0.016
Additional outflow from Wuhan – Jan 25–Feb 6 (x_4)	0.015

Supplementary Table 18 | Importance of different outflow variables by RF.

Similar to the CI test, non-returning-resident outflow from Wuhan (i.e., Wuhan residents) is the most important factor in predicting confirmed cases. The returning resident outflow from Wuhan is ranked second in importance, and is 3.55 times less important than non-returning-residents. Thus, again, non-returning-resident outflow from Wuhan (i.e., likely Wuhan residents) were primarily responsible to the spread of SARS-CoV-2 in prefectures in China. Other outflow variables are much less important (during January 1-24).

When we combine resident and non-returning-resident outflow from Wuhan together (x_1+x_3), the total population outflow from Wuhan takes 70.3% of the importance score (see Supplementary Table 19**)**. The population outflow from Wuhan during January 1-24 is 41 times more important than that of the additional population outflow from Wuhan after the quarantine in predicting confirmed cases. Thus, the quarantine of Wuhan seemed relatively successful in controlling spread of the virus.

Variables	Importance (sum to 1)
Total outflow from Wuhan – Jan 1-24 (x_1+x_2)	0.703
Local population (z_1)	0.119
Local GDP (z_2)	0.057
Population outflow from Hubei – Jan 1-24 (x_3)	0.056
Additional outflow from Hubei – Jan 25–Feb 6 (x_5)	0.031
Hubei dummy	0.018
Additional outflow from Wuhan – Jan 25–Feb 6 (x_4)	0.017

Supplementary Table 19 | Importance of combined outflow variables by RF.

6. Tencent data analyses

As noted in our discussion in the main text, our methodology may be applied to any form of relatively representative population outflow data whether it is toll-booth data, traffic data, mobile app data, cell tower triangulation or GPS data, etc.

Therefore, as a robustness check, in order to show that our data was representative of national travel patterns (as discussed above), we further replicate our analyses using Tencent mobility data (i.e., mobile app geolocation data from the most popular Chinese social media and communications ecosystem) from the *chunyun* migration of 2017. We show, by replicating our general pattern of results, that our data and methodology is not sensitive to our sample; the carrier's estimation procedure; or the kind of telecommunications data.

These data are available online and describe human mobility and population outflow from Wuhan to other prefectures in China from December 3, 2016 to January 24, 2017 (Lunar New Year's Eve was January 27, 2017) using Tencent user geolocation data (e.g., from mobile phone apps such as WeChat, QQ, Tencent Map).[72] Although this Tencent mobility dataset was from 2017 and did not capture contemporaneous population outflows like our own data from 2020, we believe the general pattern of travel should be similar across years. Although one might expect inbound travel to Wuhan to have decreased as a result of the virus, Wuhan residents who were visiting other prefectures during the holiday had additional reasons to leave town in 2020. Since travel during migration and emergencies is commonly determined by the presence of social connections[73-75], we expected travel patterns to be similar across years. In addition, the fact that the Lunar New Year also occurred in late January (January 22) in 2017 helps account for seasonality effects.

We choose the same 296 prefectures for this comparative analysis between our dataset and the Tencent dataset. As hypothesized, the two datasets have a high correlation in mobility flows, *r* = 0.943. For new confirmed COVID-19 cases, the correlations with the two datasets are similar, especially at a later stage of the spread of the virus, *r* = 0.951 for the Tencent dataset and *r* = 0.952 for our telecom data. However, after including local population and GDP in the model analysis, Tencent data had *poorer* performance than our telecommunications data, with R^2 = 0.932 and 0.933 for EL and EG models respectively; compared to R^2 = 0.957 and 0.958 for the respective models using the mobile telecommunications data. Detailed results for the Exponential-Logistic model (6) and Exponential-Gompertz model (7) are provided in Supplementary Table 20.

The high predictive utility of the Tencent data also implies that *chunyun* migration patterns out of Wuhan from 2017 was very similar (but not identical) to *chunyun* or Wuhan population outflow in 2020 (as we hypothesized). This result also has implications for policymakers, particularly in China, since it suggests that (1) prior years' non-emergency holiday travel patterns may be predictive of future travel patterns during emergencies (particularly if mobility during those holidays is motivated by the location of family ties), and (2) traditional cultural migration patterns are highly robust to disruption, for better or worse. Future researchers could examine if migrations in other countries, such as Christmas travel in Western countries, Thanksgiving travel in the United States, or recurring religious pilgrimages, are similarly predictive of post-disaster travel patterns and investigate their role in the spread of epidemic outbreaks.

	Tencent mobility data from 2017 *chunyun*		Telecom mobility data from main analyses	
	Exponential-Logistic (EL)	Exponential-Gompertz (EG)	Exponential-Logistic	Exponential-Gompertz
R^2	0.932	0.933	0.957	0.958
α	0.244	0.250	0.930	0.559
γ/a	0.277	0.004	0.274	0.004
ω/b	3.408	0.849	3.386	0.850
β_1	2.693	2.693	1.360	1.360
β_2	-0.085	-0.085	0.106	0.106
β_3	0.115	0.115	0.273	0.273
Fixed effects	Yes	Yes	Yes	Yes
N	7,992	7,992	7,992	7,992

Supplementary Table 20 | Comparison between Tencent data with our data in model fitting.

7. Code for model estimation and data availability

Our data for our main analyses is available online at this journal’s website, as supplementary information.

Our code is below.

```
import pandas as pd
from sklearn.metrics import r2_score
from scipy import stats
import numpy as np
import warnings
from lmfit import Model  # solve non-linear optimization by Levenberg-Marquardt algorithm

'''
We use the source code library of lmfit in python
“LMFIT: Non-linear least-square minimization and curve-fitting for Python” (Newville et al. 2016)
to estimate the parameters in our models. The relevant codes are provided as follows.
'''

data_path = "pneumonia_panel_296_cities(submit).csv"
INF = 99999999999999
NAN = 0

def standardization(data, variables):
    for var in variables:
        try:
            ln_var = "%s(Ln)" % var
            mean_var = "%s(Ln_Mean)" % var
            std_var = "%s(Ln_Std)" % var
            data[ln_var] = np.log(data[var] + 1)
            group = data.groupby("day", as_index=False)[ln_var].mean()
            group.columns = ["day", mean_var]
            data = pd.merge(data, group, on="day", how="left")
            group = pd.DataFrame(data.groupby("day")[ln_var].std())
            group.columns = [std_var]
            group["day"] = group.index
            group.reset_index(drop=True, inplace=True)
            data = pd.merge(data, group, on="day", how="left")
        except Exception as e:
            print(e)
    return data

def normalization(data, variables):
    for var in variables:
        sum_var = "%s(Sum)" % var
        group = data.groupby("day", as_index=False)[var].mean()
        group.columns = ["day", sum_var]
        data = pd.merge(data, group, on="day", how="left")
    return data

def province_dummies(data):
    return np.array(data[["Shanghai", "Yunnan", "Neimenggu", "Beijing", "Jilin", "Sichuan",
"Tianjin", "Ningxia", "Anhui","Shandong", "Shanxi", "Guangdong", "Guangxi", "Xinjiang",
"Jiangsu", "JIangxi", "Hebei", "Henan", "Zhejiang", "Hainan", "Hubei", "Hunan", "Gansu",
"Fujian", "Xizang", "Guizhou", "Liaoning", "Chongqing", "Shaanxi", "Qinghai",
"Heilongjiang"]].values)

# Daily Exponential Static Model
def exponential_model(X, alpha, beta_1, beta_2, beta_3, beta_4, beta_5, lambda_1,
lambda_2, lambda_3, lambda_4, lambda_5, lambda_6, lambda_7, lambda_8, lambda_9,
lambda_10, lambda_11, lambda_12, lambda_13, lambda_14, lambda_15, lambda_16,
lambda_17, lambda_18, lambda_19, lambda_20, lambda_21, lambda_22, lambda_23,
lambda_24, lambda_25, lambda_26, lambda_27, lambda_28, lambda_29, lambda_30,
lambda_31):
    # province effect fixed
    lambdas = [lambda_1, lambda_2, lambda_3, lambda_4, lambda_5, lambda_6, lambda_7,
lambda_8, lambda_9, lambda_10, lambda_11, lambda_12, lambda_13, lambda_14, lambda_15,
lambda_16, lambda_17, lambda_18, lambda_19, lambda_20, lambda_21, lambda_22,
lambda_23, lambda_24, lambda_25, lambda_26, lambda_27, lambda_28, lambda_29,
lambda_30, lambda_31]
    betas = [beta_1, beta_2, beta_3, beta_4, beta_5]
    N = len(X) - 31
    fix = 1
    for i in range(len(lambdas)):
        fix = fix * np.exp(lambdas[i] * X[i + N])
    R = alpha
    for i in range(N):
        R = R * np.exp(betas[i] * X[i])
    R = R * fix
    R[np.isinf(R)] = INF
    R[np.isnan(R)] = NAN
    return R

def exponential_static_model_estimate(date="2020-01-28", y="confirmed",
x=["wuhan_outflow", "gdp", "population"]):
    data = pd.read_csv(data_path)
    province = pd.read_csv("province_fix.csv")
    data = pd.merge(data, province, on="province", how="left")
    data.fillna(0, inplace=True)
    data = standardization(data, x)
    data = data[data["date"] == date]
    Y = data[y]
    X = []
    for k in x:
        X.append((data["%s(Ln)" % k] - data["%s(Ln_Mean)" % k]) / data["%s(Ln_Std)" % k])
    fix = province_dummies(data)
    X.append(fix.T)
    X = np.vstack(X)
    model = Model(exponential_model)
    params = model.make_params(alpha=0, beta_1=0, beta_2=0, beta_3=0, beta_4=0, beta_5=0,
lambda_1=0, lambda_2=0, lambda_3=0, lambda_4=0, lambda_5=0, lambda_6=0, lambda_7=0,
lambda_8=0, lambda_9=0, lambda_10=0, lambda_11=0, lambda_12=0, lambda_13=0,
lambda_14=0, lambda_15=0, lambda_16=0, lambda_17=0, lambda_18=0, lambda_19=0,
lambda_20=0, lambda_21=0, lambda_22=0, lambda_23=0, lambda_24=0, lambda_25=0,
lambda_26=0, lambda_27=0, lambda_28=0, lambda_29=0, lambda_30=0, lambda_31=0)
    result = model.fit(Y, X=X, params=params)
    Y_pred = result.best_fit
    data["confirmed_pred"] = Y_pred
    r2 = r2_score(Y, Y_pred)
    print(result.fit_report())
    print("R square of Exponential model is ", r2)
    data["actual-pred"] = data[y] - data["%s_pred" % y]
    data["risk_index"] = (data["actual-pred"] - data["actual-pred"].mean()) / data["actual-
pred"].std()
    x = data["actual-pred"]
    mean, std = x.mean(), x.std(ddof=1)
    conf_intveral = stats.norm.interval(0.9, loc=mean, scale=std)
    print(conf_intveral[1], conf_intveral[0])
    data = data.sort_values("actual-pred", ascending=False)
    data = data[["city_cn", "city_en", "province", y, "%s_pred" % y, "actual-pred", "risk_index"]]
    data.to_csv("exponential_prediction_%s.csv" % date, encoding="utf-8-sig", index=False)

# Exponential-Logistic Dynamic Model
def exponential_logistic_model(X, gamma, omega, alpha, beta_1, beta_2, beta_3, beta_4,
beta_5, lambda_1, lambda_2, lambda_3, lambda_4, lambda_5, lambda_6, lambda_7,
lambda_8, lambda_9, lambda_10, lambda_11, lambda_12, lambda_13, lambda_14,
lambda_15, lambda_16, lambda_17, lambda_18, lambda_19, lambda_20, lambda_21,
lambda_22, lambda_23, lambda_24, lambda_25, lambda_26, lambda_27, lambda_28,
lambda_29, lambda_30, lambda_31):
    lambdas = [lambda_1, lambda_2, lambda_3, lambda_4, lambda_5, lambda_6, lambda_7,
lambda_8, lambda_9, lambda_10, lambda_11, lambda_12, lambda_13, lambda_14, lambda_15,
lambda_16, lambda_17, lambda_18, lambda_19, lambda_20, lambda_21, lambda_22,
lambda_23, lambda_24, lambda_25, lambda_26, lambda_27, lambda_28, lambda_29,
lambda_30, lambda_31]
    betas = [beta_1, beta_2, beta_3, beta_4, beta_5]
    N = len(X) - 31
    fix = 1
    for i in range(len(lambdas)):
        fix = fix * np.exp(lambdas[i] * X[i + N])
    R = alpha / (1 + np.exp(-1 * gamma * X[0] + omega))
    for i in range(1, N):
        R *= np.exp(betas[i - 1] * X[i])
    R *= fix
    R[np.isinf(R)] = INF
    R[np.isnan(R)] = NAN
    return R

# Exponential-Gompertz Dynamic Model
def exponential_gompertz_model(X, a, b, alpha, beta_1, beta_2, beta_3, beta_4, beta_5,
lambda_1, lambda_2, lambda_3, lambda_4, lambda_5, lambda_6, lambda_7, lambda_8,
lambda_9, lambda_10, lambda_11, lambda_12, lambda_13, lambda_14, lambda_15,
lambda_16, lambda_17, lambda_18, lambda_19, lambda_20, lambda_21, lambda_22,
lambda_23, lambda_24, lambda_25, lambda_26, lambda_27, lambda_28, lambda_29,
lambda_30, lambda_31):
    lambdas = [lambda_1, lambda_2, lambda_3, lambda_4, lambda_5, lambda_6, lambda_7,
lambda_8, lambda_9, lambda_10, lambda_11, lambda_12, lambda_13, lambda_14, lambda_15,
lambda_16, lambda_17, lambda_18, lambda_19, lambda_20, lambda_21, lambda_22,
lambda_23, lambda_24, lambda_25, lambda_26, lambda_27, lambda_28, lambda_29,
lambda_30, lambda_31]
    betas = [beta_1, beta_2, beta_3, beta_4, beta_5]
    N = len(X) - 31
    fix = 1
    for i in range(len(lambdas)):
        fix = fix * np.exp(lambdas[i] * X[i + N])
    R = alpha * np.power(a, np.power(b, X[0]))
    for i in range(1, N):
        R *= np.exp(betas[i - 1] * X[i])
    R *= fix
    R[np.isinf(R)] = INF
    R[np.isnan(R)] = NAN
    return R

# Exponential-Richards Dynamic Model
def exponential_richards_model(X, g, r, ti, alpha, beta_1, beta_2, beta_3, beta_4, beta_5,
lambda_1, lambda_2, lambda_3, lambda_4, lambda_5, lambda_6, lambda_7, lambda_8,
lambda_9, lambda_10, lambda_11, lambda_12, lambda_13, lambda_14, lambda_15,
lambda_16, lambda_17, lambda_18, lambda_19, lambda_20, lambda_21, lambda_22,
lambda_23, lambda_24, lambda_25, lambda_26, lambda_27, lambda_28, lambda_29,
lambda_30, lambda_31):
    lambdas = [lambda_1, lambda_2, lambda_3, lambda_4, lambda_5, lambda_6, lambda_7,
lambda_8, lambda_9, lambda_10, lambda_11, lambda_12, lambda_13, lambda_14, lambda_15,
lambda_16, lambda_17, lambda_18, lambda_19, lambda_20, lambda_21, lambda_22,
lambda_23, lambda_24, lambda_25, lambda_26, lambda_27, lambda_28, lambda_29,
lambda_30, lambda_31]
    betas = [beta_1, beta_2, beta_3, beta_4, beta_5]
    N = len(X) - 31
    fix = 1
    for i in range(len(lambdas)):
        fix = fix * np.exp(lambdas[i] * X[i + N])
    R = alpha / np.power((1 + g * np.exp(-r * (X[0] - ti))), 1 / g)
    for i in range(1, N):
        R *= np.exp(betas[i - 1] * X[i])
    R *= fix
    R[np.isinf(R)] = INF
    R[np.isnan(R)] = NAN
    return R

def exponential_dynamic_model_estimate(end_day=27, y="confirmed",
x=["wuhan_outflow", "gdp", "population"]):
    data = pd.read_csv(data_path)
    data.fillna(0, inplace=True)
    data = standardization(data, x)
    data = data[data["day"] <= end_day]
    prov = pd.read_csv("province_fix.csv")
    data = pd.merge(data, prov, on="province", how="left")
    data.fillna(0, inplace=True)
    Y = data[y]
    T = data["day"]
    X = [T]
    for k in x:
        X.append((data["%s(Ln)" % k] - data["%s(Ln_Mean)" % k]) / data["%s(Ln_Std)" % k])
    fix = province_dummies(data)
    X.append(fix.T)
    X = np.vstack(X)
    # fit the EL model
    ELmodel = Model(exponential_logistic_model)
    params = ELmodel.make_params(gamma=0.5, omega=1, alpha=1, beta_1=0, beta_2=0,
beta_3=0, beta_4=0, beta_5=0, lambda_1=0, lambda_2=0, lambda_3=0, lambda_4=0,
lambda_5=0, lambda_6=0, lambda_7=0, lambda_8=0, lambda_9=0, lambda_10=0,
lambda_11=0, lambda_12=0, lambda_13=0, lambda_14=0, lambda_15=0, lambda_16=0,
lambda_17=0, lambda_18=0, lambda_19=0, lambda_20=0, lambda_21=0, lambda_22=0,
lambda_23=0, lambda_24=0, lambda_25=0, lambda_26=0, lambda_27=0, lambda_28=0,
lambda_29=0, lambda_30=0, lambda_31=0)
    result = ELmodel.fit(Y, X=X, params=params)
    Y_pred = result.best_fit
    r2 = r2_score(Y, Y_pred)
    print(result.fit_report())
    print("R square of EL model is ", r2)
    data["confirmed_pred"] = Y_pred
    # fit the EG model
    EGmodel = Model(exponential_gompertz_model)
    params = EGmodel.make_params(a=0.5, b=0.5, alpha=1, beta_1=0, beta_2=0, beta_3=0,
beta_4=0, beta_5=0, lambda_1=0, lambda_2=0, lambda_3=0, lambda_4=0, lambda_5=0,
lambda_6=0, lambda_7=0, lambda_8=0, lambda_9=0, lambda_10=0, lambda_11=0,
lambda_12=0, lambda_13=0, lambda_14=0, lambda_15=0, lambda_16=0, lambda_17=0,
lambda_18=0, lambda_19=0, lambda_20=0, lambda_21=0, lambda_22=0, lambda_23=0,
lambda_24=0, lambda_25=0, lambda_26=0, lambda_27=0, lambda_28=0, lambda_29=0,
lambda_30=0, lambda_31=0)
    params["a"].set(min=0, max=1)
    params["b"].set(min=0, max=1)
    result = EGmodel.fit(Y, X=X, params=params)
    Y_pred = result.best_fit
    r2 = r2_score(Y, Y_pred)
    print(result.fit_report())
    print("R square of EG model is ", r2)
    # fit the ER model
    ERmodel = Model(exponential_richards_model)
    params = ERmodel.make_params(g=0.5, r=1, ti=1, alpha=1, beta_1=0, beta_2=0, beta_3=0,
beta_4=0, beta_5=0, lambda_1=0, lambda_2=0, lambda_3=0, lambda_4=0, lambda_5=0,
lambda_6=0, lambda_7=0, lambda_8=0, lambda_9=0, lambda_10=0, lambda_11=0,
lambda_12=0, lambda_13=0, lambda_14=0, lambda_15=0, lambda_16=0, lambda_17=0,
lambda_18=0, lambda_19=0, lambda_20=0, lambda_21=0, lambda_22=0, lambda_23=0,
lambda_24=0, lambda_25=0, lambda_26=0, lambda_27=0, lambda_28=0, lambda_29=0,
lambda_30=0, lambda_31=0)
    result = ERmodel.fit(Y, X=X, params=params)
    Y_pred = result.best_fit
    r2 = r2_score(Y, Y_pred)
    print(result.fit_report())
    print("R square of ER model is ", r2)


# Daily Power Static Model
def power_model(X, alpha, beta_1, beta_2, beta_3, beta_4, beta_5, lambda_1, lambda_2,
lambda_3, lambda_4, lambda_5, lambda_6, lambda_7, lambda_8, lambda_9, lambda_10,
lambda_11, lambda_12, lambda_13, lambda_14, lambda_15, lambda_16, lambda_17,
lambda_18, lambda_19, lambda_20, lambda_21, lambda_22, lambda_23, lambda_24,
lambda_25, lambda_26, lambda_27, lambda_28, lambda_29, lambda_30, lambda_31):
    lambdas = [lambda_1, lambda_2, lambda_3, lambda_4, lambda_5, lambda_6, lambda_7,
lambda_8, lambda_9, lambda_10, lambda_11, lambda_12, lambda_13, lambda_14, lambda_15,
lambda_16, lambda_17, lambda_18, lambda_19, lambda_20, lambda_21, lambda_22,
lambda_23, lambda_24, lambda_25, lambda_26, lambda_27, lambda_28, lambda_29,
lambda_30, lambda_31]
    betas = [beta_1, beta_2, beta_3, beta_4, beta_5]
    N = len(X) - 31
    fix = 1
    for i in range(len(lambdas)):
        fix = fix * np.exp(lambdas[i] * X[i + N])
    R = alpha
    for i in range(N):
        R *= np.power(X[i], betas[i])
    R *= fix
    R[np.isinf(R)] = INF
    R[np.isnan(R)] = NAN
    return R

def power_static_model_estimate(date="2020-01-28", y="confirmed",
x=["wuhan_outflow", "gdp", "population"]):
    data = pd.read_csv(data_path)
    province = pd.read_csv("province_fix.csv")
    data = pd.merge(data, province, on="province", how="left")
    data.fillna(0, inplace=True)
    data = normalization(data, x)
    data = data[data["date"] == date]
    Y = data[y]
    X = []
    for k in x:
        X.append(data[k] / data["%s(Sum)" % k])
    fix = province_dummies(data)
    X.append(fix.T)
    X = np.vstack(X)
    model = Model(power_model)
    params = model.make_params(alpha=1, beta_1=0.5, beta_2=0.5, beta_3=0.5, beta_4=0.5,
beta_5=0.5, lambda_1=0, lambda_2=0, lambda_3=0, lambda_4=0, lambda_5=0, lambda_6=0,
lambda_7=0, lambda_8=0, lambda_9=0, lambda_10=0, lambda_11=0, lambda_12=0,
lambda_13=0, lambda_14=0, lambda_15=0, lambda_16=0, lambda_17=0, lambda_18=0,
lambda_19=0, lambda_20=0, lambda_21=0, lambda_22=0, lambda_23=0, lambda_24=0,
lambda_25=0, lambda_26=0, lambda_27=0, lambda_28=0, lambda_29=0, lambda_30=0,
lambda_31=0)
    result = model.fit(Y, X=X, params=params)
    Y_pred = result.best_fit
    data["confirmed_pred"] = Y_pred
    r2 = r2_score(Y, Y_pred)
    print(result.fit_report())
    print("R square of Power model is ", r2)

# Power-Logistic Dynamic Model
def power_logistic_model(X, gamma, omega, alpha, beta_1, beta_2, beta_3, beta_4, beta_5,
lambda_1, lambda_2, lambda_3, lambda_4, lambda_5, lambda_6, lambda_7, lambda_8,
lambda_9, lambda_10, lambda_11, lambda_12, lambda_13, lambda_14, lambda_15,
lambda_16, lambda_17, lambda_18, lambda_19, lambda_20, lambda_21, lambda_22,
lambda_23, lambda_24, lambda_25, lambda_26, lambda_27, lambda_28, lambda_29,
lambda_30, lambda_31):
    lambdas = [lambda_1, lambda_2, lambda_3, lambda_4, lambda_5, lambda_6, lambda_7,
lambda_8, lambda_9, lambda_10, lambda_11, lambda_12, lambda_13, lambda_14, lambda_15,
lambda_16, lambda_17, lambda_18, lambda_19, lambda_20, lambda_21, lambda_22,
lambda_23, lambda_24, lambda_25, lambda_26, lambda_27, lambda_28, lambda_29,
lambda_30, lambda_31]
    betas = [beta_1, beta_2, beta_3, beta_4, beta_5]
    N = len(X) - 31
    fix = 1
    for i in range(len(lambdas)):
        fix = fix * np.exp(lambdas[i] * X[i + N])
    R = alpha / (1 + np.exp(-1 * gamma * X[0] + omega))
    for i in range(1, N):
        R *= np.power(X[i], betas[i - 1])
    R *= fix
    R[np.isinf(R)] = INF
    R[np.isnan(R)] = NAN
    return R

# Power-Gompertz Dynamic Model
def power_gompertz_model(X, a, b, alpha, beta_1, beta_2, beta_3, beta_4, beta_5,
lambda_1, lambda_2, lambda_3, lambda_4, lambda_5, lambda_6, lambda_7, lambda_8,
lambda_9, lambda_10, lambda_11, lambda_12, lambda_13, lambda_14, lambda_15,
lambda_16, lambda_17, lambda_18, lambda_19, lambda_20, lambda_21, lambda_22,
lambda_23, lambda_24, lambda_25, lambda_26, lambda_27, lambda_28, lambda_29,
lambda_30, lambda_31):
    lambdas = [lambda_1, lambda_2, lambda_3, lambda_4, lambda_5, lambda_6, lambda_7,
lambda_8, lambda_9, lambda_10, lambda_11, lambda_12, lambda_13, lambda_14, lambda_15,
lambda_16, lambda_17, lambda_18, lambda_19, lambda_20, lambda_21, lambda_22,
lambda_23, lambda_24, lambda_25, lambda_26, lambda_27, lambda_28, lambda_29,
lambda_30, lambda_31]
    betas = [beta_1, beta_2, beta_3, beta_4, beta_5]
    N = len(X) - 31
    fix = 1
    for i in range(len(lambdas)):
        fix = fix * np.exp(lambdas[i] * X[i + N])
    R = alpha * np.power(a, np.power(b, X[0]))
    for i in range(1, N):
        R *= np.power(X[i], betas[i - 1])
    R *= fix
    R[np.isinf(R)] = INF
    R[np.isnan(R)] = NAN
    return R

# Power-Richards Dynamic Model
```

```
def power_richards_model(X, g, r, ti, alpha, beta_1, beta_2, beta_3, beta_4, beta_5,
lambda_1, lambda_2, lambda_3, lambda_4, lambda_5, lambda_6, lambda_7, lambda_8,
lambda_9, lambda_10, lambda_11, lambda_12, lambda_13, lambda_14, lambda_15,
lambda_16, lambda_17, lambda_18, lambda_19, lambda_20, lambda_21, lambda_22,
lambda_23, lambda_24, lambda_25, lambda_26, lambda_27, lambda_28, lambda_29,
lambda_30, lambda_31):
    lambdas = [lambda_1, lambda_2, lambda_3, lambda_4, lambda_5, lambda_6, lambda_7,
lambda_8, lambda_9, lambda_10, lambda_11, lambda_12, lambda_13, lambda_14, lambda_15,
lambda_16, lambda_17, lambda_18, lambda_19, lambda_20, lambda_21, lambda_22,
lambda_23, lambda_24, lambda_25, lambda_26, lambda_27, lambda_28, lambda_29,
lambda_30, lambda_31]
    betas = [beta_1, beta_2, beta_3, beta_4, beta_5]
    N = len(X) - 31
    fix = 1
    for i in range(len(lambdas)):
        fix = fix * np.exp(lambdas[i] * X[i + N])
    R = alpha / np.power((1 + g * np.exp(-r * (X[0] - ti))), 1 / g)
    for i in range(1, N):
        R *= np.power(X[i], betas[i - 1])
    R *= fix
    R[np.isinf(R)] = INF
    R[np.isnan(R)] = NAN
    return R

def power_dynamic_model_estimate(end_day=27, y="confirmed", x=["wuhan_outflow",
"gdp", "population"]):
    data = pd.read_csv(data_path)
    data.fillna(0, inplace=True)
    data = normalization(data, x)
    data = data[data["day"] <= end_day]
    province = pd.read_csv("province_fix.csv")
    data = pd.merge(data, province, on="province", how="left")
    Y = data[y]
    T = data["day"]
    X = [T]
    for k in x:
        X.append(data[k] / data["%s(Sum)" % k])
    fix = province_dummies(data)
    X.append(fix.T)
    X = np.vstack(X)
    # fit the PL model
    PLmodel = Model(power_logistic_model)
    params = PLmodel.make_params(gamma=1, omega=1, alpha=0.5, beta_1=0.5, beta_2=0.5,
beta_3=0.5, beta_4=0.5, beta_5=0.5, lambda_1=0, lambda_2=0, lambda_3=0, lambda_4=0,
lambda_5=0, lambda_6=0, lambda_7=0, lambda_8=0, lambda_9=0, lambda_10=0,
lambda_11=0, lambda_12=0, lambda_13=0, lambda_14=0, lambda_15=0, lambda_16=0,
lambda_17=0, lambda_18=0, lambda_19=0, lambda_20=0, lambda_21=0, lambda_22=0,
lambda_23=0, lambda_24=0, lambda_25=0, lambda_26=0, lambda_27=0, lambda_28=0,
lambda_29=0, lambda_30=0, lambda_31=0)
    result = PLmodel.fit(Y, X=X, params=params)
    Y_pred = result.best_fit
    r2 = r2_score(Y, Y_pred)
    print(result.fit_report())
    print("R square of PL model is ", r2)
    # fit the PG model
    PGmodel = Model(power_gompertz_model)
    params = PGmodel.make_params(a=0.5, b=0.5, alpha=0.5, beta_1=0.5, beta_2=0.5,
beta_3=0.5, beta_4=0.5, beta_5=0.5, lambda_1=0, lambda_2=0, lambda_3=0, lambda_4=0,
lambda_5=0, lambda_6=0, lambda_7=0, lambda_8=0, lambda_9=0, lambda_10=0,
lambda_11=0, lambda_12=0, lambda_13=0, lambda_14=0, lambda_15=0, lambda_16=0,
lambda_17=0, lambda_18=0, lambda_19=0, lambda_20=0, lambda_21=0, lambda_22=0,
lambda_23=0, lambda_24=0, lambda_25=0, lambda_26=0, lambda_27=0, lambda_28=0,
lambda_29=0, lambda_30=0, lambda_31=0)
    params["a"].set(min=0, max=1)
    params["b"].set(min=0, max=1)
    result = PGmodel.fit(Y, X=X, params=params)
    Y_pred = result.best_fit
    r2 = r2_score(Y, Y_pred)
    print(result.fit_report())
    print("R square of PG model is ", r2)
    # fit the PR model
    PRmodel = Model(power_richards_model)
    params = PRmodel.make_params(g=0.5, r=1, ti=1, alpha=0.5, beta_1=0.5, beta_2=0.5,
beta_3=0.5, beta_4=0.5, beta_5=0.5, lambda_1=0, lambda_2=0, lambda_3=0, lambda_4=0,
lambda_5=0, lambda_6=0, lambda_7=0, lambda_8=0, lambda_9=0, lambda_10=0,
lambda_11=0, lambda_12=0, lambda_13=0, lambda_14=0, lambda_15=0, lambda_16=0,
lambda_17=0, lambda_18=0, lambda_19=0, lambda_20=0, lambda_21=0, lambda_22=0,
lambda_23=0, lambda_24=0, lambda_25=0, lambda_26=0, lambda_27=0, lambda_28=0,
lambda_29=0, lambda_30=0, lambda_31=0)
    result = PRmodel.fit(Y, X=X, params=params)
    Y_pred = result.best_fit
    r2 = r2_score(Y, Y_pred)
    print(result.fit_report())
    print("R square of PR model is ", r2)

def negative_number_to_zero(x):
    if x < 0:
        return 0
    else:
        return x

# df(x)/dt, Exponential-Logistic Dynamic Increased Model
```

```
def exponential_logistic_increased_model(X, gamma, omega, alpha, beta_1, beta_2, beta_3,
beta_4, beta_5, lambda_1, lambda_2, lambda_3, lambda_4, lambda_5, lambda_6,
lambda_7, lambda_8, lambda_9, lambda_10, lambda_11, lambda_12, lambda_13,
lambda_14, lambda_15, lambda_16, lambda_17, lambda_18, lambda_19, lambda_20,
lambda_21, lambda_22, lambda_23, lambda_24, lambda_25, lambda_26, lambda_27,
lambda_28, lambda_29, lambda_30, lambda_31):
    lambdas = [lambda_1, lambda_2, lambda_3, lambda_4, lambda_5, lambda_6, lambda_7,
lambda_8, lambda_9, lambda_10, lambda_11, lambda_12, lambda_13, lambda_14, lambda_15,
lambda_16, lambda_17, lambda_18, lambda_19, lambda_20, lambda_21, lambda_22,
lambda_23, lambda_24, lambda_25, lambda_26, lambda_27, lambda_28, lambda_29,
lambda_30, lambda_31]
    betas = [beta_1, beta_2, beta_3, beta_4, beta_5]
    N = len(X) - 31
    fix = 1
    for i in range(len(lambdas)):
        fix = fix * np.exp(lambdas[i] * X[i + N])
    R = alpha * np.exp(-1 * gamma * X[0] + omega) * gamma / np.power((1 + np.exp(-1 *
gamma * X[0] + omega)), 2)
    for i in range(1, N):
        R *= np.exp(betas[i - 1] * X[i])
    R *= fix
    R[np.isinf(R)] = INF
    R[np.isnan(R)] = NAN
    return R

# df(x)/dt, Exponential-Gompertz Dynamic Increased Model
def exponential_gompertz_increased_model(X, a, b, alpha, beta_1, beta_2, beta_3, beta_4,
beta_5, lambda_1, lambda_2, lambda_3, lambda_4, lambda_5, lambda_6, lambda_7,
lambda_8, lambda_9, lambda_10, lambda_11, lambda_12, lambda_13, lambda_14,
lambda_15, lambda_16, lambda_17, lambda_18, lambda_19, lambda_20, lambda_21,
lambda_22, lambda_23, lambda_24, lambda_25, lambda_26, lambda_27, lambda_28,
lambda_29, lambda_30, lambda_31):
    lambdas = [lambda_1, lambda_2, lambda_3, lambda_4, lambda_5, lambda_6, lambda_7,
lambda_8, lambda_9, lambda_10, lambda_11, lambda_12, lambda_13, lambda_14, lambda_15,
lambda_16, lambda_17, lambda_18, lambda_19, lambda_20, lambda_21, lambda_22,
lambda_23, lambda_24, lambda_25, lambda_26, lambda_27, lambda_28, lambda_29,
lambda_30, lambda_31]
    betas = [beta_1, beta_2, beta_3, beta_4, beta_5]
    N = len(X) - 31
    fix = 1
    for i in range(len(lambdas)):
        fix = fix * np.exp(lambdas[i] * X[i + N])
    R = alpha * np.power(a, np.power(b, X[0])) * np.power(b, X[0]) * np.log(a) * np.log(b)
    for i in range(1, N):
        R *= np.exp(betas[i - 1] * X[i])
    R *= fix
    R[np.isinf(R)] = INF
    R[np.isnan(R)] = NAN
    return R

def exponential_dynamic_increased_model_estimate(end_day=27, y="confirmed",
x=["wuhan_outflow", "gdp", "population"]):
    data = pd.read_csv(data_path)
    data.fillna(0, inplace=True)
    data = standardization(data, x)
    data = data[data["day"] <= end_day]
    prov = pd.read_csv("province_fix.csv")
    data = pd.merge(data, prov, on="province", how="left")
    data.fillna(0, inplace=True)
    data[y] = data[y] - data["%s_pre" % y]
    data[y] = data[y].map(negative_number_to_zero)
    Y = data[y]
    T = data["day"]
    X = [T]
    for k in x:
        X.append((data["%s(Ln)" % k] - data["%s(Ln_Mean)" % k]) / data["%s(Ln_Std)" % k])
    fix = province_dummies(data)
    X.append(fix.T)
    X = np.vstack(X)
    # fit the EL model
    ELmodel = Model(exponential_logistic_increased_model)
    params = ELmodel.make_params(gamma=0.5, omega=1, alpha=1, beta_1=0, beta_2=0,
beta_3=0, beta_4=0, beta_5=0, lambda_1=0, lambda_2=0, lambda_3=0, lambda_4=0,
lambda_5=0, lambda_6=0, lambda_7=0, lambda_8=0, lambda_9=0, lambda_10=0,
lambda_11=0, lambda_12=0, lambda_13=0, lambda_14=0, lambda_15=0, lambda_16=0,
lambda_17=0, lambda_18=0, lambda_19=0, lambda_20=0, lambda_21=0, lambda_22=0,
lambda_23=0, lambda_24=0, lambda_25=0, lambda_26=0, lambda_27=0, lambda_28=0,
lambda_29=0, lambda_30=0, lambda_31=0)
    result = ELmodel.fit(Y, X=X, params=params)
    Y_pred = result.best_fit
    r2 = r2_score(Y, Y_pred)
    print(result.fit_report())
    print("R square of EL model is ", r2)
    # fit the EG model
    EGmodel = Model(exponential_gompertz_increased_model)
    params = EGmodel.make_params(a=0.5, b=0.5, alpha=1, beta_1=0, beta_2=0, beta_3=0,
beta_4=0, beta_5=0, lambda_1=0, lambda_2=0, lambda_3=0, lambda_4=0, lambda_5=0,
lambda_6=0, lambda_7=0, lambda_8=0, lambda_9=0, lambda_10=0, lambda_11=0,
lambda_12=0, lambda_13=0, lambda_14=0, lambda_15=0, lambda_16=0, lambda_17=0,
lambda_18=0, lambda_19=0, lambda_20=0, lambda_21=0, lambda_22=0, lambda_23=0,
lambda_24=0, lambda_25=0, lambda_26=0, lambda_27=0, lambda_28=0, lambda_29=0,
lambda_30=0, lambda_31=0)
```

```
    params["a"].set(min=0, max=1)
    params["b"].set(min=0, max=1)
    result = EGmodel.fit(Y, X=X, params=params)
    Y_pred = result.best_fit
    r2 = r2_score(Y, Y_pred)
    print(result.fit_report())
    print("R square of EG model is ", r2)

# df(x)/dt, Power-Logistic Dynamic Increased Model
def power_logistic_increased_model(X, gamma, omega, alpha, beta_1, beta_2, beta_3,
beta_4, beta_5, lambda_1, lambda_2, lambda_3, lambda_4, lambda_5, lambda_6,
lambda_7, lambda_8, lambda_9, lambda_10, lambda_11, lambda_12, lambda_13,
lambda_14, lambda_15, lambda_16, lambda_17, lambda_18, lambda_19, lambda_20,
lambda_21, lambda_22, lambda_23, lambda_24, lambda_25, lambda_26, lambda_27,
lambda_28, lambda_29, lambda_30, lambda_31):
    lambdas = [lambda_1, lambda_2, lambda_3, lambda_4, lambda_5, lambda_6, lambda_7,
lambda_8, lambda_9, lambda_10, lambda_11, lambda_12, lambda_13, lambda_14, lambda_15,
lambda_16, lambda_17, lambda_18, lambda_19, lambda_20, lambda_21, lambda_22,
lambda_23, lambda_24, lambda_25, lambda_26, lambda_27, lambda_28, lambda_29,
lambda_30, lambda_31]
    betas = [beta_1, beta_2, beta_3, beta_4, beta_5]
    N = len(X) - 31
    fix = 1
    for i in range(len(lambdas)):
        fix = fix * np.exp(lambdas[i] * X[i + N])
    R = alpha * np.exp(-1 * gamma * X[0] + omega) * gamma / np.power((1 + np.exp(-1 *
gamma * X[0] + omega)), 2)
    for i in range(1, N):
        R *= np.power(X[i], betas[i - 1])
    R *= fix
    R[np.isinf(R)] = INF
    R[np.isnan(R)] = NAN
    return R

# df(x)/dt, Power-Gompertz Dynamic Increased Model
def power_gompertz_increased_model(X, a, b, alpha, beta_1, beta_2, beta_3, beta_4,
beta_5, lambda_1, lambda_2, lambda_3, lambda_4, lambda_5, lambda_6, lambda_7,
lambda_8, lambda_9, lambda_10, lambda_11, lambda_12, lambda_13, lambda_14,
lambda_15, lambda_16, lambda_17, lambda_18, lambda_19, lambda_20, lambda_21,
lambda_22, lambda_23, lambda_24, lambda_25, lambda_26, lambda_27, lambda_28,
lambda_29, lambda_30, lambda_31):
    lambdas = [lambda_1, lambda_2, lambda_3, lambda_4, lambda_5, lambda_6, lambda_7,
lambda_8, lambda_9, lambda_10, lambda_11, lambda_12, lambda_13, lambda_14, lambda_15,
lambda_16, lambda_17, lambda_18, lambda_19, lambda_20, lambda_21, lambda_22,
lambda_23, lambda_24, lambda_25, lambda_26, lambda_27, lambda_28, lambda_29,
lambda_30, lambda_31]
```

```
    betas = [beta_1, beta_2, beta_3, beta_4, beta_5]
    N = len(X) - 31
    fix = 1
    for i in range(len(lambdas)):
        fix = fix * np.exp(lambdas[i] * X[i + N])
    R = alpha * np.power(a, np.power(b, X[0])) * np.power(b, X[0]) * np.log(a) * np.log(b)
    for i in range(1, N):
        R *= np.power(X[i], betas[i - 1])
    R *= fix
    R[np.isinf(R)] = INF
    R[np.isnan(R)] = NAN
    return R

def power_dynamic_increased_model_estimate(end_day=27, y="confirmed",
x=["wuhan_outflow", "gdp", "population"]):
    data = pd.read_csv(data_path)
    data.fillna(0, inplace=True)
    data = normalization(data, x)
    data = data[data["day"] <= end_day]
    prov = pd.read_csv("province_fix.csv")
    data = pd.merge(data, prov, on="province", how="left")
    data.fillna(0, inplace=True)
    data[y] = data[y] - data["%s_pre" % y]
    data[y] = data[y].map(negative_number_to_zero)
    Y = data[y]
    T = data["day"]
    X = [T]
    for k in x:
        X.append(data[k] / data["%s(Sum)" % k])
    fix = province_dummies(data)
    X.append(fix.T)
    X = np.vstack(X)
    # fit the PL model
    PLmodel = Model(power_logistic_increased_model)
    params = PLmodel.make_params(gamma=0.5, omega=1, alpha=1, beta_1=0, beta_2=0,
beta_3=0, beta_4=0, beta_5=0, lambda_1=0, lambda_2=0, lambda_3=0, lambda_4=0,
lambda_5=0, lambda_6=0, lambda_7=0, lambda_8=0, lambda_9=0, lambda_10=0,
lambda_11=0, lambda_12=0, lambda_13=0, lambda_14=0, lambda_15=0, lambda_16=0,
lambda_17=0, lambda_18=0, lambda_19=0, lambda_20=0, lambda_21=0, lambda_22=0,
lambda_23=0, lambda_24=0, lambda_25=0, lambda_26=0, lambda_27=0, lambda_28=0,
lambda_29=0, lambda_30=0, lambda_31=0)
    result = PLmodel.fit(Y, X=X, params=params)
    Y_pred = result.best_fit
    r2 = r2_score(Y, Y_pred)
    print(result.fit_report())
    print("R square of PL model is ", r2)
    # fit the PG model
    PGmodel = Model(power_gompertz_increased_model)
    params = PGmodel.make_params(a=0.5, b=0.5, alpha=1, beta_1=0, beta_2=0, beta_3=0,
beta_4=0, beta_5=0, lambda_1=0, lambda_2=0, lambda_3=0, lambda_4=0, lambda_5=0,
lambda_6=0, lambda_7=0, lambda_8=0, lambda_9=0, lambda_10=0, lambda_11=0,
lambda_12=0, lambda_13=0, lambda_14=0, lambda_15=0, lambda_16=0, lambda_17=0,
lambda_18=0, lambda_19=0, lambda_20=0, lambda_21=0, lambda_22=0, lambda_23=0,
lambda_24=0, lambda_25=0, lambda_26=0, lambda_27=0, lambda_28=0, lambda_29=0,
lambda_30=0, lambda_31=0)
    params["a"].set(min=0, max=1)
    params["b"].set(min=0, max=1)
    result = PGmodel.fit(Y, X=X, params=params)
    Y_pred = result.best_fit
    r2 = r2_score(Y, Y_pred)
    print(result.fit_report())
    print("R square of PG model is ", r2)

if __name__ == "__main__":
    exponential_static_model_estimate()
    exponential_dynamic_model_estimate()
    exponential_dynamic_increased_model_estimate()
    power_static_model_estimate()
    power_dynamic_model_estimate()
    power_dynamic_increased_model_estimate()
```

References

30. Pew Research Center. Retrieved from https://www.pewresearch.org/fact-tank/2017/03/16/china-outpaces-india-in-internet-access-smartphone-ownership (2017).

31. Deloitte. *Chinese consumers at the forefront of digital technologies: China Mobile Consumer Survey 2018*. Retrieved from: https://www2.deloitte.com/content/dam/Deloitte/cn/Documents/technology-media-telecommunications/deloitte-cn-2018-mobile-consumer-survey-en-190121.pdf (2019).

32. http://en.people.cn/n3/2019/0119/c90000-9539458.html

33. Barbosa H. et al. (2018) Human mobility: Models and applications. *Phys. Rep.* 734: 1-74.

34. Song, C., Koren, T., Wang, P., & Barabási, A. L. (2010). Modelling the scaling properties of human mobility. *Nat. Phys. 6*(10), 818-823.

35. Song, C., Qu, Z., Blumm, N. & Barabási, A. L. Limits of predictability in human mobility. *Science* **327**, 1018–1021 (2010).

36. Simini, F., González, M. C., Maritan, A. & Barabási, A. L. A universal model for mobility and migration patterns. *Nature* **484**, 96–100 (2012).

37. Yan X., Wang W., Gao Z., & Lai Y Universal model of individual and population mobility on diverse spatial scales, *Nat. Commun.* **8**, 1639 (2017).

38. Bengtsson, L., Lu, X., Thorson, A., Garfield, R. & von Schreeb, J. Improved Response to Disasters and Outbreaks by Tracking Population Movements with Mobile Phone Network Data: A Post-Earthquake Geospatial Study in Haiti. *PLoS Med* **8**, p. e1001083 (2011).

39. Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. L., Noor, A. M., Snow, R. W., & Buckee, C. O. Quantifying the impact of human mobility on malaria. *Science* **338**, 267-70 (2012).

40. Bengtsson, L., J. Gaudart, X. Lu, S. Moore, E. Wetter, K. Sallah, S. Rebaudet, & R. Piarroux, Using Mobile Phone Data to Predict the Spatial Spread of Cholera. *Sci. Rep.* **5**, 8923 (2015).

41. Finger, F., Genolet, T., Mari, L., de Magny, G. C., Manga, N. M., Rinaldo, A., & Bertuzzo, E. Mobile phone data highlights the role of mass gatherings in the spreading of cholera outbreaks. *Proc. Natl. Acad. Sci. USA* **113**, 6421–6426 (2016).

42. Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., ... & Yu, T. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet*. (2020).

43. Yang, Y., Lu, Q., Liu, M., Wang, Y., Zhang, A., Jalali, N., … & Fang, L. Epidemiological and clinical features of the 2019 novel coronavirus outbreak in China. Preprint at https://www.medrxiv.org/content/10.1101/2020.02.10.20021675v1 (2020).

44. Chan, J. F. W., Yuan, S., Kok, K. H., To, K. K. W., Chu, H., Yang, J., ... & Tsoi, H. W. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet*. (2020).

45. World Health Organization. *Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19)*. Retrieved at https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf (2020).

46. https://www.worldometers.info/coronavirus/coronavirus-incubation-period/

47. Chinese Center for Disease Control and Prevention. Retrieved from http://www.gov.cn/xinwen/2020-02/12/content_5477538.htm

48. Newville, M., Stensitzki, T., Allen, D. B., Rawlik, M., Ingargiola, A., & Nelson, A. LMFIT: Non-linear least-square minimization and curve-fitting for Python. *Astrophysics Source Code Library* (2016).
49. Laskawski, M. S., & Kazala, R. Identification of Parameters of Inertial Models Using the Levenberg-Marquardt Method. In *2018 Conference on Electrotechnology: Processes, Models, Control and Computer Science (EPMCCS)* (pp. 1-5). IEEE (2018, November).
50. Zwietering, M. H., Jongenburger, I., Rombouts, F. M. & Van't Riet, K. J. A. E. M. Modeling of the bacterial growth curve. *Appl. Environ. Microbiol.* **56**, 1875-1881. (1990).
51. Nishiura, H., Tsuzuki, S., Yuan, B., Yamaguchi, T., & Asai, Y. Transmission dynamics of cholera in Yemen, 2017: a real time forecasting. *Theor. Biol. Med. Model.* **14**, 14 (2017).
52. Bürger R., Chowell G., Lara-Dı́ıaz L. Comparative analysis of phenomenological growth models applied to epidemic outbreaks. *Math. Biosci. Eng*. **16**, 4250–4273 (2019).
53. Roosa, K., Lee, Y., Luo, R., Kirpich, A., Rothenberg, R., Hyman, J. M., Yan, P., & Chowell, G. Short-term Forecasts of the COVID-19 Epidemic in Guangdong and Zhejiang, China: February 13-23, 2020. *J Clin. Med.* **9**, 596 (2020).
54. https://tech.sina.cn/2020-02-08/detail-iimxyqvz1180050.d.html
55. http://www.bjnews.com.cn/news/2020/02/11/687835.html
56. Zipf, G. The P1 P2 / D hypothesis: On the intercity movement of persons. *Amer. Sociol. Rev.* **11**, 677–686 (1946).
57. Viboud, C., Bjornstad, O.N., Smith, D.L., Simonsen, L., Miller, M.A., & Grenfell, B.T. Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science* **312**, 447-451 (2006).
58. Balcan D, Colizza V, Gonçalves B, Hu H, Ramasco JJ, & Vespignani A. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proc. Natl. Acad. Sci. USA* **106**, 21484-9 (2009).
59. Li, X., Tian, H., Lai, D., & Zhang, Z. Validation of the gravity model in predicting the global spread of influenza. *International Journal of Environmental Research and Public Health* **8**, 3134–3143 (2011).
60. Barrios, J. M., Verstraeten, W. W., Maes, P., Aerts, J. M., Farifteh, J., & Coppin, P. Using the gravity model to estimate the spatial spread of vector-borne diseases. *International Journal of Environmental Research and Public Health* **9**, 4346–4364 (2012).
61. Wesolowski, A., Qureshi, T., Boni, M. F., Sundsøy, P. R., Johansson, M. A., Rasheed, S. B., Engø-Monsen, K., & Buckee, C. O. Impact of human mobility on the emergence of dengue epidemics in Pakistan. *Proc. Natl. Acad. Sci. USA* **112**, 11887–11892 (2015).
62. Brockmann, D. & Helbing, D. The hidden geometry of complex, network-driven contagion phenomena. *Science* 342, 1337-1342 (2013).
63. Eggo, R. M., Cauchemez, S., & Ferguson, N. M. Spatial dynamics of the 1918 influenza pandemic in England, Wales and the United States. *J. R. Soc. Interface* **8**, 233–243 (2011).
64. Charu, V., Zeger, S., Gog, J., Bjørnstad, O. N., Kissler, S., Simonsen, L., Grenfell, B. T., & Viboud, C. Human mobility and the spatial transmission of influenza in the United States. *PLoS Comput. Biol.* **13**, e1005382 (2017).

65. Bonney, P. J., Malladi, S., Boender, G. J., Weaver, J. T., Ssematimba, A., Halvorson, D. A., & Cardona, C. J. Spatial transmission of H5N2 highly pathogenic avian influenza between Minnesota poultry premises during the 2015 outbreak. *PloS One* **13**, e0204262 (2018).
66. Churakov, M., Villabona-Arenas, C. J., Kraemer, M., Salje, H., & Cauchemez, S. Spatio-temporal dynamics of dengue in Brazil: Seasonal travelling waves and determinants of regional synchrony. *PLoS Neglect. Tropical Diseases* **13**, e0007012 (2019).
67. Fang, L. Q., Yang, Y., Jiang, … & Cao, W. C. Transmission dynamics of Ebola virus disease and intervention effectiveness in Sierra Leone. *Proc. Natl. Acad. Sci. USA* **113**, 4488–4493 (2016).
68. Tang, X., Yang, Y., Yu, H. J., Liao, Q. H., & Bliznyuk, N. A Spatio-Temporal Modeling Framework for Surveillance Data of Multiple Infectious Pathogens with Small Laboratory Validation Sets. *J. Am. Stat. Assoc.* **114**, 1561–1573 (2019).
69. Zhang, K., Peters, J., Janzing, D., & Schölkopf, B. Kernel-based conditional independence test and application in causal discovery, *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, July 2011*, 804–813 (2012).
70. Grömping, U. Variable importance assessment in regression: linear regression versus random forest. *Am. Stat.*, **63**, 308-319 (2009).
71. Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. Permutation importance: a corrected feature importance measure. *Bioinformatics* **26**, 1340-1347 (2010).
72. Du, Z., Wang, L., Cauchemez, S., Xu, X., Wang, X., Cowling, B. J. & Meyers, L. A. (2020). Risk of 2019 novel coronavirus importations throughout China prior to the Wuhan quarantine. Prepint at https://www.medrxiv.org/content/10.1101/2020.01.28.20019299v3 (2020).
73. Massey, D. S. & García España, F. The Social Process of International Migration, *Science* **237**, 733-738 (1987).
74. Palloni, A, Massey, D. S., Ceballos, M., Espinosa, K. & Spittel, M. Social Capital and International Migration: A Test Using Information on Family Networks. *Am. J. Sociol.* **106**, 1262-1298 (2001).
75. Korinek, K., Entwisle, B. & Jampaklay, A. Through thick and thin: Layers of social ties and urban settlement among Thai migrants. *Am. Sociol. Rev.* **70**, 779-800 (2005).