

Crowding and the shape of COVID-19 epidemics

Benjamin Rader^{1,2,16}, Samuel V. Scarpino^{1,3,4,5,16}✉, Anjalika Nande^{1,6}, Alison L. Hill^{1,6,7}, Ben Adlam⁶, Robert C. Reiner^{1,8,9}, David M. Pigott^{1,8,9}, Bernardo Gutierrez^{1,10,11}, Alexander E. Zarebski¹⁰, Munik Shrestha³, John S. Brownstein^{1,12}, Marcia C. Castro^{1,13}, Christopher Dye^{1,10}, Huaiyu Tian¹⁴, Oliver G. Pybus^{1,10,15}✉ and Moritz U. G. Kraemer^{1,10}✉

The coronavirus disease 2019 (COVID-19) pandemic is straining public health systems worldwide, and major non-pharmaceutical interventions have been implemented to slow its spread^{1–4}. During the initial phase of the outbreak, dissemination of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was primarily determined by human mobility from Wuhan, China^{5,6}. Yet empirical evidence on the effect of key geographic factors on local epidemic transmission is lacking⁷. In this study, we analyzed highly resolved spatial variables in cities, together with case count data, to investigate the role of climate, urbanization and variation in interventions. We show that the degree to which cases of COVID-19 are compressed into a short period of time (peakedness of the epidemic) is strongly shaped by population aggregation and heterogeneity, such that epidemics in crowded cities are more spread over time, and crowded cities have larger total attack rates than less populated cities. Observed differences in the peakedness of epidemics are consistent with a meta-population model of COVID-19 that explicitly accounts for spatial hierarchies. We paired our estimates with globally comprehensive data on human mobility and predict that crowded cities worldwide could experience more prolonged epidemics.

Predicting the epidemiology of the COVID-19 pandemic is a priority for guiding epidemic responses around the world. China has undergone its first epidemic wave, and, remarkably, cities across the country are now reporting few or no locally acquired cases⁸. Analyses have indicated that the spread of COVID-19 from Hubei to the rest of China was driven primarily by human mobility from Wuhan^{6,9}, and that the stringent measures to restrict human movement and public gatherings within and among cities in China were associated with bringing local epidemics under control⁵. Key uncertainties remain as to which geographic factors drive the local transmission dynamics of COVID-19, and initial analysis suggests a limited role of climate in determining epidemic growth¹⁰.

Spatial heterogeneity in infectious disease transmission can be influenced by local differences in population or human movements, such that high local population densities might catalyze the spread of new pathogens due to higher contact rates with susceptible

individuals^{11,12}. For respiratory pathogens, the temporal clustering of cases in an epidemic (that is, the shortest period during which most cases are observed) varies with increased indoor crowding and socio-economic and climatic factors^{13–18}. The temporal concentration of cases is minimized when incidence is spread evenly across time and increases as incidence becomes more concentrated in particular days, as has been observed for influenza¹³. In any given location, a higher temporal concentration of cases might require a larger surge capacity in the public health system¹⁹, especially for an emerging respiratory pathogen such as COVID-19 (ref. ²⁰).

Results

Spatial population structure predicts the shape of epidemics of COVID-19. China and Italy provide detailed epidemiological time series for COVID-19 (refs. ^{2,21,22}) across a wide range of geographic contexts; hence, the outbreaks in these countries provide an opportunity to evaluate the role of local factors in shaping epidemic behavior. We used daily epidemiological data from Chinese cities^{23,24} and Italian provinces, climate and population data and the response to local interventions as measured by human mobility data from Baidu²⁵ and the COVID-19 Aggregated Mobility Research Dataset (<https://www.google.com/covid19/mobility/>) to identify drivers of transmission, with a focus on how the temporal clustering of cases differs between prefectures in China and provinces in Italy. A summary of the main findings, limitations and policy implications of our study is shown in Table 1.

We used daily incidence data of confirmed COVID-19 cases aggregated at the prefectural level ($n=293$) in China (Fig. 1a) and at the province level ($n=108$) in Italy. Prefectures and provinces are administrative units that typically have one urban center (Fig. 1b). We aggregated daily individual-level data collected from official government reports²². Epidemiological data in each prefecture were truncated to exclude dates before the first and after the last day of reported cases during the first epidemic. Cases reported after March 1, 2020, that were imported from outside China were excluded from the analysis. All epidemiological data from Hubei Province were excluded because of the lack of prefecture-level epidemiological data and issues with consistent reporting before January 20, 2020. The shape of epidemic curves varied among prefectures, with some

¹Computational Epidemiology Lab, Boston Children's Hospital, Boston MA, USA. ²Department of Epidemiology, Boston University School of Public Health, Boston MA, USA. ³Network Science Institute, Northeastern University, Boston MA, USA. ⁴ISI Foundation, Turin, Italy. ⁵Santa Fe Institute, Santa Fe NM, USA. ⁶Program for Evolutionary Dynamics, Harvard University, Cambridge MA, USA. ⁷Institute for Computational Medicine, Johns Hopkins University, Baltimore MD, USA. ⁸Department of Health Metrics, University of Washington, Seattle WA, USA. ⁹Institute for Health Metrics and Evaluation, University of Washington, Seattle WA, USA. ¹⁰Department of Zoology, University of Oxford, Oxford, UK. ¹¹School of Biological and Environmental Sciences, Universidad San Francisco de Quito USFQ, Quito, Ecuador. ¹²Harvard Medical School, Boston MA, USA. ¹³Department of Global Health and Population, Harvard T.H. Chan School of Public Health, Boston MA, USA. ¹⁴State Key Laboratory of Remote Sensing Science, College of Global Change and Earth System Science, Beijing Normal University, Beijing, China. ¹⁵Department of Pathobiology and Population Science, The Royal Veterinary College, London, UK. ¹⁶These authors contributed equally: Benjamin Rader, Samuel V. Scarpino. ✉e-mail: s.scarpino@northeastern.edu; oliver.pybus@zoo.ox.ac.uk; moritz.kraemer@zoo.ox.ac.uk

Table 1 | Policy summary

Background	There are obvious differences in the geographic distribution of COVID-19 cases within and among countries. We hypothesize that some of these differences are due to spatial variability in population crowding. Using detailed case count data from COVID-19 among cities in China and Italy, we fit multiple regression models to explain variability in the shape of epidemics among them.
Main findings and limitations	We found that cities with higher crowding have longer epidemics and higher attack rates after the first epidemic wave. Using a meta-population model that splits cities into neighborhood subunits is consistent with these findings, suggesting that the hierarchical structure and organization of cities are influential in defining their epidemics. We predict that comparatively rural areas might experience more peaked epidemics. As with all modeling studies, further data generated during the epidemic might change our parameter estimates, and large-scale serological data would help verify our findings. Further, it will be important to evaluate whether cities that have greater peak incidence might be more prone to strained healthcare systems.
Policy implications	Our results have implications for assessing the drivers of transmission of SARS-CoV-2. Spatial factors, such as crowding and population density, might elevate the risk of sustained (longer) outbreaks, even after the implementation of lockdowns. Cities that are less crowded and have lower attack rates might be more susceptible to experiencing future outbreaks if SARS-CoV-2 is successfully re-introduced.

showing a rapid rise and decline in reported cases and others showing more prolonged epidemics (Fig. 1a and Extended Data Fig. 1).

To characterize the temporal clustering of cases for each prefecture and province, we calculated the Shannon diversity index of the distribution of incident cases¹³. We defined the incidence distribution p_{ij} for a given city to be the proportion of COVID-19 cases during the first epidemic wave j that occurred on day i . The Shannon index of incidence for a given prefecture and year is given by $v_j = \left(-\sum_i p_{ij} \log p_{ij} \right)^{-1}$. Because v_j is a function of the disease incidence curve in each location, rather than of absolute incidence values, it is less sensitive to varying reporting rates among cities. The Shannon index is maximal when all cases occur on the same day and minimal when each day of the epidemic has the same number of incident cases (for example, ‘flat’ epidemic curves). It is highly correlated with alternative measures of epidemic peakedness, such as the proportion of cases that occur at the peak ± 1 d (Extended Data Fig. 2). The total attack rate of reported COVID-19 cases in each prefecture is strongly negatively correlated with the Shannon index in China (Fig. 1c); hence, less peaked epidemics have a larger total attack rate (Pearson’s $r = -0.67$, 95% confidence interval (CI), -0.73 to -0.59 , $P < 0.01$; for Italy, $R^2 = 0.33$, $P < 0.01$). We hypothesize that this variation among cities in total attack rate and the temporal clustering of cases is the result of the spatial organization of human populations.

To test this hypothesis, we used Lloyd’s index of mean crowding^{13,26}, treating the population count of each spatial grid cell as an individual unit (Fig. 1). The term ‘mean crowding’ used here is a specific geographic metric that summarizes both population density and how density is distributed across a prefecture (that is, patchiness; Fig. 1). Higher values of Lloyd’s index suggest a spatially aggregated population structure. For example, Xi’an has high values of crowding, whereas Bozhou has a similar population density but a population that is more evenly distributed across the prefecture (Fig. 1b). We performed log-linear regression modeling to determine the association between the temporal clustering of cases with socio-economic and environmental variables, including reductions in population flows during the outbreak period (Methods).

We found that the temporal clustering of cases was significantly negatively correlated with the mean number of contacts ($P < 0.01$) but positively correlated with mean population density ($P < 0.01$) and varies widely across China and Italy (Fig. 2 and Supplementary Table 1). This observation contrasts with the expectations of simple and classical epidemiological models, which predict higher peakedness in crowded areas due to the increased availability of susceptible individuals^{27,28}. The spatial scale at which this relationship is best explained was 10×10 km, but results were statistically significant

at all spatial scales between 1 and 50 km^2 (Extended Data Fig. 3; $P < 0.01$). Mean specific humidity and population mobility remained significantly negatively correlated with epidemic peakedness when included in a multivariate model with crowding (Supplementary Table 1; all $P < 0.01$).

Using weekly human mobility data, we found that within-city human mobility during the outbreak was correlated with the temporal clustering of cases—that is, prefectures that have larger reductions in mobility also have lower epidemic peakedness (Extended Data Fig. 4 and Supplementary Table 1; $P < 0.01$). When we combined mobility reduction in a model with crowding and humidity, we found that these variables each remained significant predictors of the temporal clustering of cases (Extended Data Table 1; $P < 0.01$). These results suggest that, although measures to reduce mobility can successfully lead to a flattening of the epidemic curve, population crowding is an independent contributor to the shape of epidemics in these two countries.

Our multivariate model can explain a large fraction of the variation in epidemic peakedness among Chinese cities and Italian provinces, and sensitivity analyses confirm the robustness of our results to potential noise in location-specific incidence distributions ($R^2 = 0.638$; Extended Data Fig. 2, Supplementary Table 1 and Extended Data Fig. 5). To evaluate the out-of-sample performance of our model, we 1) performed n -fold cross validation at the prefecture level in China (Spearman’s $\rho = 0.61$; 95% bootstrap CI, 0.52–0.68; $P < 0.01$); 2) used the fitted model in China to estimate peak intensity at the corresponding administrative level 2 locations—that is, province level, in Italy (Spearman’s $\rho = 0.57$; 95% bootstrap CI, 0.41–0.69; $P < 0.01$); and 3) performed n -fold cross validation at the province level in Italy (Spearman’s $\rho = 0.65$; 95% bootstrap CI, 0.52–0.76; $P < 0.01$). These results suggest that the model is robust to both within- and between-country out-of-sample testing (Extended Data Fig. 6).

To evaluate the potential effect of the temporal clustering of cases on the peak attack rate and total attack rate, we performed a simple linear regression (Supplementary Table 2). For locations that have a single peak, the attack rate at the peak is highest in two settings: 1) in crowded locations with high population size (prefectures that also experience high total attack rates); and 2) in locations that have lower population and lower crowding and, therefore, high temporal clustering of cases (Extended Data Fig. 7). Other prefectures that have low population and low crowding sometimes experience very short outbreaks with a small peak attack rate, suggesting local stochastic extinction possibly due to limited mixing between populations. We hypothesize that the observation that high peak attack rates can sometimes be found in low crowding areas is related to rare super-spreading events as observed in Bergamo, Italy, or Mulhouse, France.

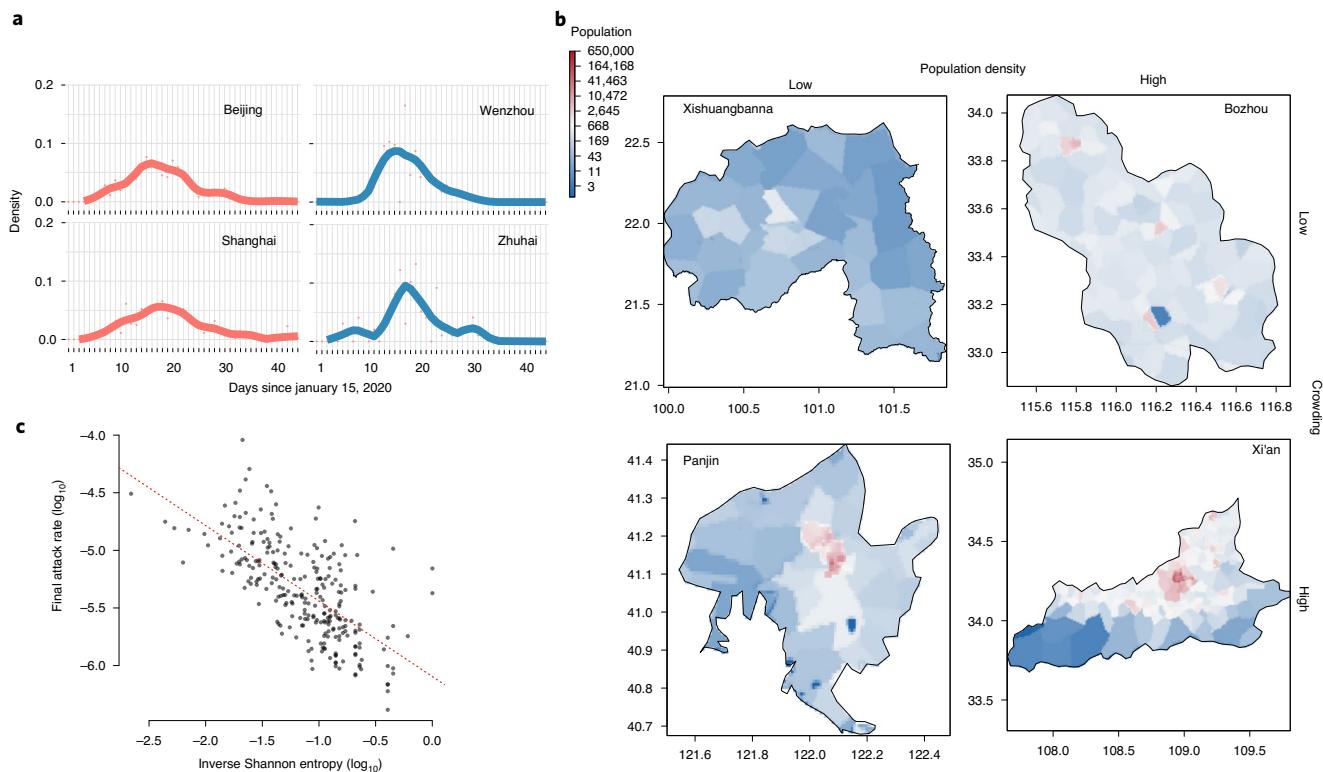


Fig. 1 | Maps of crowding in prefectures in China. **a**, Examples of epidemic curves that are normalized to show the percentage of cases across the whole epidemic that occur at each given day. Beijing and Shanghai (red) have less peaked epidemics than Wenzhou and Zhuhai. **b**, Examples of prefectures in China with different levels of crowding and population size. The color scale illustrates the estimated number of inhabitants per grid cell ($1\text{ km} \times 1\text{ km}$). **c**, Relationship between the Shannon index of the incidence curve and the final attack rate for prefectures in China.

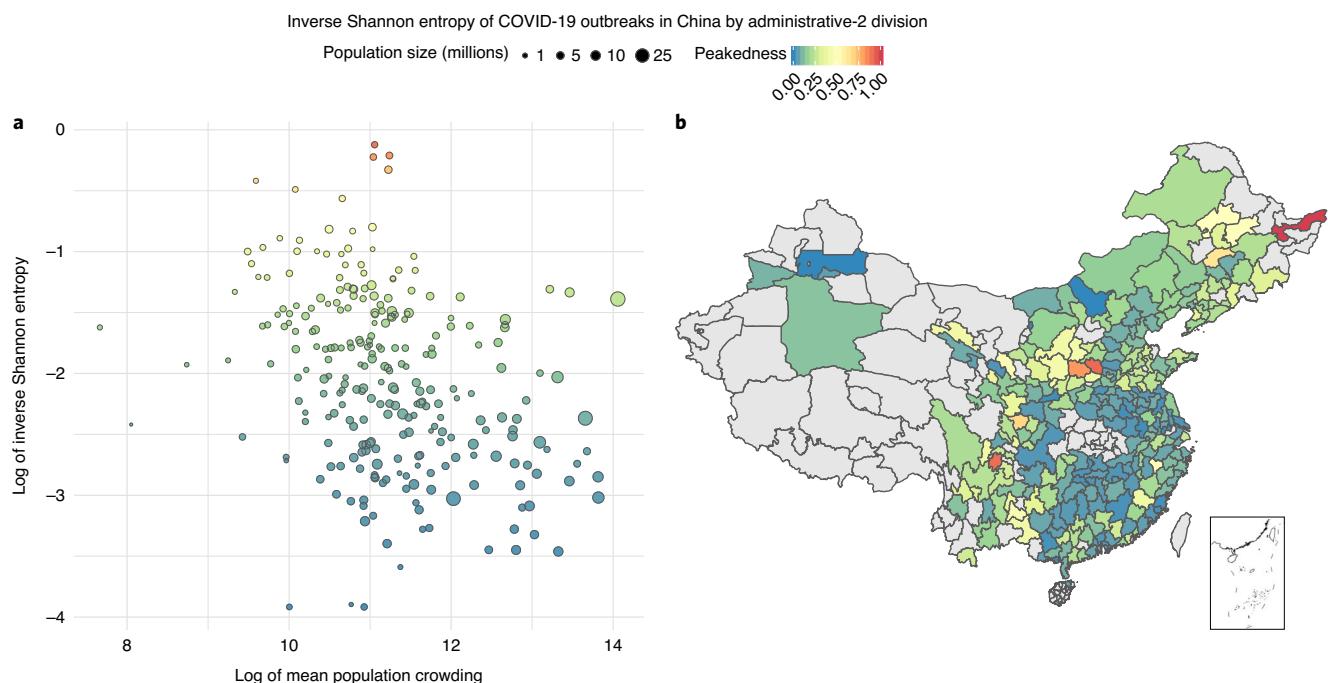


Fig. 2 | Crowding and the temporal clustering of transmission of COVID-19 in China. **a**, Negative association between \log_{10} of epidemic peakedness, as measured by Shannon's diversity index, and \log_{10} population crowding, as measured by Lloyd's mean crowding. The point sizes indicate the size of the population in each city. **b**, Map of epidemic peakedness in China at the prefectural level. Blue and green colors indicate lower peakedness; red and yellow colors indicate higher peakedness. Gray prefectures had either no reported cases or were not included in analyses due to potential inconsistencies in reporting of early cases (Hubei Province).

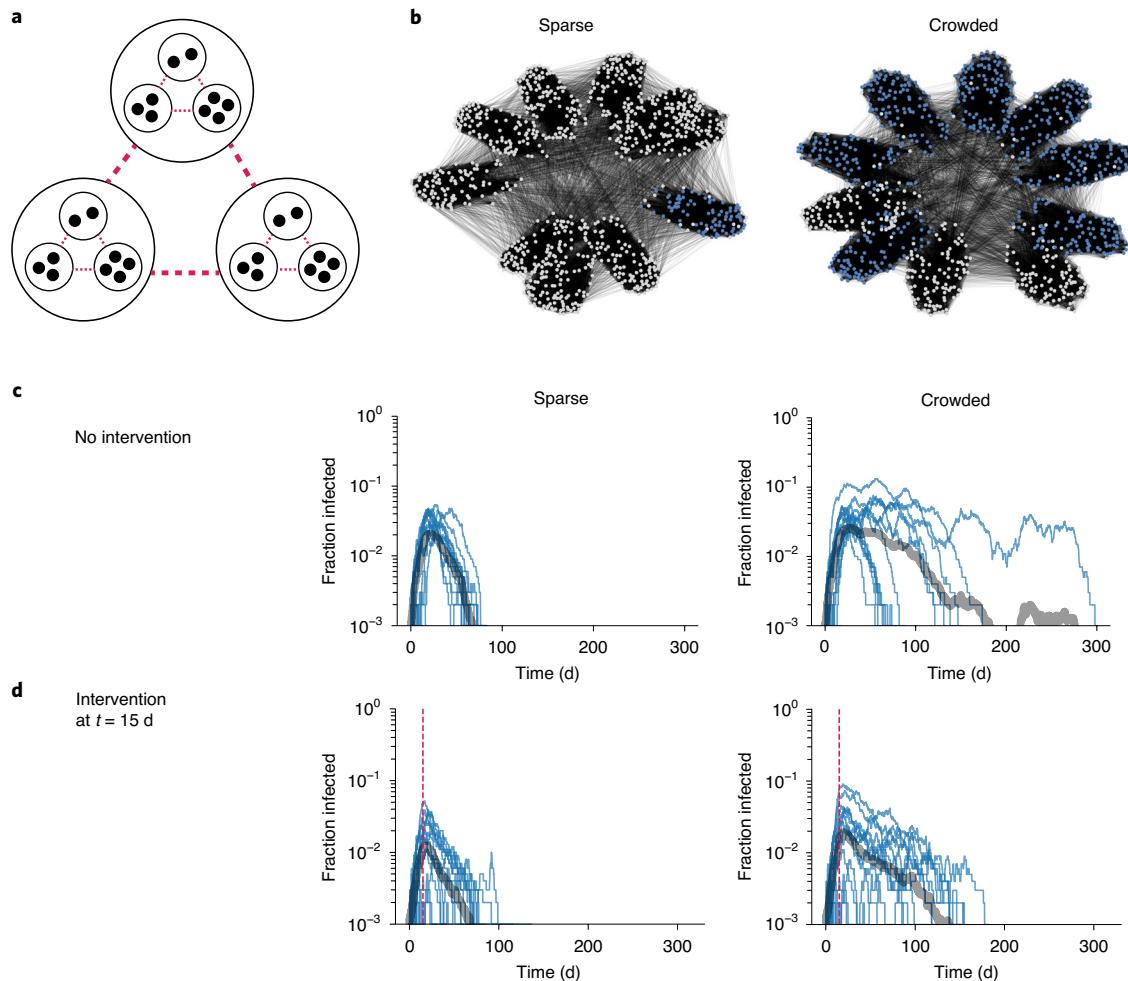


Fig. 3 | Mechanisms generating less peaked epidemics in crowded populations. **a**, Schematic of a hierarchically structured population model consisting of households and ‘neighborhoods’ within a prefecture. Transmission is more likely among contacts connected at lower spatial levels. Crowded populations have a greater number of contacts outside the household, and interventions reduce the number of these connections in both populations (pink dotted lines). **b, c**, Simulated outbreak dynamics in the absence of interventions in crowded versus sparse populations. For the networks in **b**, blue nodes are individuals who were eventually infected by the end of the outbreak. In **c**, thin blue lines show individual realizations of the model, the average shown by the thick gray line. **d**, Simulated outbreak dynamics in the presence of strong social distancing measures in crowded versus sparse populations. The intervention was implemented at day 15 (vertical dotted line) and led to a 75% reduction in contacts, similar to observed changes in contact rates in China^{35,36}. Mean values of median log epidemic peakedness (Shannon index) are -2.3 for low crowding and -2.8 for high crowding.

Simulation of COVID-19 epidemics in hierarchically structured populations. We hypothesize that the mechanism underlying our central observation—that more crowded cities experience less peaked outbreaks—is that crowding enables sustained transmission among households and through a city’s population, leading incidence to be widely distributed through time. To explore this proposed mechanism, we simulated stochastic epidemic dynamics in two types of populations. Simple, well-mixed transmission models in which contact rates are high in crowded regions were not consistent with our findings, because they predict that crowded regions would have more temporally clustered outbreaks. To capture realistic contact patterns, we created hierarchically structured populations²⁹ in which individuals had high rates of contact within their social units (which are defined broadly and could represent households, care homes, hospitals, prisons, etc); lower rates with individuals from other units but within the same neighborhoods; and relatively rare contact with other individuals in other neighborhoods within the same prefecture (Fig. 3a). These assumptions are consistent with reports that most onward transmission after lockdowns were implemented occurred in households or in other close-contact

situations^{2,30}. In this scenario, less crowded prefectures often had more peaked and shorter outbreaks that were isolated to specific neighborhoods, whereas more crowded prefectures could sustain drawn-out outbreaks of larger final size, which jumped among the more highly connected neighborhoods (Fig. 3b,c). Further, if the reproduction number of COVID-19 is over-dispersed^{31–33}, then crowding could enable local outbreaks to spread more widely due to the availability of contacts³⁴.

We also simulated outbreak dynamics under extensive social distancing measures, as observed in Chinese prefectures (75% reduction in contact rates^{35,36}). If social distancing reduces non-household contacts by the same relative amount in all locations, there will be more contacts remaining in crowded areas, because baseline contact rates are higher. Consequently, outbreaks in crowded regions could be larger and take longer to end after intervention (Fig. 3d, Fig. 1c and Extended Data Fig. 1).

Using the fitted model from China paired with globally comprehensive covariates, we extrapolated our results to cities across the world (Fig. 4). Human mobility data from Baidu were not available for locations outside of China. Therefore, we used aggregated

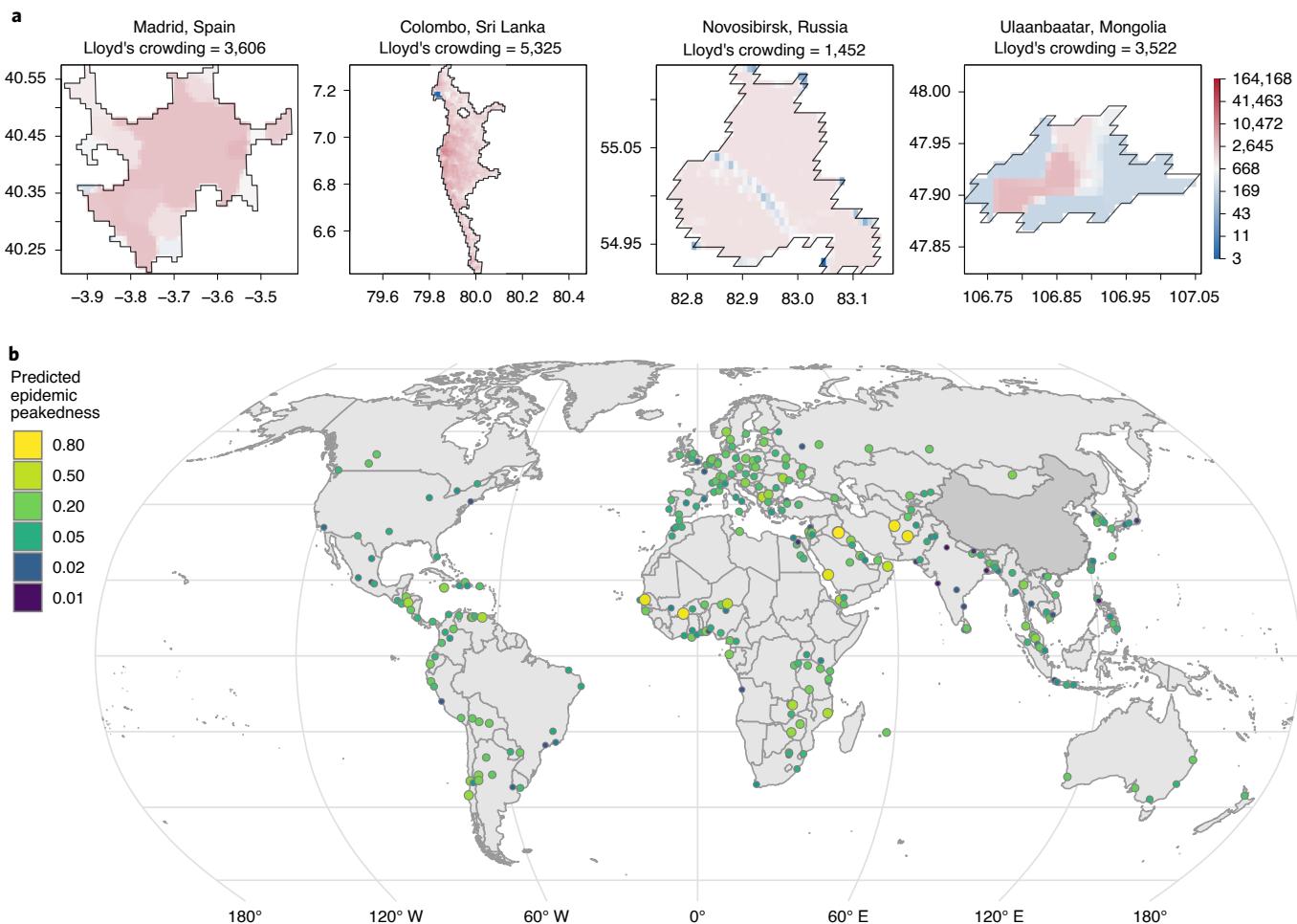


Fig. 4 | Predicted epidemic peakedness across the world. **a**, Maps of cities and their population densities at a 1 × 1-km scale. Madrid, Spain, and Colombo, Sri Lanka, have low predicted peakedness, whereas Novosibirsk, Russia, and Ulaanbaatar, Mongolia, have high predicted peakedness. **b**, Map of predicted epidemic peakedness for 310 cities across the world for which both human population data and mobility data were available for the study period.

human mobility data from Google's COVID Mobility Research Dataset (Methods) to capture relative differences in human mobility through time. At the global scale, cities in yellow are predicted to have concentrated and peaked epidemics, whereas cities in blue are predicted to have more prolonged outbreaks (Fig. 4b; a full list is provided in the Supplementary Information). In general, the epidemics in coastal cities were less peaked and were larger and more prolonged, which could be attributable to high levels of population crowding in coastal cities. These predictions rely on fitted relationships of the first epidemic curves from Chinese and Italian cities and, therefore, should be interpreted very cautiously when generalizing to other settings.

Discussion

Our findings confirm previous work on the peakedness of epidemics transmission for influenza in cities¹³. Our work provides empirical support for the role of spatial organization in determining infectious disease dynamics^{29,37} and, specifically, spatial variability in transmission parameters³⁸. Furthermore, with lower total incidence in small cities compared to larger cities, the risk of resurgence could be elevated owing to lower population immunity after the first wave of the epidemic. Higher seroprevalence for COVID-19 in urban areas³⁹ provides initial data to support these findings; however, there remains an urgent need to expand serological data collection and provide a

full picture of attack rates across cities worldwide⁴⁰. Even though our model does not account for over-dispersion in COVID-19 transmission, there is a theoretical link between the reproduction number in heterogeneous environments and Lloyd's crowding index of aggregation⁴¹, such that the reproduction number increases with higher aggregation³⁴. We report that, in dense cities, reductions in mobility tend to be larger, which potentially elevates the effectiveness of non-pharmaceutical interventions in dense cities⁴². However, assessing the effect of within-city connectivity and its spatial heterogeneity on disease dynamics will be critical to further our understanding of how COVID-19 spreads in urban areas. We found that there is an association between climatic factors and the peakedness of epidemics, but particular caution will need to be applied in interpreting these relationships outside the two studied countries (Italy and China). More work is needed to provide causal evidence for the effect of climatic factors on transmission dynamics of COVID-19 during the pandemic and post-pandemic phases¹⁰.

Currently, non-pharmaceutical interventions are the primary control strategy for COVID-19. As a result, public health measures are often focused on 'flattening the curve' to lower the risk of essential services running out of capacity. We show that spatial context, especially crowding, are important factors for assessing the shape of epidemic curves. Therefore, it will be critical to view non-pharmaceutical interventions through the perspective of

crowding—that is, how does an intervention reduce the circle of contacts of an average individual—in cities across the world.

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/s41591-020-1104-0>.

Received: 29 June 2020; Accepted: 16 September 2020;

Published online: 5 October 2020

References

- Fraher, E. P. et al. Ensuring and sustaining a pandemic workforce. *N. Engl. J. Med.* **382**, 2181–2183 (2020).
- Leung, K., Wu, J. T., Liu, D. & Leung, G. M. First-wave COVID-19 transmissibility and severity in China outside Hubei after control measures, and second-wave scenario planning: a modelling impact assessment. *Lancet* **395**, 1382–1393 (2020).
- Ji, Y., Ma, Z., Peppelenbosch, M. P. & Pan, Q. Potential association between COVID-19 mortality and health-care resource availability. *Lancet Glob. Health* **8**, e480 (2020).
- Rosenbaum, L. Facing Covid-19 in Italy—ethics, logistics, and therapeutics on the epidemic's front line. *N. Engl. J. Med.* **382**, 1873–1875 (2020).
- Tian, H. et al. An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. *Science* **368**, 638–642 (2020).
- Kraemer, M. U. G. et al. The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* **368**, 493–497 (2020).
- Lipsitch, M., Swerdlow, D. L. & Finelli, L. Defining the epidemiology of Covid-19—studies needed. *N. Engl. J. Med.* **382**, 1194–1196 (2020).
- World Health Organization. Coronavirus disease 2019 (COVID-19) Situation Report - 71 https://www.who.int/docs/default-source/coronavirus/situation-reports/2020331-sitrep-71-covid-19.pdf?sfvrsn=4360e92b_8 (2020).
- Zhao, S. et al. Quantifying the association between domestic travel and the exportation of novel coronavirus (2019-nCoV) cases from Wuhan, China in 2020: a correlational analysis. *J. Travel Med.* **27**, 1–3 (2020).
- Baker, R. E., Yang, W., Vecchi, G. A., Metcalf, C. J. E. & Grenfell, B. T. Susceptible supply limits the role of climate in the COVID-19 pandemic. *Science* **369**, 315–319 (2020).
- Rocklöv, J. & Sjödin, H. High population densities catalyse the spread of COVID-19. *J. Travel Med.* **27**, taaa038 (2020).
- Kraemer, M. U. G. et al. Big city, small world: density, contact rates, and transmission of dengue across Pakistan. *J. R. Soc. Interface* **12**, 20150468 (2015).
- Dalziel, B. D. et al. Urbanization and humidity shape the intensity of influenza epidemics in US. cities. *Science* **362**, 75–79 (2018).
- Shaman, J., Pitzer, V. E., Viboud, C., Grenfell, B. T. & Lipsitch, M. Absolute humidity and the seasonal onset of influenza in the continental United States. *PLoS Biol.* **8**, e1000316 (2010).
- Gog, J. R. et al. Spatial transmission of 2009 pandemic influenza in the US. *PLoS Comput. Biol.* **10**, e1003635 (2014).
- Shaman, J. & Kohn, M. Absolute humidity modulates influenza survival, transmission, and seasonality. *Proc. Natl Acad. Sci. USA* **106**, 3243–3248 (2009).
- Chetty, R. et al. The association between income and life expectancy in the United States, 2001–2014. *JAMA* **315**, 1750–1766 (2016).
- Kissler, S. M., Tedijanto, C., Goldstein, E., Grad, Y. H. & Lipsitch, M. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *Science* **21**, 1–9 (2020).
- Crawford, J. M. et al. Laboratory surge response to pandemic (H1N1) 2009 outbreak, New York City Metropolitan Area, USA. *Emerg. Infect. Dis.* **16**, 8–13 (2010).
- Grasselli, G., Pesenti, A. & Cecconi, M. Critical care utilization for the COVID-19 outbreak in Lombardy, Italy. *JAMA* **323**, 1545–1546 (2020).
- Li, Q. et al. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *N. Engl. J. Med.* **382**, 1199–1207 (2020).
- Xu, B. et al. Epidemiological data from the COVID-19 outbreak, real-time case information. *Sci. Data* **7**, 106 (2020).
- Xu, B. et al. Epidemiological data from the COVID-19 outbreak, real-time case information. *figshare* <https://doi.org/10.6084/m9.figshare.11949279> (2020).
- Xu, B. & Kraemer, M. U. G. Open access epidemiological data from the COVID-19. *Lancet Infect. Dis.* **3099**, 30119 (2020).
- Aurora Big Data. 2017 Mobile Map App Research Report: Which of the Highest, the Baidu, and Tencent Is Strong? <https://baijiahao.baidu.com/s?id=1590386747028939917&wfr=spider&for=pc>. (2017)
- Lloyd, M. 'Mean crowding'. *J. Anim. Ecol.* **36**, 1–30 (1967).
- May, R. M. & Anderson, R. M. Spatial heterogeneity and the design of immunization programs. *Math. Biosci.* **72**, 83–111 (1984).
- Anderson, R. M. & May, R. M. *Infectious Diseases of Humans: Dynamics and Control* (Oxford University Press, 1991).
- Watts, D. J., Muhamad, R., Medina, D. C. & Dodds, P. S. Multiscale, resurgent epidemics in a hierarchical metapopulation model. *Proc. Natl Acad. Sci. USA* **102**, 11157–11162 (2005).
- Report of the WHO–China Joint Mission on Coronavirus Disease 2019 (COVID-19) <https://www.who.int/docs/default-source/coronavirus/who-china-joint-mission-on-covid-19-final-report.pdf> 16–24 (2020).
- Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E. & Getz, W. M. Superspreading and the effect of individual variation on disease emergence. *Nature* **438**, 355–359 (2005).
- Kucharski, A. J. et al. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *Lancet Infect. Dis.* **3099**, 1–7 (2020).
- Riou, J. & Althaus, C. L. Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. *Euro. Surveill.* **25**, 1–5 (2020).
- Southwood, T. R. in *Ecological Methods* (ed Southwood, T. R.) 7–69 (Springer Netherlands, 1978).
- Zhang, J. et al. Changes in contact patterns shape the dynamics of the COVID-19 outbreak in China. *Science* **368**, 1481–1486 (2020).
- Lai, S. et al. Effect of non-pharmaceutical interventions to contain COVID-19 in China. *Nature* **585**, 410–413 (2020).
- Sattenspiel, L. Simulating the effect of quarantine on the spread of the 1918–19 flu in central Canada. *Bull. Math. Biol.* **65**, 1–26 (2003).
- Meyers, L. A. Contact network epidemiology: bond percolation applied to infectious disease prediction and control. *Bull. Am. Math. Soc.* **44**, 63–87 (2006).
- Kissler, S. M. et al. Reductions in commuting mobility predict geographic differences in SARS-CoV-2 prevalence in New York City. *Nat. Commun.* **16**, 4674 (2020).
- Lipsitch, M., Swerdlow, D. L. & Finelli, L. Defining the epidemiology of Covid-19 — studies needed. *N. Engl. J. Med.* **382**, 1194–1196 (2020).
- Mat, N. F. C., Edinur, H. A., Razab, M. K. A. A. & Safuan, S. A single mass gathering resulted in massive transmission of COVID-19 infections in Malaysia with further international spread. *J. Travel Med.* **27**, taaa059 (2020).
- Flaxman, S. et al. Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries. *Nature* **584**, 257–261 (2020).

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

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2020

Methods

Epidemiological data. No officially reported line list was available for cases in China. We used a standardized protocol⁴³ to extract individual-level data from December 1, 2019, to March 30, 2020. Sources were mainly official reports from provincial, municipal or national health governments. Data included basic demographics (age and sex), travel histories and key dates (dates of onset of symptoms, hospitalization and confirmation). Data were entered by a team of data curators on a rolling basis, and technical validation and geo-positioning protocols were applied continuously to ensure validity. A detailed description of the methodology is available²². Lastly, total numbers were matched with officially reported data from China and other government reports. Daily case counts from Italian provinces ($n = 107$) were extracted from the Presidenza del Consiglio dei Ministri Dipartimento della Protezione Civile (<https://github.com/pcm-dpc/COVID-19>).

Estimating epidemic peakedness. Epidemic peakedness was estimated for each prefecture by calculating the inverse Shannon entropy of the distribution of COVID-19 cases. Inverse Shannon entropy was used to fit time series of other respiratory infections (influenza)¹³. The inverse Shannon entropy⁻¹ of incidence for a given prefecture in 2020 is then given by $v_j = (-\sum_i p_{ij} \log p_{ij})^{-1}$. Because v_j is a function of incidence distribution in each location rather than raw incidence, it is invariant under differences in overall reporting rates between cities or attack rates. We then assessed how peakedness $v \propto \sum_j v_j$ varied across geographic areas in China. As an alternative measure of temporal clustering of cases, we estimated the proportion of cases at the peak ± 1 d (Extended Data Fig. 2).

Proxies for COVID-19 interventions using within-city human mobility data from China. Estimates of within-city reductions of human mobility between the period before and after the lockdown was implemented on January 23, 2020, were extracted from Lai et al.³⁶. Daily measures of human mobility were extracted from the Baidu Qianxi web platform to estimate the proportion of daily movement within prefectures in China. Relative mobility volume was available from January 2, 2020, to January 25, 2020. For each city, change in relative mobility was defined by $m_i = m_i(\text{lockdown})/m_i(\text{baseline})$, where m_i is defined as mobility in prefecture i . Baidu's mapping service is estimated to have a 30% market share in China, and more data can be found³⁶.

Data on drivers of transmission of COVID-19. Prefecture-specific population counts and densities were derived from the 2020 Gridded Population of The World, a modeled continuous surface of population estimated from national census data and the United Nations World Population Prospects⁴⁴. Population counts are defined at a 30-arc-second resolution (approximately 1 km × 1 km at the equator) and extracted within administrative 2 level cartographic boundaries defined by the National Bureau of Statistics of China. Lloyd's mean crowding, $\frac{\sum_i (q_i - 1)q_i}{\sum_i q_i}$, was estimated for each prefecture, where q_i represents the population count of each non-zero pixel within a prefecture's boundary and the resulting value estimates an individual's mean number of expected neighbors^{13,45}. When fitting the models, we consider the numerator $[\sum_i (q_i - 1)q_i]$, which we refer to as 'contacts', and the denominator $\sum_i q_i$ (that is, population size) as separate predictors. We note that a negative slope for 'contacts' and a positive slope for 'population' support a negative coefficient for Lloyd's mean crowding.

Daily temperature (°F), relative humidity (%) and atmospheric pressure (Pa) at the centroid of each prefecture was provided by The Dark Sky Company via the Dark Sky API and aggregated across a variety of data sources. Specific humidity (kg/kg) was then calculated using the R package `humidity`⁴⁶. Meteorological variables for each prefecture were then averaged across the entirety of the study period.

Statistical analysis. We normalized the values of epidemic peakedness between 0 and 1 and, for all non-zero values, fit a generalized linear model of the form

$$\log(Y_j) \sim \beta_0 + \beta_1 \log(C_j) + \beta_2 \log(h_j) + \beta_3 \log(P_j) + \beta_4 \log(f_j) + \beta_5 \log(t_j)$$

where, for each prefecture j , Y is the scaled inverse Shannon entropy measure of epidemic peakedness derived from the COVID-19 time series; C is the mean number of contacts^{26,46}; h is the mean specific humidity over the reporting period in kg/kg; P is the estimated population density; f is the relative change in population flows within each prefecture; and t is daily mean temperature.

Projecting epidemic peakedness in cities around the world. We selected 310 urban centers from the European Commission Global Human Settlement Urban Centre Database and their included cartographic boundaries⁴⁷. To ensure global coverage, up to the five most populous cities in each country were selected from the 1,000 most populous urban centers recorded in the database. Population count, crowding and meteorological variables were then estimated following identical procedures used to calculate these variables in the Chinese prefectures. Weather measurements were averaged over the 2-month period starting on February 1, 2020.

The parameters from the model of epidemic peakedness predicted by humidity, crowding and population size (Supplementary Table 1, model 6) were used to

estimate relative peakedness in the 310 urban centers. A full list of predicted epidemic peakedness values can be found in Supplementary Table 3.

Global human mobility data. We used the Google COVID-19 Aggregated Mobility Research Dataset, which contains anonymized relative mobility flows aggregated over users who have turned on the Location History setting, which is off by default. This is similar to the data used to show how busy certain types of places are in Google Maps, helping identify when a local business tends to be the most crowded. The mobility flux is aggregated per week, between pairs of approximately 5-km² cells worldwide, and for the purpose of this study aggregated for 310 cities worldwide. We calculated both mobility within each city's shapefile and mobility coming into each city. For each city, change in relative mobility was defined by $m_i = m_i(\text{April})/m_i(\text{December})$, where m_i is defined as mobility in city i .

To produce this data set, machine learning was applied to log data to automatically segment it into semantic trips⁴⁸. To provide strong privacy guarantees, all trips were anonymized and aggregated using a differentially private mechanism⁴⁹ to aggregate flows over time (<https://policies.google.com/technologies/anonymization>). This research is done on the resulting heavily aggregated and differentially private data. No individual user data were ever manually inspected; only heavily aggregated flows of large populations were handled.

All anonymized trips were processed in aggregate to extract their origin and destination location and time. For example, if users traveled from location a to location b within time interval t , the corresponding cell (a, b, t) in the tensor would be $n \pm \text{err}$, where err is Laplacian noise. The automated Laplace mechanism adds random noise drawn from a zero-mean Laplace distribution and yields (ϵ, δ) -differential privacy guarantee of $\epsilon = 0.66$ and $\delta = 2.1 \times 10^{-29}$. The parameter ϵ controls the noise intensity in terms of its variance, whereas δ represents the deviation from pure ϵ -privacy. The closer they are to zero, the stronger the privacy guarantees. Each user contributes, at most, one increment to each partition. If they go from a region a to another region b multiple times in the same week, they contribute only once to the aggregation count.

These results should be interpreted in light of several important limitations. First, the Google mobility data are limited to smartphone users who have opted in to Google's Location History feature, which is off by default. These data might not be representative of the population as whole, and, furthermore, their representativeness might vary by location. Importantly, these limited data are viewed only through the lens of differential privacy algorithms, specifically designed to protect user anonymity and obscure fine detail. Moreover, comparisons across, rather than within, locations are descriptive only because these regions can differ in substantial ways.

Simulating epidemic dynamics. We simulated a simple stochastic SIR model of infection spread on weighted networks created to represent hierarchically structured populations. Individuals were first assigned to households using the distribution of household sizes in China (data from the United Nations Population Division; mean, 3.4 individuals). Households were then assigned to 'neighborhoods' of ~100 individuals, and all neighborhood members were connected with a lower weight. A randomly chosen 10% of individuals were given 'external' connections to individuals outside the neighborhood. The total population size was $n = 1,000$. Simulations were run for 300 d, and averages were taken over 20 iterations. The SIR model used a per-contact transmission rate of $\beta = 0.15$ per day and recovery rate $\gamma = 0.1$ per day. For the simulations without interventions, the weights were $w_{HH} = 1$, $w_{NH} = 0.01$ and $w_{EX} = 0.001$ for the crowded prefecture and $w_{EX} = 0.0001$ for the less crowded prefecture. For the simulations with interventions, the household and neighborhood weights were the same, but we used $w_{EX} = 0.01$ for the crowded prefecture and $w_{EX} = 0.001$ for the 'sparse' prefecture. The intervention reduced the weight of all connections outside the household by 75%.

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

Data availability

We collated epidemiological data from publicly available data sources (news articles, press releases and published reports from public health agencies) that are described in full in ref. ²². Epidemiological and spatial data used in this study are available via Github (https://github.com/Emergent-Epidemics/COVID_crowding). The Google COVID-19 Aggregated Mobility Research Dataset used for this study is available with permission from Google. Code and data are also available at <https://zenodo.org/record/4056578#.X3IFF5NKiek>.

Code availability

The code associated with the data analysis and statistics is available from https://github.com/Emergent-Epidemics/COVID_crowding. The simulation code is available from <https://github.com/alsnhll/SIRNestedNetwork>. Code and data are also available at <https://zenodo.org/record/4056578#.X3IFF5NKiek>.

References

43. Ramshaw, R. E. et al. A database of geopositioned Middle East respiratory syndrome coronavirus occurrences. *Sci. Data* **6**, 318 (2019).
44. Doxsey-Whitfield, E. et al. Taking advantage of the improved availability of census data: a first look at the gridded population of the world, version 4. *Pap. Appl. Geogr.* **1**, 226–234 (2015).
45. Reiczigel, J., Lang, Z., Rózsa, L. & Tóthmérész, B. Properties of crowding indices and statistical tools to analyse parasite crowding data. *J. Parasitol.* **91**, 245–252 (2005).
46. Wade, M. J., Fitzpatrick, C. L. & Lively, C. M. 50-year anniversary of Lloyd's 'mean crowding': ideas on patchy distributions. *J. Anim. Ecol.* **87**, 1221–1226 (2018).
47. Florczyk, A. et al. GHS-UCDB R2019A - GHS Urban Centre Database 2015, multitemporal and multidimensional attributes <https://data.jrc.ec.europa.eu/dataset/53473144-b88c-44bc-b4a3-4583ed1f547e> (2019).
48. Bassolas, A. et al. Hierarchical organization of urban mobility and its connection with city livability. *Nat. Commun.* **10**, 4817 (2019).
49. Wilson, R. J. et al. Differentially private SQL with bounded user contribution. Preprint at <https://arxiv.org/abs/1909.01917> (2019).

Acknowledgements

The authors thank K. Cordiano for statistical assistance. We thank the members of the Open COVID-19 Data Working Group. The members of the group are listed in the Supplementary Note. B.R. acknowledges funding from Google.org. M.U.G.K. acknowledges funding from the European Commission H2020 program (MOOD project) and a Branco Weiss Fellowship. O.G.P., M.U.G.K., A.E.Z. and H.T. acknowledge funding from the Oxford Martin School. H.T. acknowledges funding from the Beijing Science and Technology Planning Project (Z201100005420010). A.L.H. and A.N. acknowledge funding from the National Institutes of Health (DP5OD019851). The funding bodies had no role in study design, data collection and analysis,

preparation of the manuscript or the decision to publish. All authors saw and approved the manuscript.

Author contributions

M.U.G.K., O.G.P. and S.V.S. conceived the research. B.R., A.L.H., A.N., B.A., S.V.S. and M.U.G.K. analyzed the data. B.R. and S.V.S. analyzed the human mobility data. C.D., O.G.P., M.U.G.K. and S.V.S. interpreted the data. M.U.G.K. wrote the first draft of the manuscript. All authors contributed to the interpretation of results and manuscript writing.

Competing interests

SVS is a paid consultant with Pandefense Advisory and Booze Allen Hamilton; is on the advisory board for BioFire Diagnostics Trend Surveillance, which includes paid consulting; and holds unexercised options in Iliad Biotechnologies. These entities provided no financial support associated with this research, did not have a role in the design of this study, and did not have any role during its execution, analyses, interpretation of the data and/or decision to submit.

Additional information

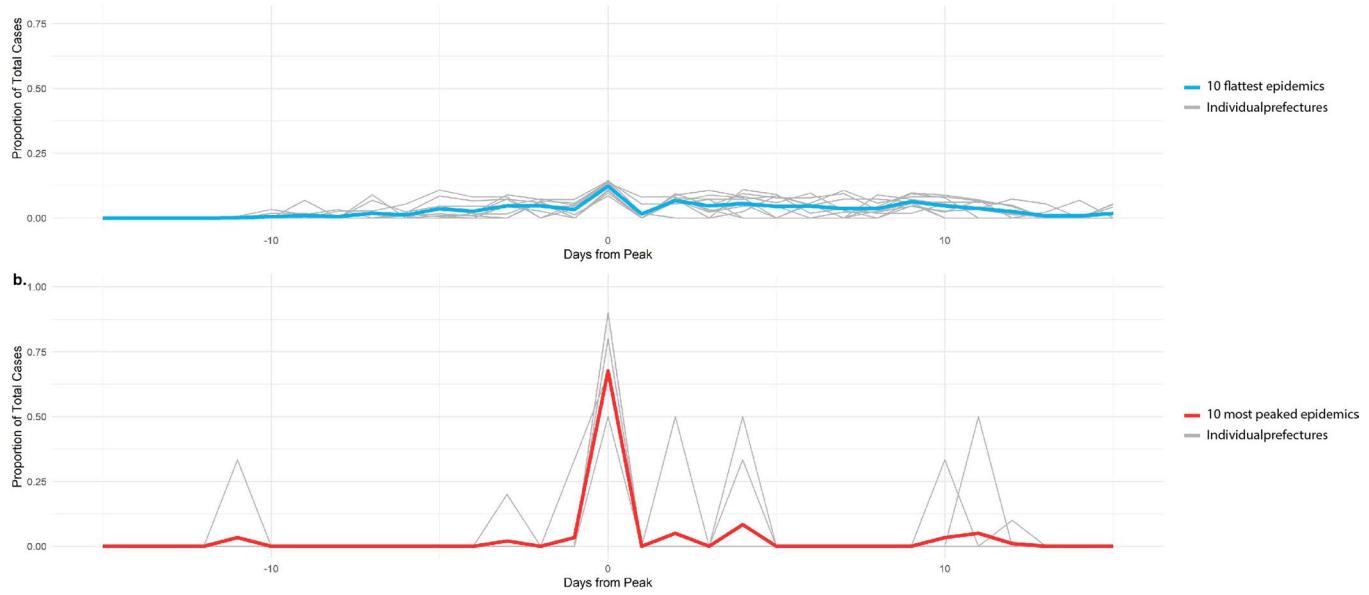
Extended data is available for this paper at <https://doi.org/10.1038/s41591-020-1104-0>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41591-020-1104-0>.

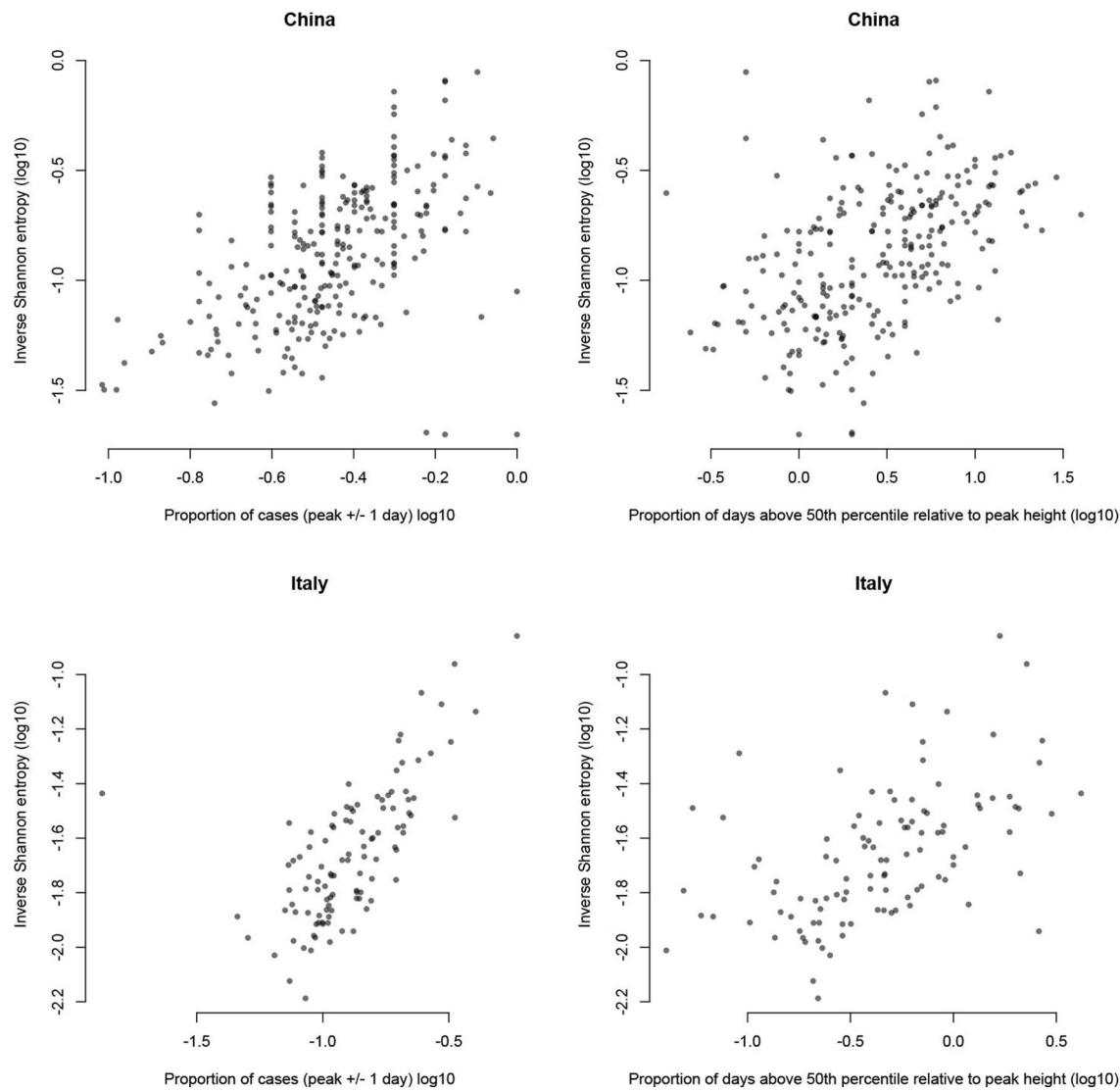
Correspondence and requests for materials should be addressed to S.V.S., O.G.P. or M.U.G.K.

Peer review information Jennifer Sargent was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

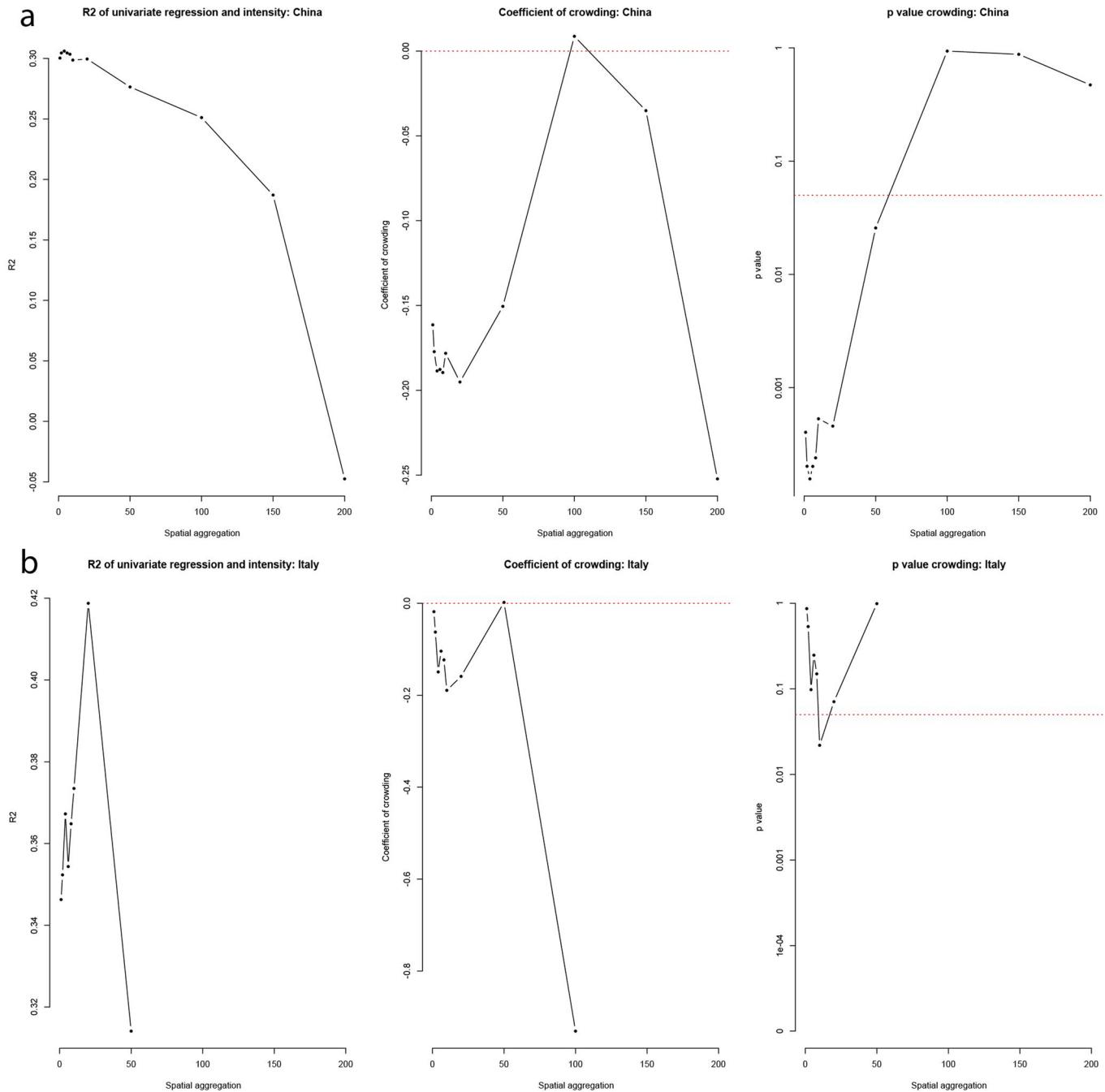
Reprints and permissions information is available at www.nature.com/reprints.



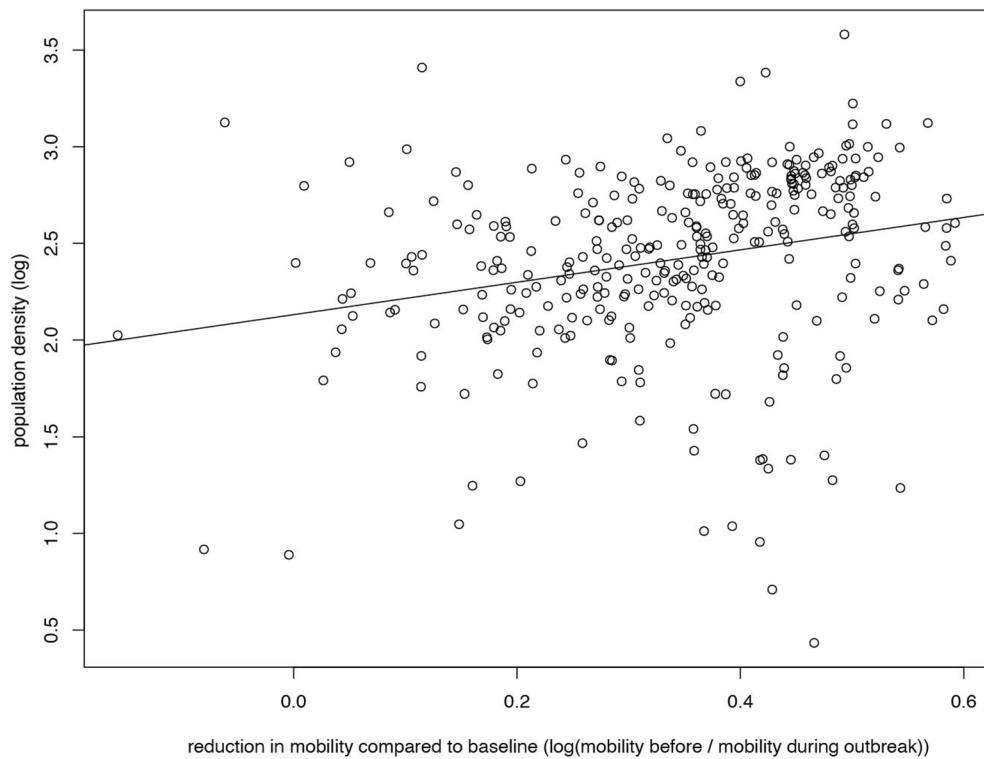
Extended Data Fig. 1 | Proportion of daily cases in prefectures in China. **a**, shows the ten flattest epidemics and **b**, shows the most peaked epidemics. Red and blue curves indicate the average across these prefectures.



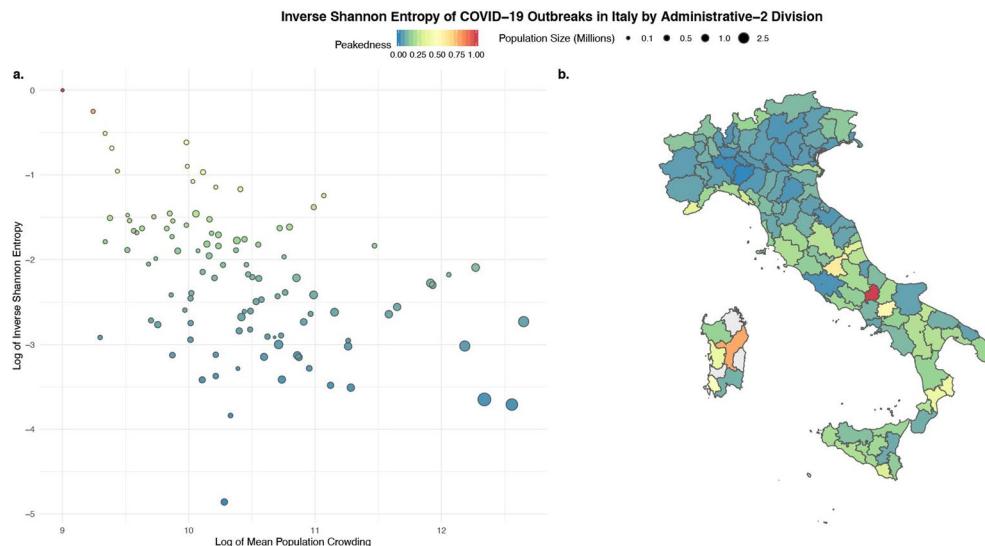
Extended Data Fig. 2 | Left panels: Proportion of cases at the peak (± 1 day) vs. inverse Shannon entropy for prefectures in China ($n=262$) and regions in Italy ($n=107$). Right panels: Proportion of days in the epidemic curve that had cases above the 50th percentile, normalized by the largest reported number of cases versus inverse Shannon entropy for prefectures in China ($n=262$) and regions in Italy ($n=107$).



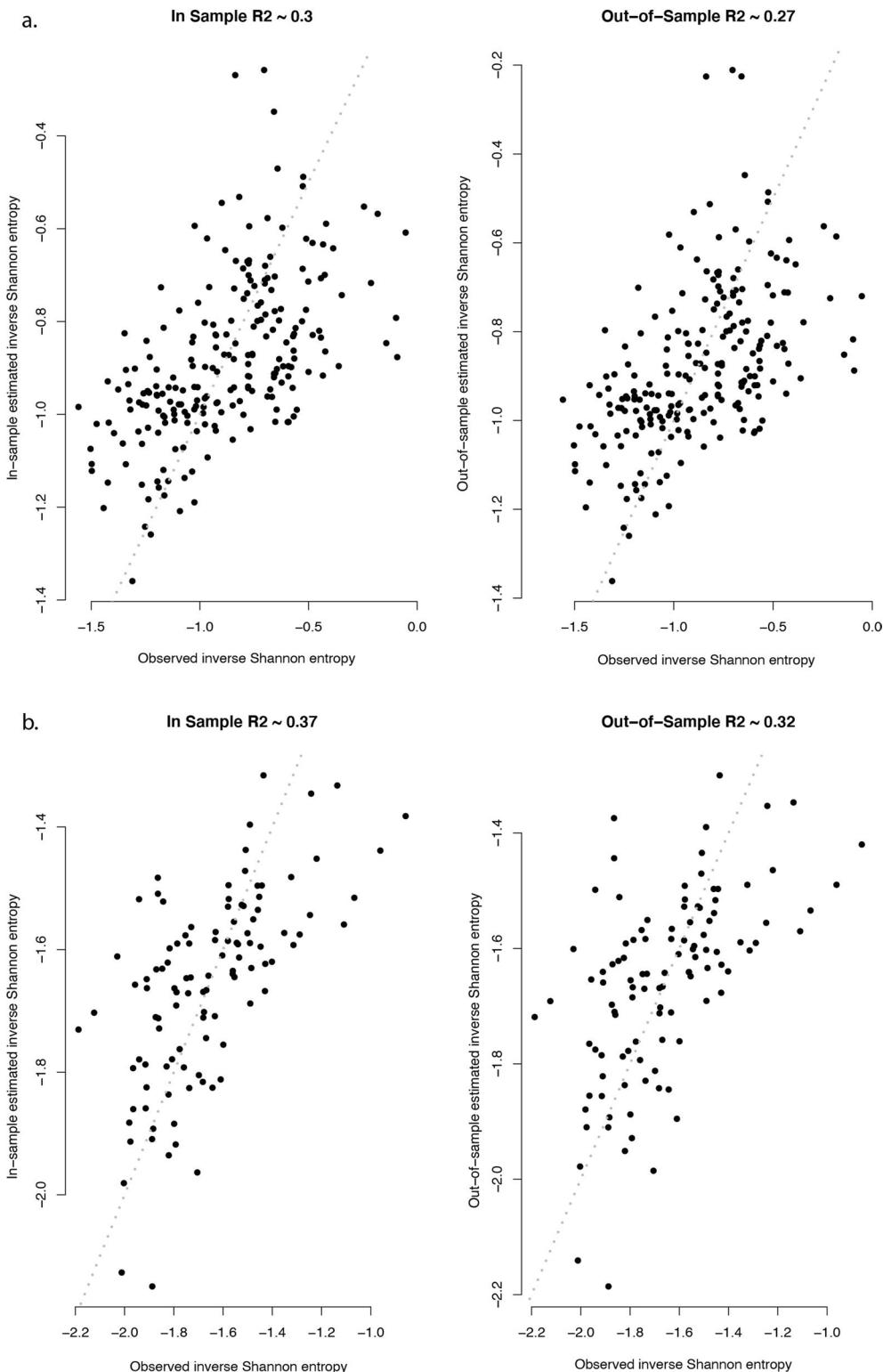
Extended Data Fig. 3 | Epidemic peakedness is well explained by covariates at spatial scales from 1 – 50 km in China **a**, and Italy **b**.



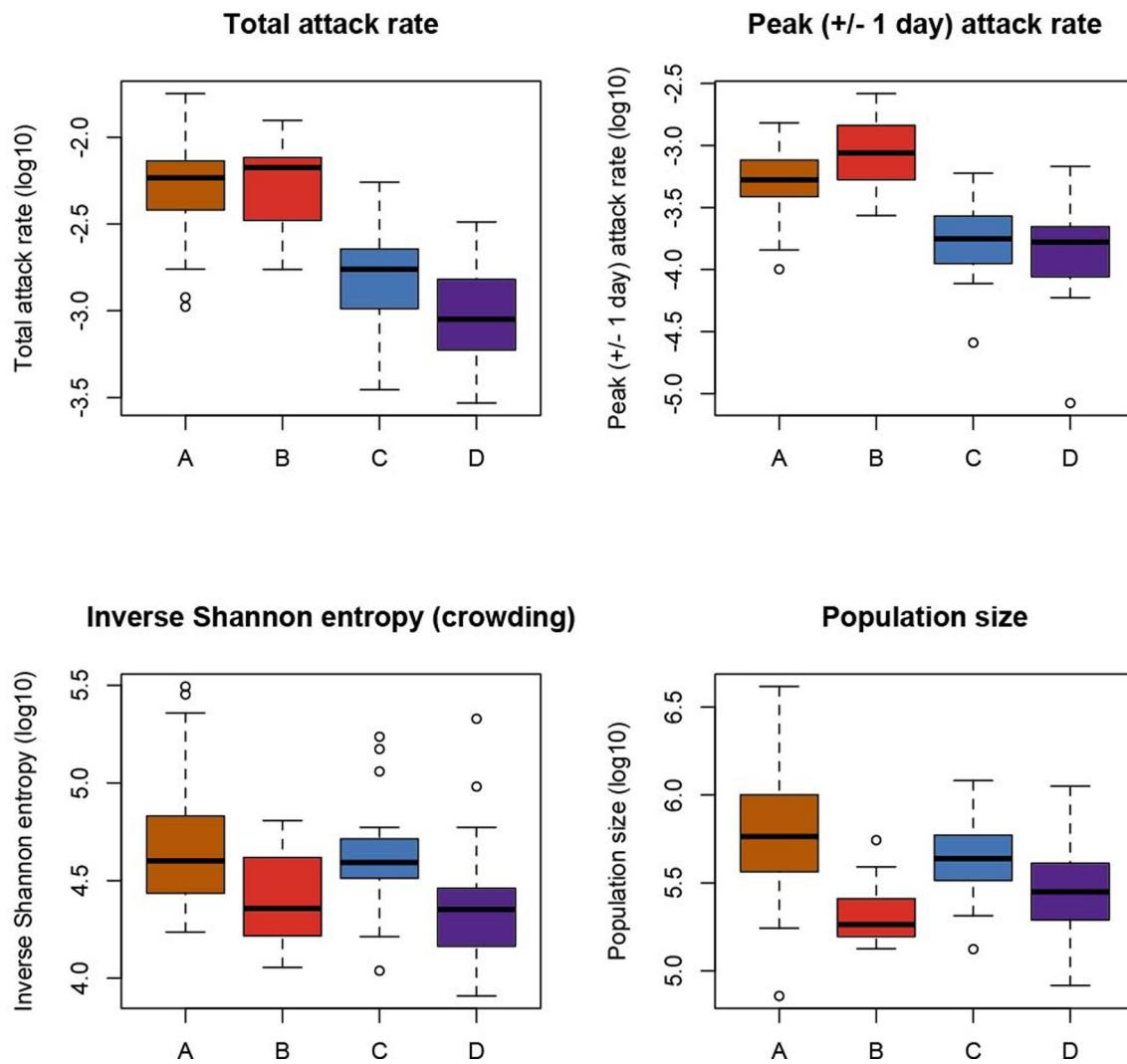
Extended Data Fig. 4 | Relationship between population density and reduction in mobility in cities in China. Each dot represents one city in China. Data on human mobility are extracted from Baidu Inc. and are available from Lai et al. 2020³⁶.



Extended Data Fig. 5 | Crowding and the temporal clustering of transmission of COVID-19 in Italy. **a**, negative association between log₁₀ of epidemic peakedness, as measured by Shannon's diversity index (Methods), and log population crowding, as measured by Lloyd's mean crowding (Methods). The point sizes indicate the size of the population in each city. **b**, Map of epidemic peakedness in Italy at the provincial level. Blue and green colours indicate lower peakedness and red and yellow colours higher peakedness. Grey prefectures had either no or very limited amount of reported cases. Values were rescaled so that Shannon index in each province = (Shannon index – min(Shannon index)) / (max(Shannon index) – min(Shannon index)).



Extended Data Fig. 6 | Out-of-sample prediction (n -fold cross validation) over all prefectures in China **a, b**, and all provinces in Italy **c, d**. In-sample prediction in China ($R^2 = 0.32$) compares well to out-of-sample predictions ($R^2 = 0.28$). In-sample prediction in Italy ($R^2 = 0.38$) compares well to out-of-sample predictions ($R^2 = 0.32$).



Extended Data Fig. 7 | Total attack rate, peak attack rate, inverse Shannon entropy and population sizes for: **a**, shows prefectures in China with low peakedness and high variance (as measured by the variance in the first difference of the time series of daily new cases). These prefects have high population and high crowding. **b**, shows prefectures in China with high intensity and high variance. **c**, shows prefectures in China that have low peakedness and low variance. **d**, shows prefectures in China that have high peakedness and low variance.

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 our [Editorial Policies](#) 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

Daily temperature (°F), relative humidity (%) and atmospheric pressure (Pa) at the centroid of each prefecture was provided by The Dark Sky Company via the Dark Sky API and aggregated across a variety of data sources (<https://darksky.net/>). For all other data sources no code was used.

Data analysis

R statistical software was used for the analysis: R version 3.5.1

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 and 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

We collated epidemiological data from publicly available data sources (news articles, press releases and published reports from public health agencies) which are described in full here²². Epidemiological and spatial data used in this study is available via Github (https://github.com/Emergent-Epidemics/covid_hierarchy). The Google COVID-19 Aggregated Mobility Research Dataset used for this study is available with permission from Google, LLC.

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

Ecological, evolutionary & environmental sciences study design

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

Study description

We use empirical epidemiological and spatial data to understand the intensity of transmission of COVID-19 across cities in China. Models are then used to predict epidemic intensity across a set of global cities.

Research sample

293 cities in China were used that had reported transmission of COVID-19. Prefectures in Hubei were excluded because they are distinct in their epidemiology due to the early spread of the virus there. Time series epidemiological data for Italy were available at the province level. The complete sample for data from Italy was included in this analysis.

Sampling strategy

No sampling strategy was performed. However, prefectures in Hubei were excluded. Data were sufficient in their spatial and temporal distribution for the analysis performed in this manuscript.

Data collection

Data collection is described in detail in an associated publication: <https://www.nature.com/articles/s41597-020-0448-0>

Timing and spatial scale

Epidemiological data for China was used from January 1 - March 30, 2020. Spatial scale was prefecture level in China (n = 293). Data from Italy was available at the prefecture level. Both datasets covered the first initial epidemic from arrival to establishment and subsequent decline in cases.

Data exclusions

Prefectures in Hubei were excluded. The epidemiological situation in Hubei is markedly different due to the early spread of the virus out of Wuhan (Hubei province).

Reproducibility

All code and data is available so that results can be reproduced.

Randomization

No randomization was performed and was not necessary for the correlative analysis performed in this work.

Blinding

No blinding was performed and necessary. Data was aggregated to daily case counts and did not include any individual level data.

Did the study involve field work? Yes No

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
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging