Population flow drives spatio-temporal distribution of COVID-19 in China

https://doi.org/10.1038/s41586-020-2284-y

Received: 18 February 2020

Accepted: 21 April 2020

Published online: 29 April 2020

Check for updates

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

Sudden, large-scale and diffuse human migration can amplify localized outbreaks of disease into widespread epidemics[1–4]. Rapid and accurate tracking of aggregate population flows may therefore be epidemiologically informative. Here we use 11,478,484 counts of mobile phone data from 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 severe acute respiratory syndrome coronavirus 2 (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 operationalize 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 important in the context of the outbreak of COVID-19 in China and the rest of the world. This outbreak emerged in Wuhan (a prefecture-level city in the province of Hubei) in the run-up to the Chinese Lunar New Year's Eve on 24 January 2020, which is associated with the annual Chunyun mass migration (which can involve as many as three billion trips). The potential scale and range of the diffusion of the outbreak was particularly alarming given the position of Wuhan 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 conceptualized epidemiological morbidity as a function of the movement of the human population from a disease epicentre. We therefore normalize disease risk to the population inflow from Wuhan rather than to the size of the local population.

Our approach differs from previous studies in which individual mobility and disease spread[1–4,12,13] was linked, as we used real-time data about actual movement, focussed on aggregate population flows rather than individual tracking, and implemented a new modelling approach. That is, other recent studies on COVID-19 have used historical population flow data (for example, data on Chunyun migrations from previous years) to estimate case exportation during the current outbreak[14–18]. However, the benefits of observing rather than estimating population movements are substantial as 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, in contrast to previous approaches to epidemiological modelling[12–18], we take advantage of detailed data about the population flow that emanated from 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 the total aggregate population outflow from Wuhan before the region was quarantined on 23 January 2020, we used country-wide data (provided by a major national carrier) that tracked all of the movements out of Wuhan between 1 January and 24 January 2020. The onset of symptoms of the first recorded case of COVID-19 in Wuhan was 1 December 2019; by 19 February 2020—the end of our study period—74,576 infected cases had been verified in mainland China according to data from the China Center for Disease Control

[1]Faculty of Business and Economics, The University of Hong Kong, Hong Kong, 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, USA.
✉e-mail: jmjia@cuhk.edu.cn

and Prevention[19–21]. Our time period includes the time at which the news about the outbreak initially appeared (on 31 December 2019 and 9 January 2020) and the annual Lunar New Year migration (which culminated on 24 January 2020). The dataset included any mobile phone user who had spent at least 2 h 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 customer count of the carrier and their extrapolated count of total population movement. We use the latter in our primary analyses and the former as a robustness check (Supplementary Information).

We defined population flow as the total aggregate count of people who entered any given prefecture from Wuhan during the whole observation period (1–24 January 2020). Because Wuhan (population of 11.08 million people in 2018) is a major transportation hub, many of these people were travellers passing through rather than residents. The definition is also weighted by the number of transits through Wuhan since some people may have entered and exited Wuhan on several occasions in January (especially if they lived in neighbouring 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 24 January—Lunar New Year's Eve (outbound holiday travel is typically completed before this evening)—and January 23, when Wuhan was quarantined. We analysed the efficacy of the quarantine (Fig. 1b, c), which was manifested in a reduction of 52% and 38% in inter- and intra-provincial population outflow, respectively, on 23 January 2020 compared with 22 January 2020 (when there were 546,324 and 141,208 counts of intra- and extra-provincial travel, respectively), and a further reduction of 94% and 84% on 24 January 2020 compared with 23 January 2020. With the imposition of the quarantine—first in Wuhan (and two neighbouring prefectures) at 10:00 on 23 January 2020, and then in 12 other prefectures in Hubei by the end of the day on 24 January 2020—population outflow from Wuhan almost completely stopped (the average daily outflow thereafter was just 1,087 people to all prefectures outside of Hubei, which probably comprised government workers).

We combined the population flow dataset with the count and geographical location of confirmed cases of COVID-19 nationwide (Fig. 1a), which used consistent and stringently enforced case ascertainment during this period. As of 19 February 2020, there were 74,576 infected cases in mainland China, of which 29,549 occurred outside of Wuhan and there were 2,118 fatalities (according to data from the China Center for Disease Control and Prevention).

Population flow from Wuhan was hypothesized to export the virus to other locations, where it caused local outbreaks (that is, either by importation or community transmission (refs. [19–21])). Indeed, we find a strong correlation between total population flow and the 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 1 to 24 January 2020, and the correlation increases over time from $r = 0.522$ on 24 January 2020 to $r = 0.919$ on 5 February 2020, and increases further to $r = 0.952$ on 19 February 2020 ($P < 0.001$ for all) (Fig. 2a–c). As 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 different time windows of population outflow were used (Extended Data Fig. 1). The correlation between population outflow from Hubei province (excluding Wuhan itself) and the number of infections in each prefecture (Fig. 2c) followed a similar pattern but was substantially weaker; this correlation increased from $r = 0.365$ on 24 January 2020 to $r = 0.583$ on 19 February 2020.

For completeness we compared the predictive strength of aggregate population outflow to other factors—such as the relative frequency of Baidu search engine queries for virus-related terms in each prefecture (for example, novel coronavirus, flu, SARS, atypical pneumonia and surgical mask)[22–24], the gross domestic product (GDP) and population size of each prefecture, and other movement variables. Each of these factors became less predictive of local outbreak size over time, either for the number of cumulative cases or the number of daily reported cases (Fig. 2c, d and Extended Data Figs. 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 regions and their populations. Here, we use a special case of the gravity model with only the population variable for the 'recipient' prefecture as 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). However, 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[8,25].

Furthermore, we tested a gravity model to predict the infection count. Although the population size of the recipient prefecture and distance were significant predictors ($P < 0.001$), a mediation analysis shows that population flow from Wuhan mediates the effect of distance. Figure 2c, d illustrates 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. By contrast, the predictive strength of the distance from Wuhan, population size and GDP (an alternative source of gravity) of each prefecture shows no increases or decreases over time. There is no advantage to using distance to estimate population flow and infection spread when the actual population flow is observable, as in our case.

Next, we used two sets of models—one cross-sectional and one dynamic model—to statistically model and benchmark the extent to which aggregate population outflow from Wuhan predicts the spread and distribution of infections with SARS-CoV-2 across mainland China. We developed what we call a risk source model that leverages observed population flow data to operationalize the risk emanating from the epidemic source.

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

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

in which y_i is the number of the cumulative (or daily) confirmed cases in prefecture i (depending on the model); x_{1i} is the cumulative population outflow from Wuhan to prefecture i from 1 to 24 January 2020; 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. 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 $I_{ik} = 0$ (Supplementary Information).

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

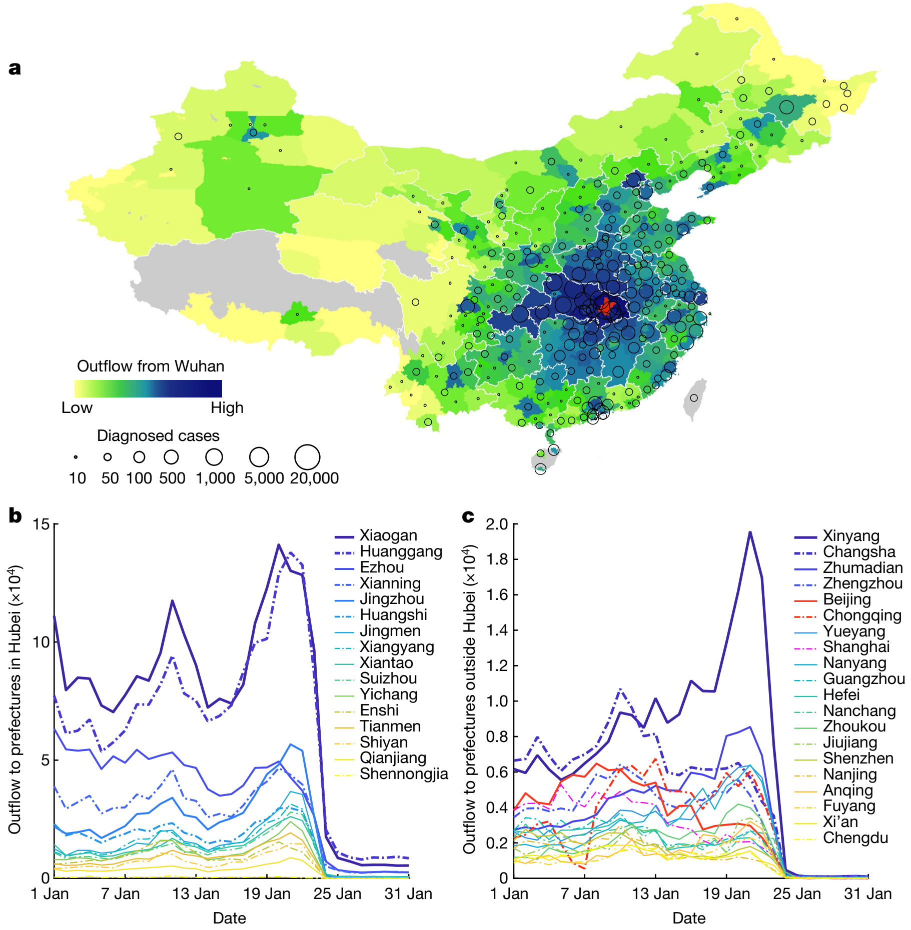


Fig. 1 | Geographical distribution of population outflow and confirmed COVID-19 cases as of 19 February 2020. a, There is a high overlap between the geographical distribution of aggregate population outflow from Wuhan until 24 January 2020 (in red) and the number of confirmed cases of COVID-19 in other 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 more than 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.

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 predictive strength of the model, is that population flow from Wuhan to other prefectures fundamentally determines the eventual distribution of total infections in 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 the number of predicted and confirmed cases on any given date (Supplementary Information). A higher-than-expected level of infection suggests more community transmission (that is, 'under-performing' 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, as they could have implemented highly successful public health measures (or may be prone to inaccurate data reporting). For example, Extended Data Fig. 4 identifies prefectures with transmission risk index values above the upper bound of the 90% confidence interval on 29 January, and the crossing of this threshold 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 decreases 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 (for example,

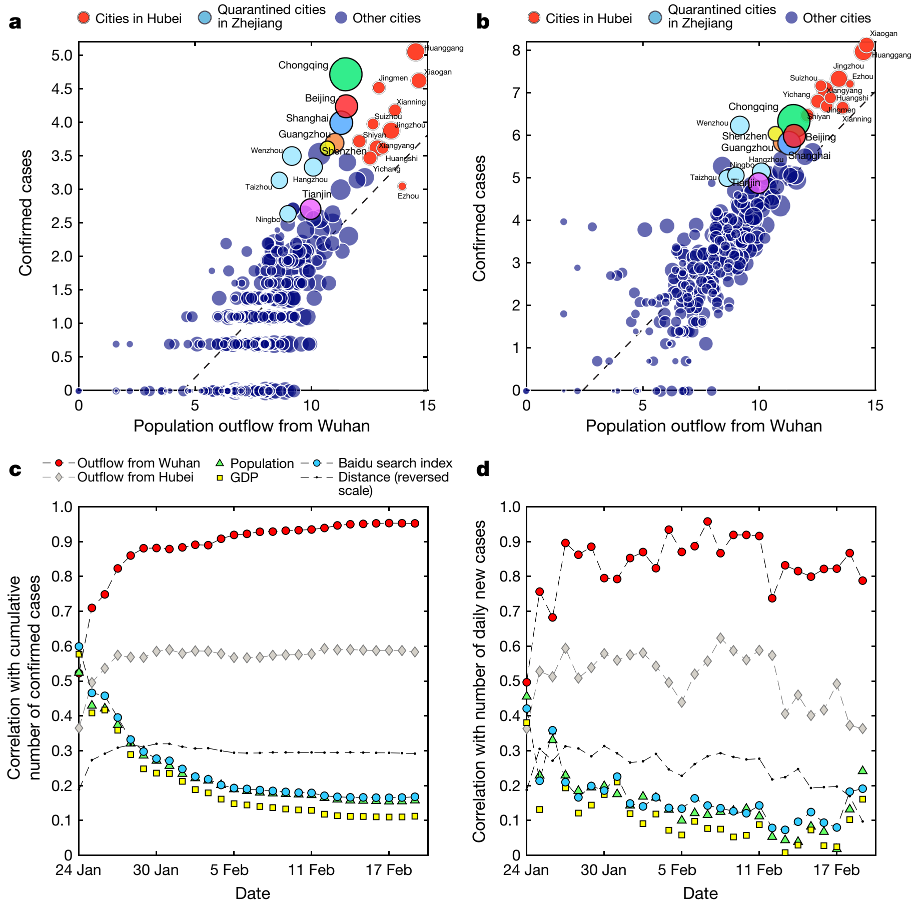


Fig. 2 | Factors correlated with confirmed COVID-19 cases. a, **b**, The relationship between the log-transformed aggregate population outflow from Wuhan (up to 24 January 2020) and the log-transformed number of confirmed cases by prefecture on 26 January 2020 (**a**) and 19 February 2020 (**b**). Red 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 colours. **c**, Relationship over time between the number of confirmed cases (cumulative until 19 February 2020) and the cumulative population inflow (up to 24 January 2020) from Wuhan, the cumulative inflow from Hubei province excluding Wuhan, the frequency of Baidu search terms related to the virus, the GDP, population and distance from Wuhan of the prefectures. Over time, the correlation between population outflow from Wuhan and the number of infected cases increases from Pearson's $r = 0.522$ on 24 January 2020 to $r = 0.952$ on 19 February ($n = 296$ prefectures). The decrease in the predictive strength of online search behaviour might reflect information saturation, while the decrease in the 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. **d**, The correlation with daily infections is consistent throughout the period with Pearson's r ranging from 0.496 on 24 January 2020 to a peak of 0.926 on 4 February 2020 ($n = 296$ prefectures). Fluctuations probably indicate lags in the reporting of cases (that are smoothed in **c**); weaker correlations on the last few days reflect that more than 90% of prefectures outside of Hubei reported no new cases.

logistic, generalized logistic or Gompertz functions[26,27]) that epidemics typically follow:

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

in which $\lambda(t|x_i)$ 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 $x_i = \{x_{1i}, x_{2i}, \ldots, x_{mi}\}$ are the realized values of the covariates for prefecture i; the other notation is the same as model (1).

This model extends our risk source model to a dynamic context; it incorporates all infected cases across all locales and dates to statistically derive the COVID-19 epidemic curve and growth pattern across mainland China. We used the same method as before to estimate the parameters (Supplementary Information). When using only the single variable of total population outflow from Wuhan (from 1 to 24 January 2020) to each other prefecture, we observe $R^2 = 0.927$ for the exponential–logistic model (Fig. 3a); 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 above to contrast the expected and observed outcomes to gauge epidemiological risk. Here, model predictions serve as reference patterns across time (Extended Data Figs. 5, 6). The differences in the growth trends between the number of 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 Data Fig. 7 and Supplementary Table 11). Indeed, our model identifies a list of statistically significant underperforming prefectures; in most of these cases, we observed the subsequent imposition of quarantine (Extended Data Figs. 5, 6, Supplementary Information and Supplementary Table 12). On the other hand, prefectures with lower trends than expected might have had more successful public health measures. Figure 3b shows the dynamic shifts in the risk index score for selected prefectures,

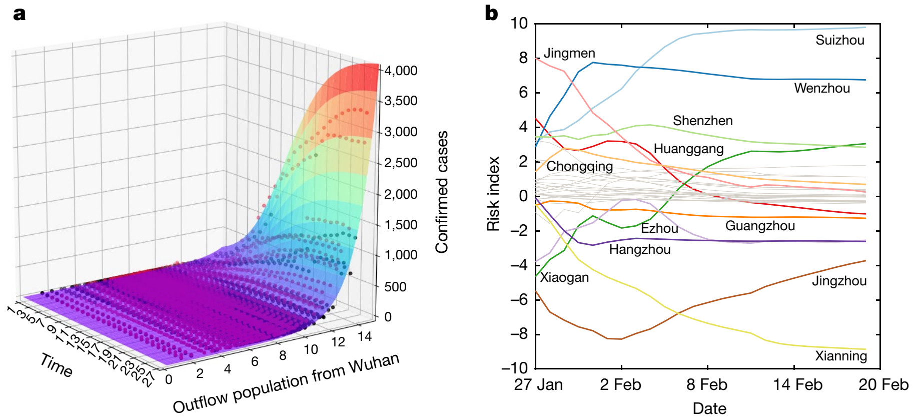


Fig. 3 | Predictive model based on population outflow. a, The surface indicates the fitted performance of our epidemiological model (see model (5) in the Supplementary Information) with a single variable x_{1i}, which indicates the outflow population from Wuhan to prefecture *i* (log transformed) for all prefectures, with *t* as the number of days after outbound Chunyun is over (that is, $t = 1$ is 24 January 2020). The dots represent the actual number of confirmed cases for a given x_{1i} and *t*. Red dots represent prefectures in which the reported number of confirmed cases is greater than the values predicted by the model; black dots are all other cases, $R^2 = 0.930$ ($n = 7{,}992$ data points). See Extended Data Fig. 8 for a robustness check. **b**, Risk scores over time provide a dynamic picture of shifting transmission risks in different prefectures.

which enables the monitoring of prefectures to analyse which prefectures performed better in controlling the transmission risk over time.

In summary, 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 up to 24 January 2020. The geographical flow of people anticipates the subsequent location, intensity and timing of outbreaks in the rest of mainland China up to 19 February 2020. These data outperform other measures, such as population size, wealth or distance from the risk source. We modelled 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 nonlinear estimations; generates a non-arbitrary, source-linked risk score; and is easily adapted to other empirical contexts. Notably, the multiplicative functional form can also accommodate multiple risk sources—for example, for countries in which there are multiple disease epicentres. As an example, we evaluated the distinct impact of population flow from Hubei (excluding Wuhan) as an alternative risk source in our models, and found that it had little impact on the spread and growth of COVID-19 in the country (Supplementary Tables 6, 10).

We focused on the relative strength of the outbreak in each area, rather than the absolute number of cases, although 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 geographical 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 (for example, 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, for the allocation of resources and manpower across specific geographical 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[22–24] or the use of network sensors (that is, the monitoring of people who are at heightened risk of falling ill given their network position)[28]. Our approach relies on data regarding population flow. Indeed, historical (that is, 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.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2284-y.

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. et al. Ebola: mobility data. *Science* **346**, 433 (2014).
3. Brockmann, D. & Helbing, D. The hidden geometry of complex, network-driven contagion phenomena. *Science* **342**, 1337–1342 (2013).
4. Balcan, D. et al. 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., González, 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. Y., Wang, W. X., Gao, Z. Y. & Lai, Y. C. Universal model of individual and population mobility on diverse spatial scales. *Nat. Commun*. **8**, 1639 (2017).
10. Csáji, B. C. et al. Exploring the mobility of mobile phone users. *Physica A* **392**, 1459–1473 (2013).
11. Wesolowski, A. et al. Quantifying the impact of human mobility on malaria. *Science* **338**, 267–270 (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. et al. 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. et al. Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. *Nat. Med*. **26**, 506–510 (2020).
16. Chinazzi, M. et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* **368**, 395–400 (2020).
17. Du, Z. et al. Risk for transportation of coronavirus disease from Wuhan to other cities in China. *Emerg. Infect Dis*. **26**, 1049–1052 (2020).
18. Li, R. et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2). *Science* **368**, 489–493 (2020).
19. Wu, F. et al. A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).
20. Zhu, N. et al. A novel coronavirus from patients with pneumonia in China, 2019. *New Engl. J. Med*. **382**, 727–733 (2020).
21. Chan, J. F.-W. et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet* **395**, 514–523 (2020).
22. Ginsberg, J. et al. Detecting influenza epidemics using search engine query data. *Nature* **457**, 1012–1014 (2009).
23. Lazer, D., Kennedy, R., King, G. & Vespignani, A. The parable of Google flu: traps in big data analysis. *Science* **343**, 1203–1205 (2014).
24. Viboud, C. & Vespignani, A. The future of influenza forecasts. *Proc. Natl Acad. Sci. USA* **116**, 2802–2804 (2019).
25. Massey, D. S. & España, F. G. The social process of international migration. *Science* **237**, 733–738 (1987).
26. Bürger, R., Chowell, G. & Lara-Díiaz, L. Y. Comparative analysis of phenomenological growth models applied to epidemic outbreaks. *Math. Biosci. Eng*. **16**, 4250–4273 (2019).
27. Roosa, K. et al. Short-term forecasts of the COVID-19 epidemic in Guangdong and Zhejiang, China: February 13–23, 2020. *J. Clin. Med*. **9**, 596 (2020).
28. Christakis, N. A. & Fowler, J. H. Social network sensors for early detection of contagious outbreaks. *PLoS ONE* **5**, e12948 (2010).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Data necessary to reproduce the primary results of this study are included in the Article and its Supplementary Information.

Code availability

Code necessary to reproduce the primary results of this study is included in the Article and its Supplementary Information.

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 and 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 contributed equally to the paper. J.S.J., J.J. and N.A.C. conceived the research. J.J., Y.Y., X.L., J.S.J. and G.X. analysed the data. J.S.J. and N.A.C. wrote the paper. J.J., J.S.J. and X.L. obtained funding. All authors contributed to research design, analytical development and critical revisions.

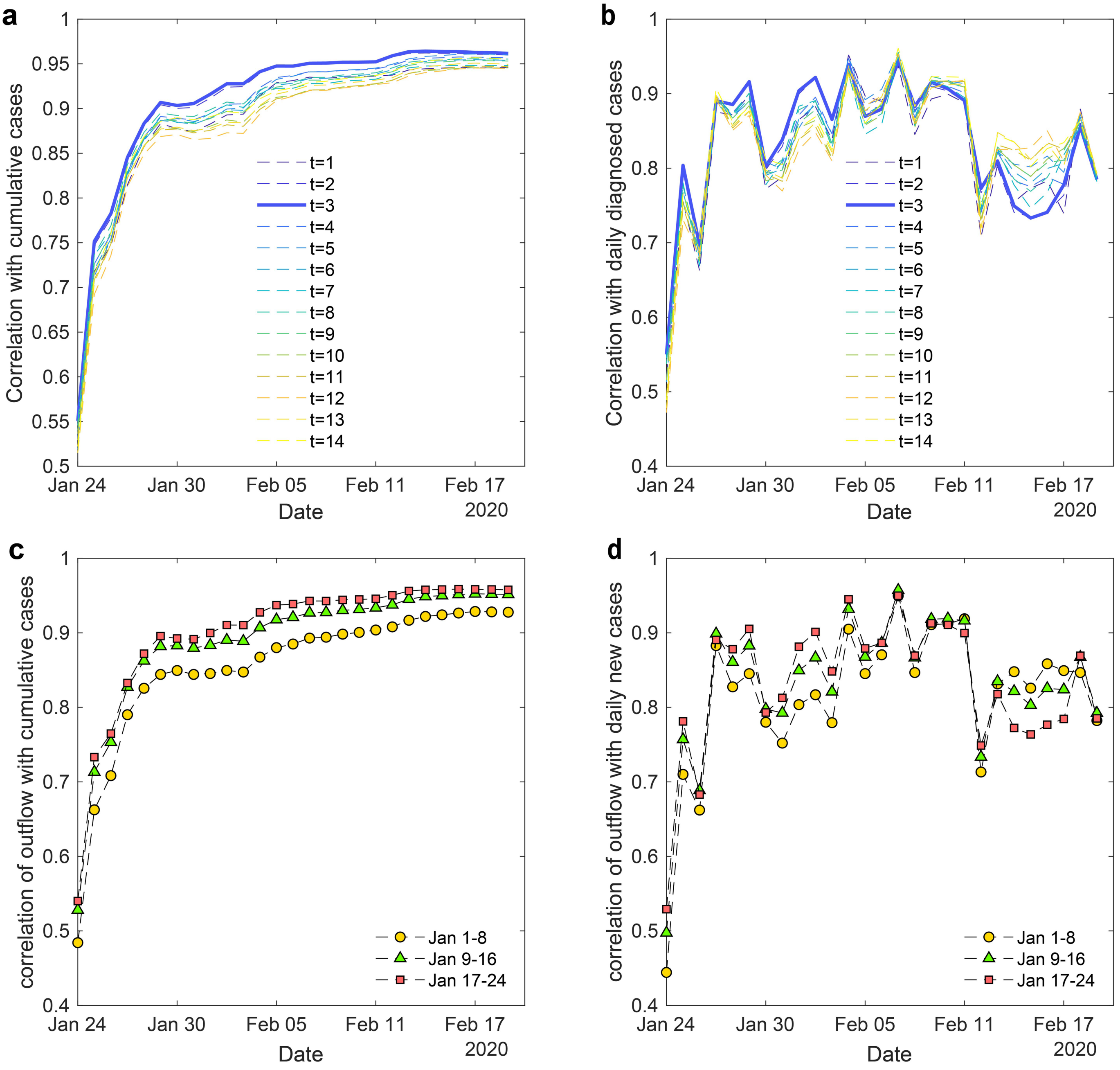
Competing interests The authors declare no competing interests.

Additional information

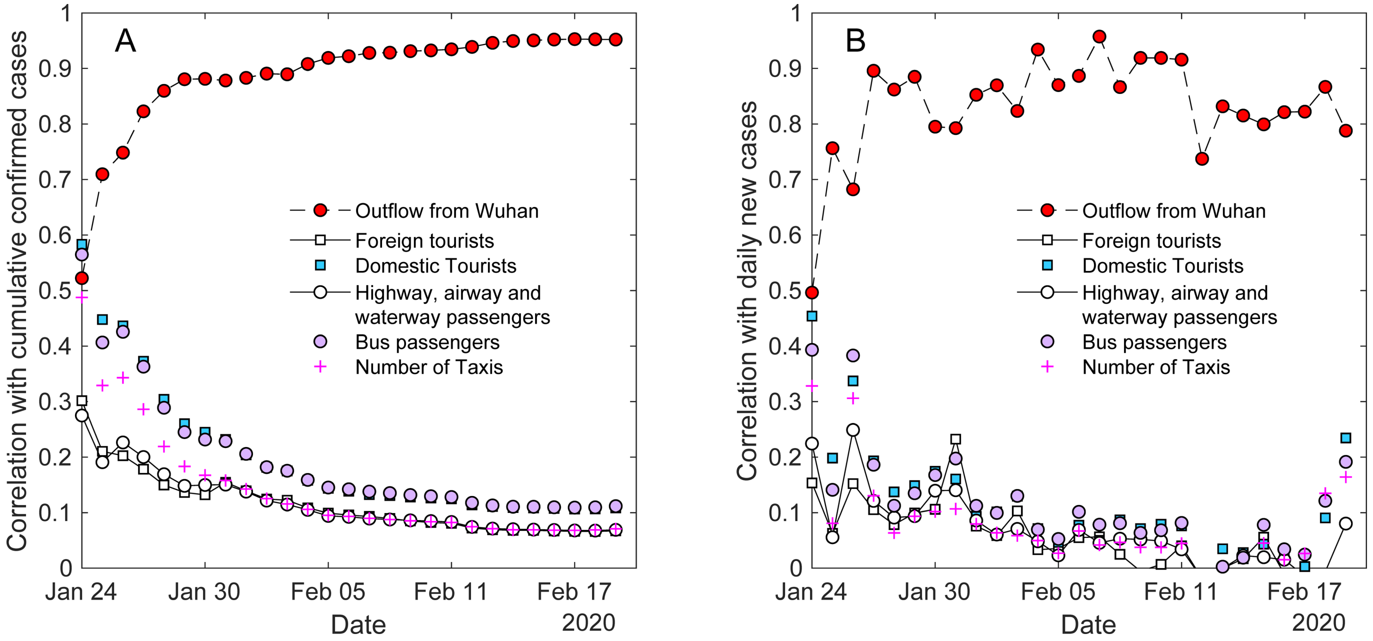
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2284-y.

Correspondence and requests for materials should be addressed to J.J.

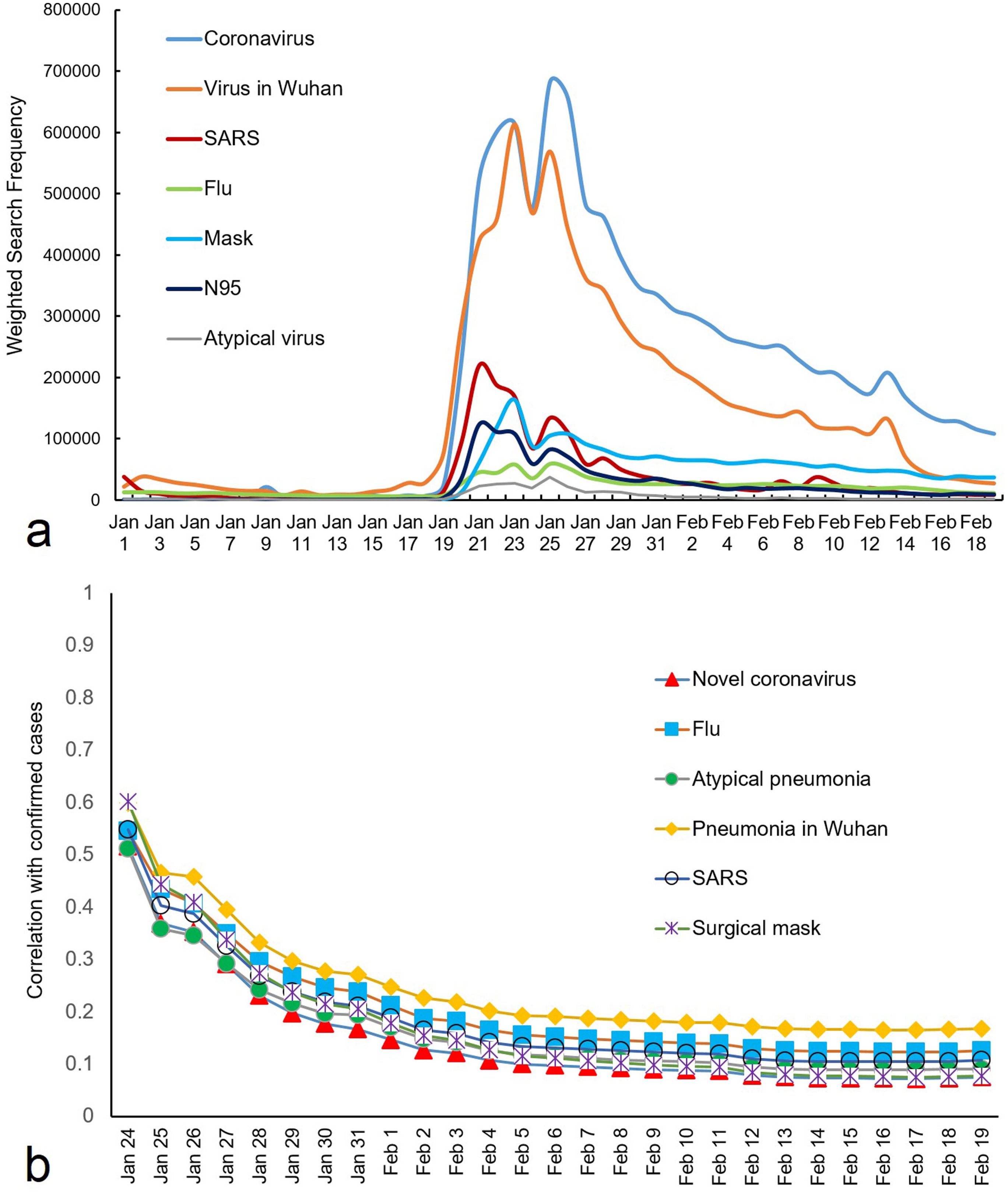
Peer review information *Nature* thanks Jukka-Pekka Onnela and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



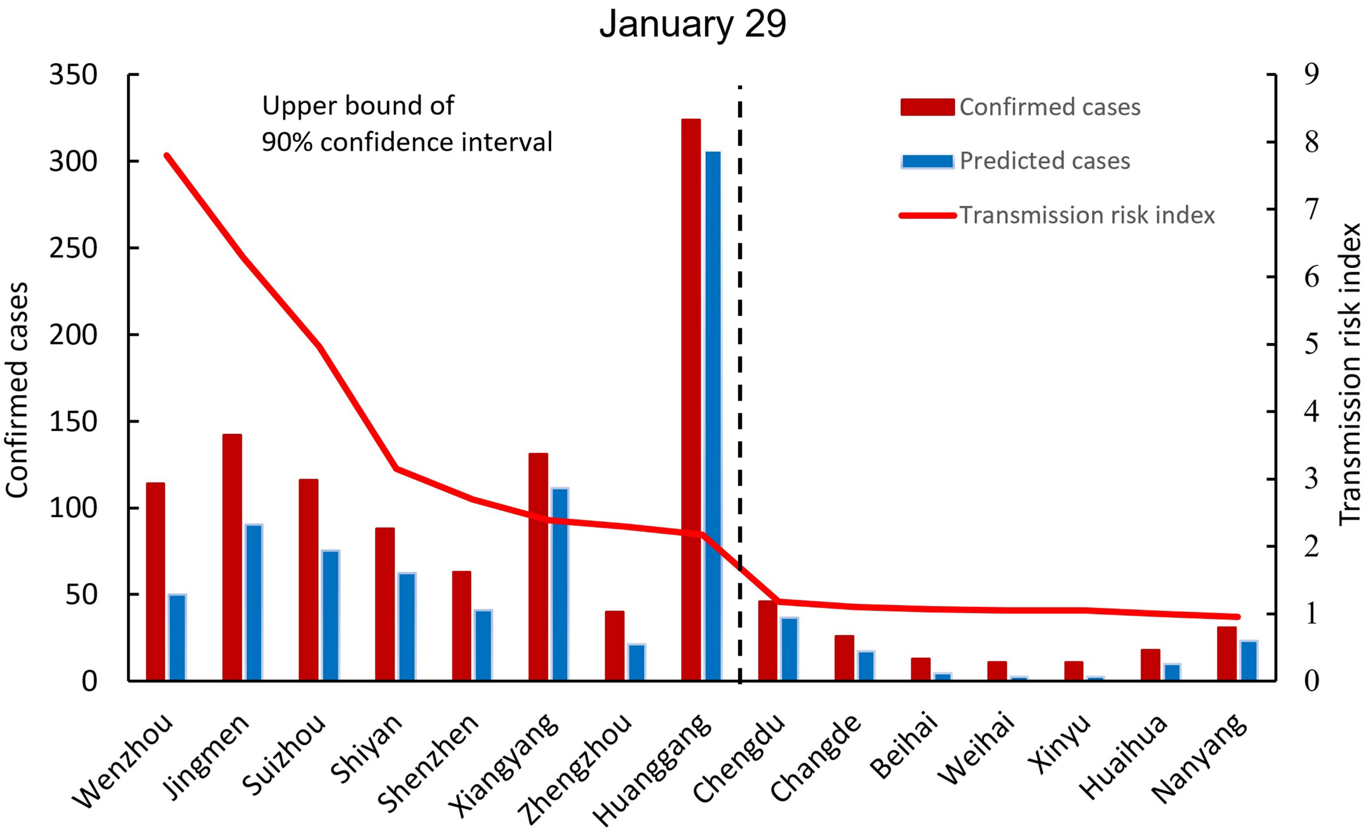
Extended Data Fig. 1 | Time-window sensitivity test for the correlational analysis. a, **b**, Pearson's correlation (*n* = 296 prefectures) between the cumulative number of confirmed cases and population outflow from Wuhan on different days ranging from 1 to 14 days before 24 January 2020 for the cumulative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



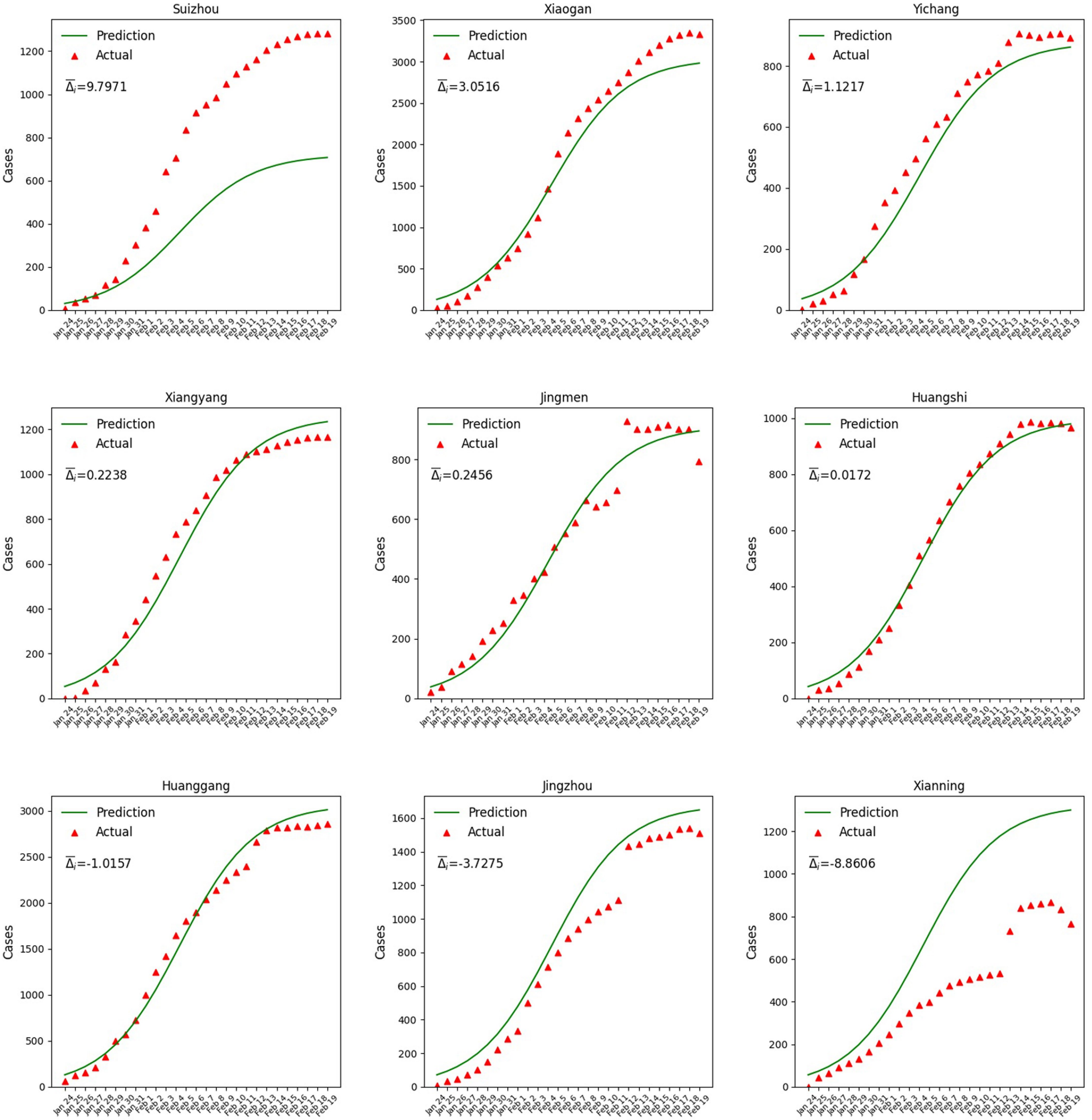
Extended Data Fig. 2 | Correlation with alternative population movement measures. a, **b**, Pearson's correlation (*n* = 296 prefectures) between alternative publicly available movement measurements from the 2018 City/Prefectures Statistical Year Book of China (with aggregate population outflow data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



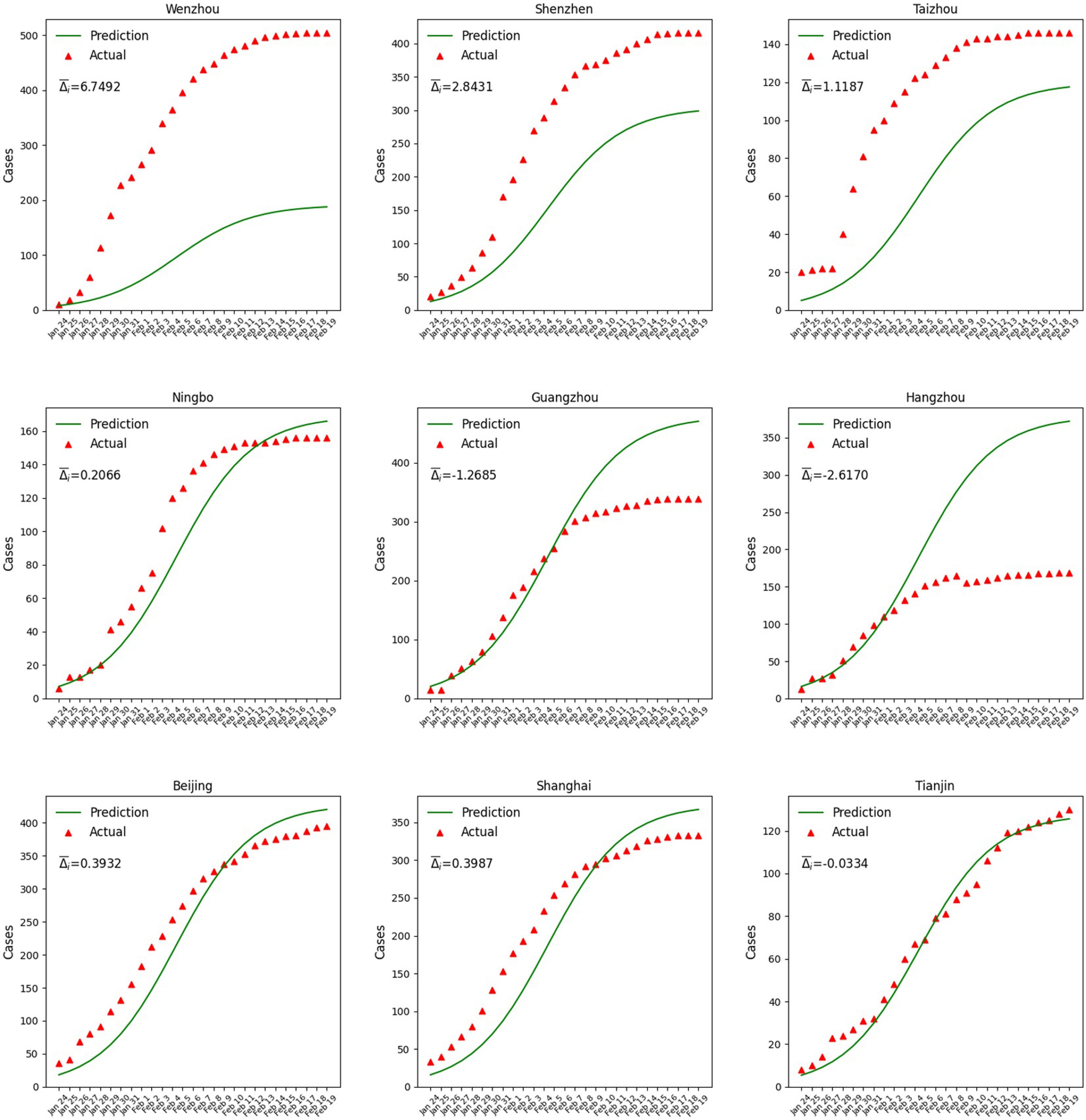
Extended Data Fig. 3 | Search terms and correlation with confirmed cases. **a**, Search frequency of Baidu search terms related to the COVID-19 outbreak: the search terms are direct translations of the Chinese keywords that Baidu users used during the study period (note the official WHO name 'COVID-19' was only announced on 11 February 2020). **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.



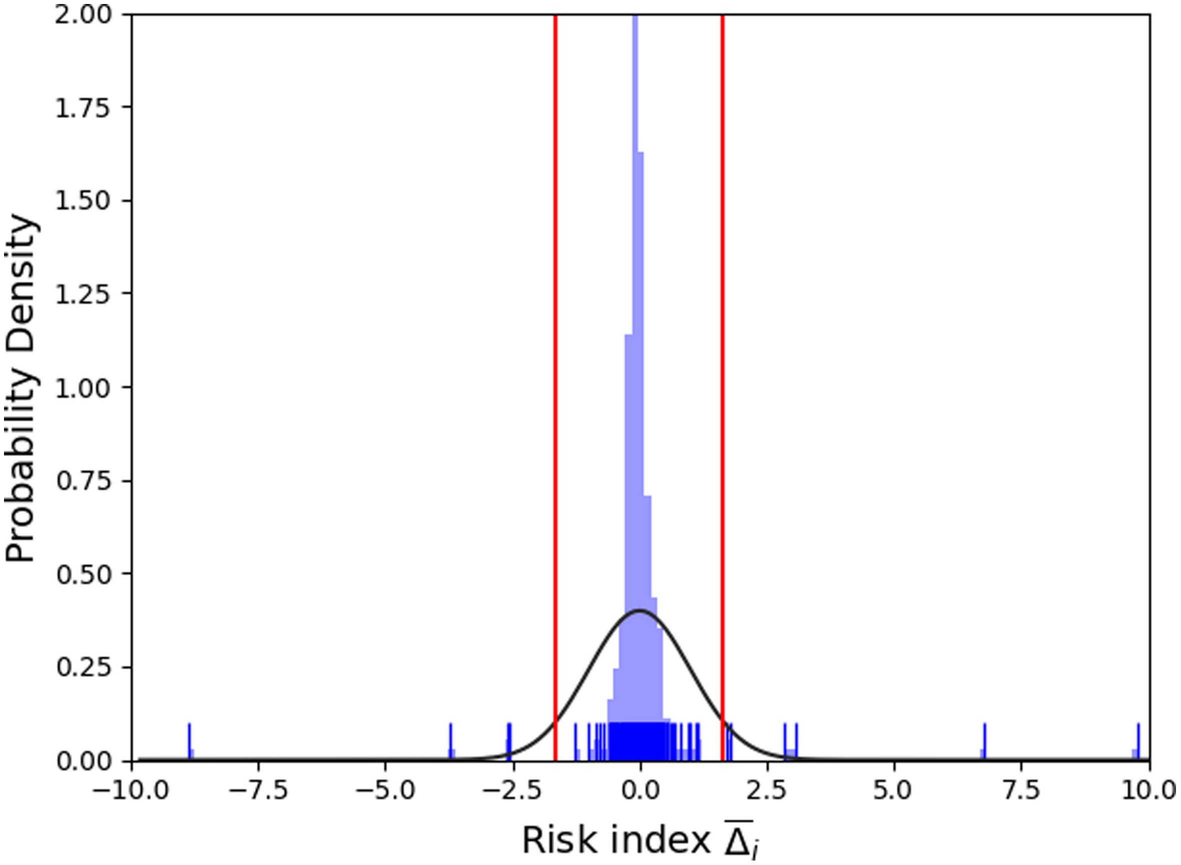
Extended Data Fig. 4 | Prefectures with a high transmission risk index on 29 January 2020. The predicted structure of the spread of the SARS-CoV-2 virus can be used as a benchmark to identify which locales deviate significantly. As model (1) predicts the number of cases in a prefecture based on the population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



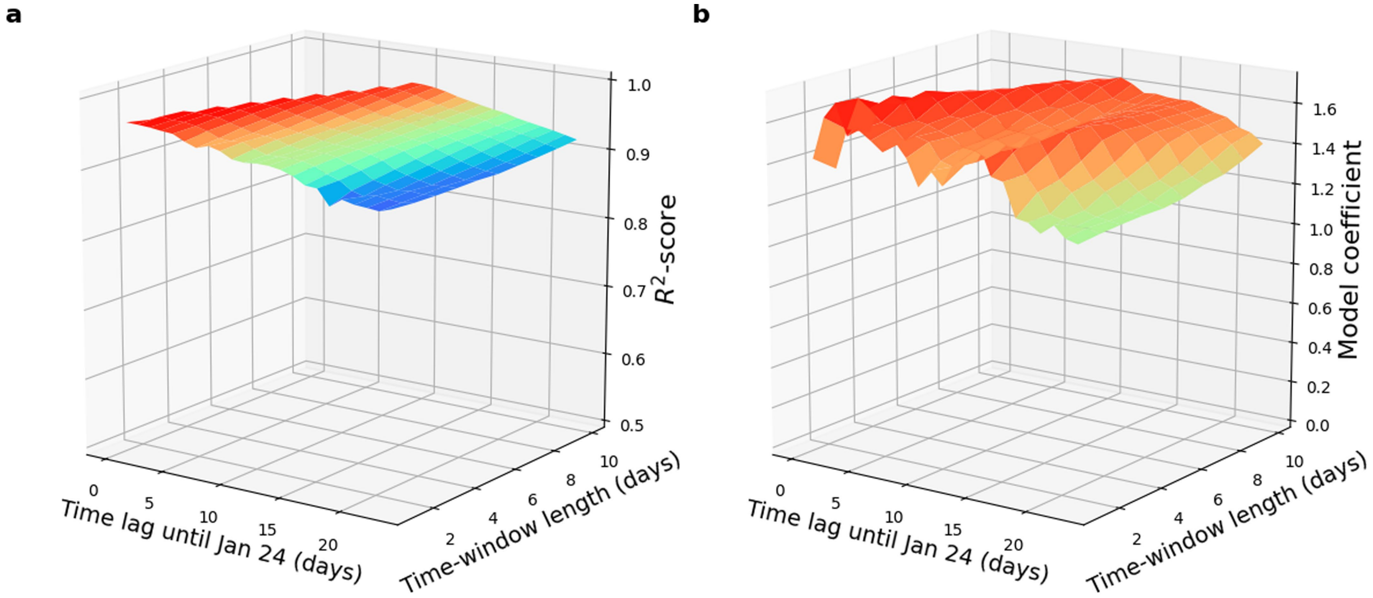
Extended Data Fig. 5 | Benchmark (predicted) versus actual virus growth in the prefectures of Hubei province. Model (2) used aggregate population outflow from Wuhan from 1 to 24 January 2020 to provide a reference growth pattern (that is, epidemic curves) for the spread of COVID-19 across time and space, without making a priori assumptions about the 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 13 February 2020 reflected a change in the infection count criteria of local governments; clinically diagnosed cases came to be included in total confirmed case counts in those prefectures (within Hubei province).



Extended Data Fig. 6 | Benchmark (predicted) versus actual virus growth in selected prefectures outside of Hubei province. Model (2) used aggregate population outflow from Wuhan from 1 to 24 January 2020 to provide a reference growth pattern (that is, epidemic curves) for the spread of COVID-19 across time and space, without making a priori assumptions about the 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 Fig. 7 | The distribution of the transmission risk index. The transmission risk index ($\overline{\Delta}_i$) is the normalized score of the integral of the differences between the actual number of confirmed infected cases and predicted numbers in our model. Prefectures above the 90% confidence interval of the index are likely to experience 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 (Supplementary Table 11).



Extended Data Fig. 8 | Robustness check of model (2) with different time lags and time-window lengths. We explored 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 24 January 2020 the time window starts. For example, analyses using time lag = 1 and time window = 2 use outflow data between 23 and 24 January 2020. The surfaces show that a more recent time lag improves the R^2 (**a**) as well as the parameter value (**b**) of the population outflow coefficient in model (2).

nature research

Corresponding author(s): Jianmin Jia

Last updated by author(s): April 16, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	□	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Aggregated mobility data extracted from mobile phones are provided by one of the largest operators in China (regarding the total number of mobile phone users in China). The data on population flows and other key covariates used in the primary analyses will be made available upon publication. The daily infection data is public information that the government releases in China.
Data analysis	We used the Levenberg–Marquardt (LM) algorithm for model estimation, and the code was from Newville et al. (2016); Software includes Matlab (R2018a, The MathWorks, Inc.), Python (V.3.7.4, Python Software Foundation), and ArcGIS (V.10.2, Esri). We will make the code available upon publication (and for review).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data and code necessary to reproduce the primary results of this study are included in this published article for release online by Nature (and in the supplementary information files).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences ☒ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description

We use data regarding outflow population from Wuahn to different prefectures in China (ascertained with mobile phone records) to explore the spread of Coronavirus, ascertained by the Chinese CDC, and to assess transmission risk in difference areas.

Research sample

We used aggregate population outflow data of all people transiting through Wuhan, China between Jan 1-24, 2020; data was provided by a major national carrier. Types of data described in SI.

Sampling strategy

We used all available population outflow data in analyses (and conducted robustness checks using all different variants/alternative measures of the population outflow data provided). N=296 prefectures based on available covariate data (for GDP and population) in statistical yearbook published by National Bureau of Statistics of China, which covered 94% of the population. Any prefectures not covered was due to lack of data availability from this official government source.

Data collection

We obtained the aggregated mobile data via our industry partner in China and linked these records, at the level of 289 Chinese prefectures, to publicly available coronavirus cases in these areas.

Timing

The mobility data was collected during the period January 1 to January 24, 2020; and the confirmed case data was collected starting from January 24 up to February 19, 2020.

Data exclusions

All data that can be matched with the China Prefectures (City) Statistical Year Book have been included in the analysis (to provide covariates for our model); smaller sparsely populated prefectures not covered by the official Statistic Bureau's yearbook were excluded.

Non-participation

NA We used aggregated data of all customers of the carrier that traveled through or were in Wuhan during the study period.

Randomization

NA This study was not an experiment, and it did not have experimental conditions.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	□	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology
☒	□	Animals and other organisms
□	☒	Human research participants
☒	□	Clinical data

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics

We used the aggregated mobile data of the Chinese phone users transiting through Wuhan in January 2020.

Recruitment

NA Population flow data was provided in aggregate form by a major Chinese carrier.

Ethics oversight

This work has been supported by the National Natural Science Foundation of China for the urgent policy research (given the pandemic). We do not use individual-level data, only anonymized aggregate flows, and this work is exempt from IRB review in China. An email to this effect was also obtained from the Yale IRB, and it has been shared with Clare Thomas, Senior Editor, Nature.

Note that full information on the approval of the study protocol must also be provided in the manuscript.