PLOS ONE

RESEARCH ARTICLE

# Population density and spreading of COVID-19 in England and Wales

**Jack Sutton**[1], **Golnaz Shahtahmassebi**[1], **Haroldo V. Ribeiro**[2], **Quentin S. Hanley**[1]*

**1** School of Science and Technology, Nottingham Trent University, Clifton Lane, Nottingham, United Kingdom, **2** Departamento de Física, Universidade Estadual de Maringá, Maringá, Brazil

* quentin.hanley@ntu.ac.uk





**Citation:** Sutton J, Shahtahmassebi G, Ribeiro HV, Hanley QS (2022) Population density and spreading of COVID-19 in England and Wales. PLoS ONE 17(3): e0261725. https://doi.org/10.1371/journal.pone.0261725

**Editor:** José S. Andrade, Jr., Universidade Federal do Ceara, BRAZIL

**Received:** July 9, 2021

**Accepted:** December 7, 2021

**Published:** March 31, 2022



**Data Availability Statement:** All relevant data are within the paper and its Supporting information files.

**Funding:** H.V.R was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) Grant Nos. 407690/2018-2 and 303121/2018-1. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

## Abstract

We investigated daily COVID-19 cases and deaths in the 337 lower tier local authority regions in England and Wales to better understand how the disease propagated over a 15-month period. Population density scaling models revealed residual variance and skewness to be sensitive indicators of the dynamics of propagation. Lockdowns and schools reopening coincided with increased variance indicative of conditions with local impact and country scale heterogeneity. University reopening and December holidays reduced variance indicative of country scale homogenisation which reached a minimum in mid-January 2021. Homogeneous propagation was associated with better correspondence with normally distributed residuals while heterogeneous propagation was more consistent with skewed models. Skewness varied from strongly negative to strongly positive revealing an unappreciated feature of community propagation. Hot spots and super-spreading events are well understood descriptors of regional disease dynamics that would be expected to be associated with positively skewed distributions. Positively skewed behaviour was observed; however, negative skewness indicative of "cold-spots" and "super-isolation" dominated for approximately 8 months during the period of study. In contrast, death metrics showed near constant behaviour in scaling, variance, and skewness metrics over the full period with rural regions preferentially affected, an observation consistent with regional age demographics in England and Wales. Regional positions relative to density scaling laws were remarkably persistent after the first 5–9 days of the available data set. The determinants of this persistent behaviour probably precede the pandemic and remain unchanged.

## Introduction

SARS-CoV-2 spread rapidly from a cluster of cases in China in late 2019 to a global pandemic on 13 March 2020. The number of confirmed cases of COVID-19 continues to grow worldwide with over 186 million cases and over 4 million deaths. SARS-CoV-2 is thought to spread by direct contact, fomites, and aerosols from both symptomatic and asymptomatic people [1–4]. During the pandemic, distancing measures and meeting size restrictions have been widely deployed to slow the spread of the disease by reducing the number and duration of interactions capable of causing infection. At scale, population density could be a proxy for

these interactions. For example, someone living in a region of high population density is expected to have a greater number of interactions compared with someone who lives in a rural setting [5].

The propagation of COVID-19 via super-spreading events has been documented [6–10] these events are reported to have fat tails and distributions presented show strong positive skew. For subsequent modelling a range of distributions have been used including Weibull [11], Poisson [12], gamma [7], and normal [13] and subsequent modelling is primarily based on population. The effects of population size on COVID-19 dynamics have been investigated previously including aspects of population density effects [14–17]. Investigations of population density effects have been limited to a relatively small number of time points aggregated over a period of time, usually a month or year [18–24]. Daily granularity of data is not easily accessible; however, the COVID-19 pandemic has provided a unique and evolving data set with daily updates for generating an extended scaling time series. These data have been influential in informing government interventions, policy decisions, and public perceptions allowing data driven informed decisions.

These daily data at relatively high regional granularity provide an opportunity to document the daily evolution of scaling metrics, descriptive statistics, and residual variance over an extended period. Here, we investigated scaling behaviour in England and Wales using daily COVID-19 cases and deaths in England and Wales Lower Tier Local Authorities (LTLAs) with population density. Additionally, we use age categories ranging from 0–4 years old to 85 + years old to investigate the impact that age demographics have on COVID-19 death. These were examined to better understand how infectious disease metrics progress over time at country scale.

## Scaling models

Urban scaling [25] considers population to predict a range of urban indicators. A variety of mathematical forms have been applied with power laws being widely used.

$$Y = Y_0 P^{\beta} 10^{\varepsilon} \tag{1}$$

Here, $Y$ is the indicator, $P$ is the population, $\beta$ is the scaling exponent, $Y_0$ is the pre-exponential factor and $\varepsilon$ are residuals that are independent and identically distributed with common $N(0, \sigma^2)$ distribution. An estimate to the parameter $\beta$, can be obtained by applying the least square method to the logarithmic version of Eq 1 (*i.e. logY vs. logP*) which aims to minimise the value $\sum \hat{\varepsilon}_i^2$.

When combining rural and urban regions, density metrics provide better models [20,21] than population. This can be described by similar power-law functions of the form

$$Y_D = Y_0 P_D^{\beta_D} 10^{\varepsilon} \tag{2}$$

where $Y_D$ is the indicator density, $P_D$ is the population density and $\beta_D$ is the density scaling exponent. Indicator and population densities are obtained by dividing them by the corresponding defined regional area, $A$ (*i.e.* $Y_D = Y/A$ *and* $P_D = P/A$). Similarly, to population scaling, when $\beta_D < 1$ scaling is sub-linear, when $\beta_D = 1$, the scaling is linear and when $\beta_D > 1$ the scaling is super-linear. When interpreting density scaling results, sub-linear scaling accelerates in rural (low-density) regions and super-linear scaling accelerates in urban (high density) areas. The log transformed data is usually fitted to the logarithmic form

$$\log(Y_D) = \log\ Y_0 + \beta_D \log(P_D) + \varepsilon \tag{3}$$

to obtain the regression model parameters.

Eq 3 recognises a linear relationship between the indicator density and population density. In some circumstances Eq 3 needs to be adjusted to account for a breakpoint to allow for a segmented fit. Empirically, the breakpoint, $d^*$, for a range of indicators usually occurs in the range of 10–70 people per hectare [19,21]. Thus, Eq 3 can be adapted to allow for such fit and is given by

$$\log(Y_D) = \begin{cases} \log\ Y_0 + \beta_L \log(P_D) + \varepsilon & d < d^* \\ \log\ Y_1 + \beta_H \log(P_D) + \varepsilon & d \geq d^* \end{cases} \tag{4}$$

Where $\beta_L$ and $Y_0$ are the exponent and pre-exponential factors below the breakpoint; $\beta_H$ and $Y_1$ are the exponent and pre-exponential factor above the breakpoint.

Residuals, $\varepsilon_i$, from the fit to the model defined in Eqs (3) and (4) are obtained using least squares method which aims to minimise the variance $\sum \varepsilon_i^2$ for

$$\varepsilon_i = log(Y_{D,i}) - log(\hat{Y}_{D,i}) \tag{5}$$

for $i = 1, \ldots, n$ and $log(\hat{y}_{D,i})$ is the estimate of $log(Y_{D,i})$. Negative values of $\varepsilon_i$ are below expectation while positive $\varepsilon_i$ are above expectation.

After obtaining residuals from the preferred model (Eqs 3 and 4), a similarity measure is computed to assess correlation. If residuals are represented as $X = (x_1, x_{2,\ldots}, x_n)$ and $Y = (y_1, y_{2,\ldots}, y_n)$ for $n$ complete set of regions between indicators then Pearson's correlation $(r(X,Y)))$ is computed as

$$r(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y} \tag{6}$$

where $cov(X,Y)$ is the covariance; $\sigma_X$ and $\sigma_Y$ is the standard deviation of $X$ and $Y$ respectively. Spearman's rank correlation $(S(R(X), R(Y))$ is less sensitive to strong outliers in comparison to Pearson's correlation and is computed as

$$(S(R(X), R(Y)) = \frac{cov(R(X), R(Y))}{\sigma_{R(X)} \sigma_{R(Y)}} \tag{7}$$

where $cov(R(X),R(Y))$ is the covariance of the rank variables; $\sigma_{R(X)}$ and $\sigma_{R(Y)}$ is the standard deviation of $X$ and $Y$ respectively.

## Residual and case density models

The distribution of residuals obtained from the England was modelled using normal and generalised logistic (GL) distributions. The latter has the form,

$$GL(x; \theta, \sigma, \alpha) = \frac{\alpha}{\sigma} \frac{e^{-\frac{x-\theta}{\sigma}}}{\left\{1 + e^{-\frac{x-\theta}{\sigma}}\right\}^{\alpha+1}} \tag{8}$$

where θ, $\sigma$ and $\alpha$ are the location, scale and shape parameters respectively such that $\alpha>0$, $\sigma>0$ and $-\infty<x<+\infty$. The first moment of the GL is $E(X) = \theta + \sigma(\Psi(\alpha)—\Psi(1))$ where $\Psi(1) \cong -0.57721$. The second moment of the GL distribution is $Var(X) = \sigma^2(\pi^2/6 + \Psi'(\alpha)$. The GL distribution was selected due to its flexibility modelling data with a range of different shapes under a single framework.

## Materials and methods

### Data sets

English and Welsh data on the number of daily COVID-19 cases and English deaths were obtained from Public Health England (PHE) (https://coronavirus.data.gov.uk/) for lower tier local authorities (LTLAs). Wales has a different methodological approach in collecting death data, and, therefore, we excluded it in any of the death analyses within this study. Meanwhile, English death statistics in this study are people who had a positive test result for COVID-19 and die within 28 days. COVID data are available in a range of time and spatial scales from both PHE and the UK Office of National Statistics (ONS). Data from PHE was available at middle super output area (7,210 regions) and lower tier local authorities (337 regions). MSOA data is updated weekly whilst LTLA data was updated daily. ONS does surveys of prevalence, however these, like MSOA data are not provided daily. We selected the daily data using LTLAs to define boundaries as the best compromise between temporal and spatial coverage as well as allowing the most up-to-date coverage. England and Wales population estimates and England 18 age categories (ranging from 0–4 years old to 85+ years old) were based on the 2011 census and regional land areas were obtained from NOMIS (https://www.nomisweb.co.uk), a database service run by the University of Durham on behalf of the UK Office for National Statistics. The shape files for LTLAs were obtained from the open geography portal (http://geoportal.statistics.gov.uk) provided by the UK Office for National Statistics and UK Data Service (https://census.ukdataservice.ac.uk). The shapefiles are available under UK open government licence v3 (https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/). LTLAs for COVID-19 cases (in England and Wales), COVID-19 mortality (England alone), population, and area were aligned in a daily time series covering the period from 01/03/2020 to 20/05/2021. All data in this study are publicly available under Crown Copyright.

The data are provided by PHE and were downloaded and formatted using R version (3.6.2). We considered regions reporting zero cases or deaths as NULL. Over the period of study PHE data provision has varied such that sometimes NULL returns were absent from reports and at other times set to 0. There has been discussion of issues associated with treatment of zeroes in the literature [26]. The population in the LTLAs for City of London (a small 289-hectare region within the greater London metropolitan area with a small resident population) and Isles of Scilly were considered small and therefore PHE combined these regions with Hackney and Cornwall, respectively. The data are limited by the conditions in place at the time they were reported. Specifically, the availability of tests was limited in the earliest period and changed greatly over the period and deaths reported by PHE were restricted to those occurring within 28 days of a positive test. The limitations created by PHE disclosure control, null data, combined regions and variations in testing are inherent in the data set.

### Statistical analysis

The data were analysed using the statistical software R version (3.6.2) [27] with the sf (0.9–1) [28], raster (3.0–12) [29], dplyr (0.8.5) [30], spData (0.3.5) [31], tmap (2.3–2) [32], ggplot2 (3.3.0) [33–36], xlsx (0.5.7) [37], gplots (3.0.4) [38], httr (1.4.2) [39], plyr (1.8.5) [40], png (0.1–7) [41], rgdal (1.5–19) [42], rgeos (0.5–5) [43], lubridate (1.7.9.2) [44], fitdistrplus (1.1–3) [45], fgarch (3042.83.2) [46], glogis (1.0–1) [47], segmented (1.3–1) [48], moments (0.14) [49], nortest (1.0–4) [50], proxy (0.4–24) [51], RColorBrewer (1.1–2) [52], psych (2.0.12) [53], car (3.0–10) and plotrix (3.7–8) [54] packages.

## Results and discussion

### Overview of regions, cases, and number of observations

England and Wales have 337 LTLAs (315 English LTLAs and 22 Welsch LTLAs) which range in area from 1213 ha (Kensington and Chelsea) up to 518,037 ha (Powys) and have populations between 37,340 (Rutland) up to 1,070,912 (Birmingham). Population densities vary from 0.25 people per hectare (p/ha) (Eden) to 138 p/ha (Islington). Not all LTLAs reported cases or deaths on each day within the period leading to variability in observations (Fig 1). This largely tracked the general progress of the pandemic with the summer months showing the fewest cases, deaths and observations. Histograms of per capita cases (Fig 2) exhibited variable shapes over the course of the pandemic with some periods showing negative skew (Fig 2(a)) while at others they were positively skewed (Fig 2(b)). The availability of testing varied widely over the 15 months which may be a confounder in some presentations; however, the daily scaling metrics, variance, and skewness will reflect the processes in place on the day and were not obviously aligned with testing or the number of observations. All daily per capita case histograms can be found in S1 Fig in the supplementary material.

### Daily progression of COVID-19

To test for scaling behaviour and to correct for the known bias of *per capita* measures daily scaling plots (Fig 3a–3d, S2 and S3 Figs) were constructed and found to be consistent with

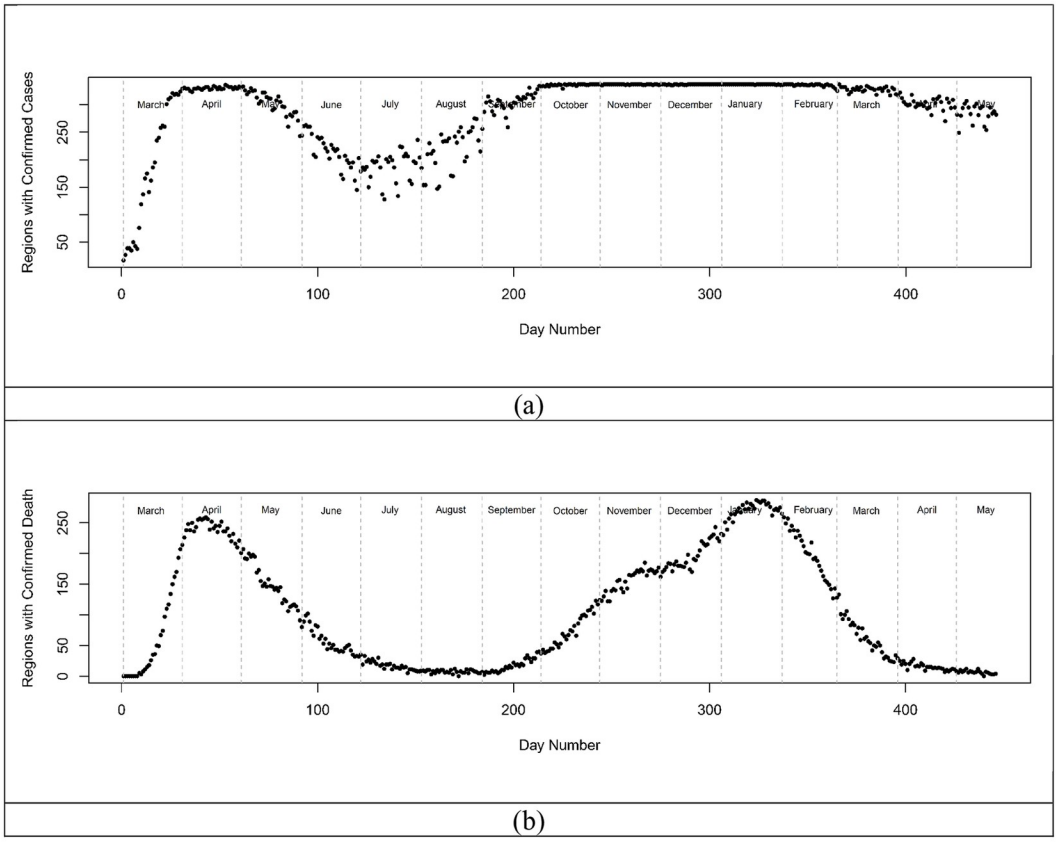



**Fig 1. Time series indicating the number of LTLAs returning cases (a) or deaths (b) over the period of study.**

https://doi.org/10.1371/journal.pone.0261725.g001

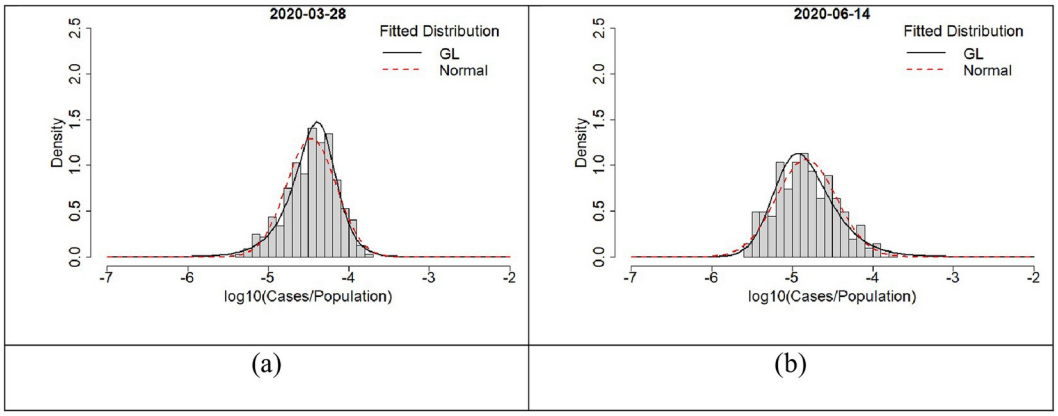



**Fig 2. Histograms of per capita cases in English and Welsh LTLAs.** Some periods within the time series showed negative skew (a) while others were positively skewed (b).

https://doi.org/10.1371/journal.pone.0261725.g002

single power-law models throughout the pandemic. The daily residuals obtained were used as scale adjusted metrics to create geomaps (Fig 3e–3h, S4 and S5 Fig). Residuals are more useful metrics that could be used to assist local interventions.

In the scaling plots (Fig 3a–3d), variability in residual variance was clear by inspection. For example, toward the end of the December holiday period (25/12/2020; Fig 3c) the data were closer to the power law than in September (16/9/2020; Fig 3b). The low variance periods represent a more homogenous presentation of cases across the regions while the higher variance periods were indicative of more heterogeneous regional cases. All daily scaling plots and corresponding geomaps can be found in S2–S5 Figs provided in the supplementary material.

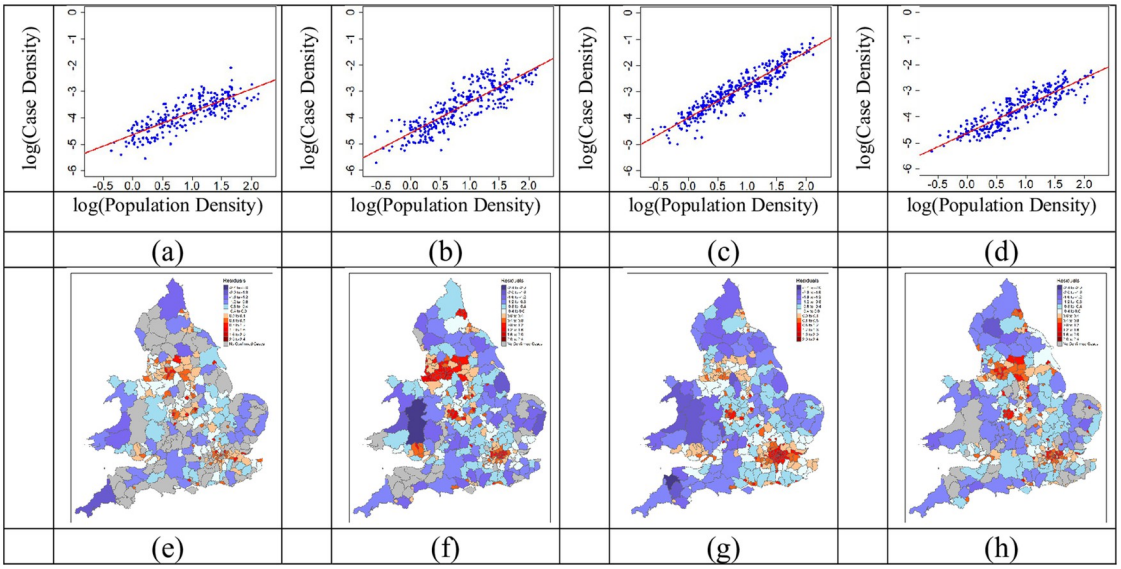



**Fig 3. Scaling plots and geoplots at different times during the pandemic.** These are recorded on the (a and e) 08/06/2020. (b and f) 16/09/2020. (c and g) 25/12/2020. (d and h) 04/04/2021. Regions that are red are above expectation and blue is below. The darker the shade the further from the scaling law. The geomaps contain public sector information licensed under the Open Government Licence v3.0.

https://doi.org/10.1371/journal.pone.0261725.g003

## Daily exponent, variance, and skewness for cases

The LTLA data were examined to assess the trajectory of scaling exponents ($\beta$), residual variance, and skewness over 15 months of the pandemic for cases (Fig 4). If $\beta < 1$, scaling behaviour is sub-linear and less population dense (rural) regions were more affected. If $\beta = 1$, the scaling is linear and rural and urban regions were proportionately affected. Finally, if $\beta > 1$, the

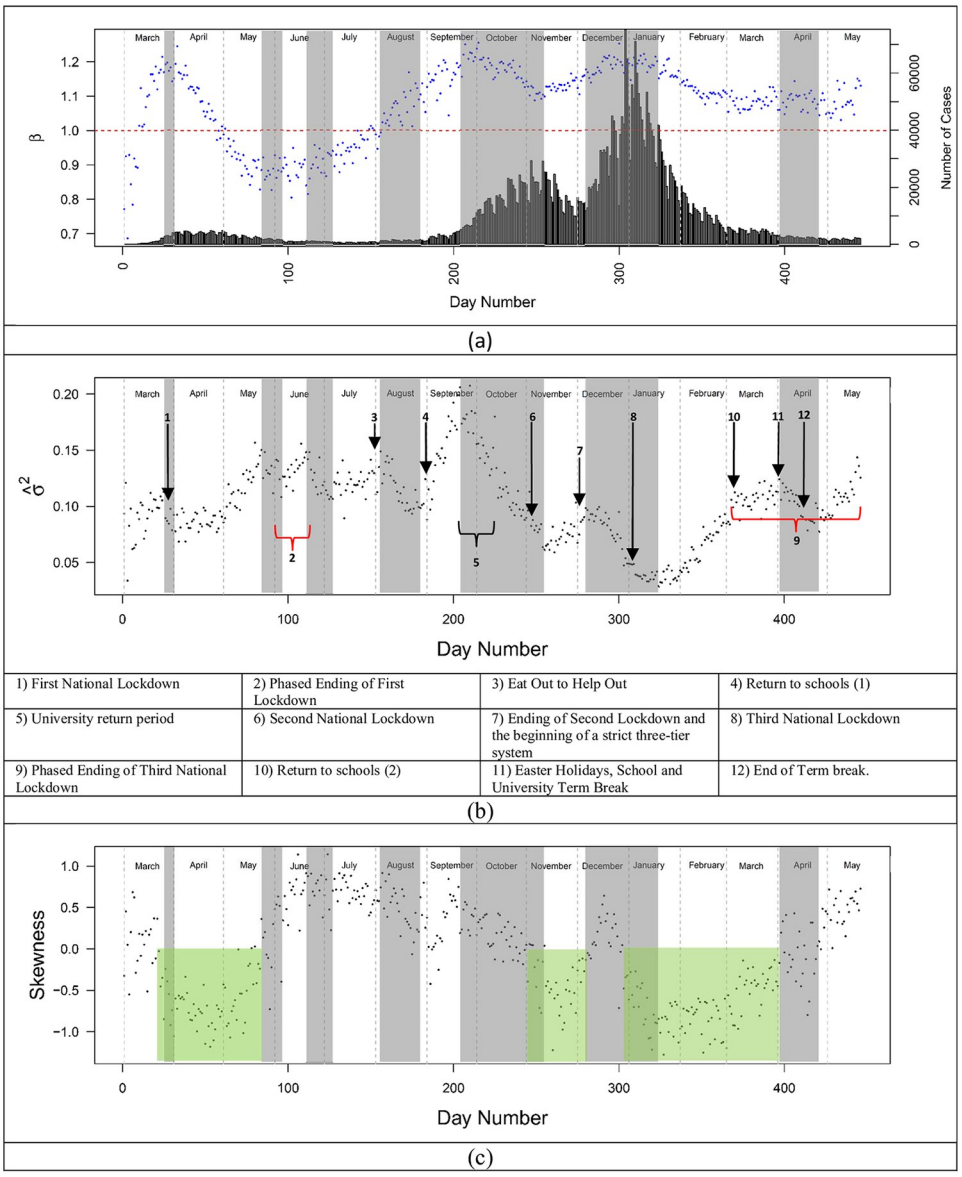



**Fig 4. Daily time series of scaling exponent and residual variance and skewness for cases between 01/03/2020 and 11/01/2021.** (a) Time series of daily scaling exponent of COVID-19 cases, (b) residual variance, and (c) residual skewness. The horizontal line in (a) indicates linear scaling. The bar chart indicates raw daily cases. The grey shading indicates periods of homogenisation. The green shaded periods in (c) correspond to negatively skewed residuals. Those regions coincide with periods of time where isolation dominate during the three national lockdowns. The remaining times were dominated by spreading (positively skewed residuals). Arrows indicate key dates/time periods and red curly brackets represent phased endings to lockdowns. The national restrictions in Wales preceded England beginning on 20/10/2020.

https://doi.org/10.1371/journal.pone.0261725.g004

scaling is super-linear and cases accelerated with population density. The scaling exponents (Fig 4a) for cases rose quickly reaching a peak near the beginning of the first lockdown (announced on the 23/03/2020) in England and Wales and declined gradually until restrictions were eased toward the end of May and early June. Although peaks in cases occurred when $\beta > 1$, super-linear scaling was not universal and the preference for cases in rural vs. urban regions reversed ($\beta$ crossing 1) three times during the period of study: early-March, late April, and the end of July 2020. Overall, during the period of study $\beta$ varied from a low near 0.7 to a high near 1.25 indicating that population density was not a simple proxy for infectious interactions.

Residual variance (Fig 4b) changed by over a factor of 4 during the 15-month period and presented a contrast to the scaling parameters. Variance remained relatively constant until late April and the later stages of the first lockdown when it increased–indicating greater regional heterogeneity. Restrictions tended to increase variance and regional heterogeneity while released restrictions tended to homogenize and reduce the distance of individual regions to the scaling law. For example, in summer 2020, regional lockdowns (Leicester and greater Manchester) in late June were followed by increasing variance. Similarly, following a short lag the January 2021 lockdown was followed by over two months of increasing variance. Release of restrictions after the second national lockdown was followed by an extended contraction of the variance. The re-opening of schools appears to be an exception to the homogenization seen in less restrictive periods of time. When schools opened in early September 2020 residual variance doubled in approximately 2 weeks. We ascribe this to cases in schools primarily having an intra-regional impact rather than leading to rapid inter-regional spreading. Although an immediate "surge" in cases was not seen, a continuation of a consistent increase in cases that began in August was observed. The increasing variance indicates heterogeneous propagation that continued until mid-September when the trend reversed until the beginning of the second national lockdown in November (05/11/2020). These observations are consistent with previous studies of COVID restrictions on mobility showing regionally heterogenous impact and reduced inter-regional interaction and movement [55,56].

The period of declining variance and homogenisation coincides with students returning to universities. There are approximately 2.4 million students studying at universities in the UK. University teaching terms have staggered start dates from the last weeks of September through the first weeks in October. These typically follow a week of orientation and social activities. In advance of orientation and the start of teaching, many students travel with their families from all parts of the UK along with a large number of students who arrive from abroad. This process changed the dynamics of propagation in England and Wales during this time. While there may be other explanations than universities re-opening, there are no other obvious country scale policy changes or processes during this time window.

Homogenisation also occurred following the release of the national restrictions (03/12/2020) and the reopening of businesses in the second national lockdown (12/04/2021). Notably, only the abrupt release of the national restrictions is associated with an obvious "surge" in cases. This includes the major holidays of Christmas and New Year's. Neither caused a "surge." They continued the propagation of the disease in a way that was consistent before and after these key dates. The general country scale homogenisation between the LTLA regions drove residual variance to the lowest levels seen over the 15-month period.

Skewness provides a further contrast to case counts, scaling exponents, and variance. We used the scaling law residuals to create a time series of skewness metrics (Fig 4c). Similar behaviour was seen in the *per capita* case distributions (Fig 2) with characteristics changing over the course of the 15-month period. When cases follow a distribution with a strong positive skew, the long positive tail of the skewed distribution is indicative of propagation with hot

spots and potential super-spreading incidents. Conversely, when the residuals are negatively skewed, this indicates a distribution better characterised by a long tail of "cold spots" or super-isolated regions.

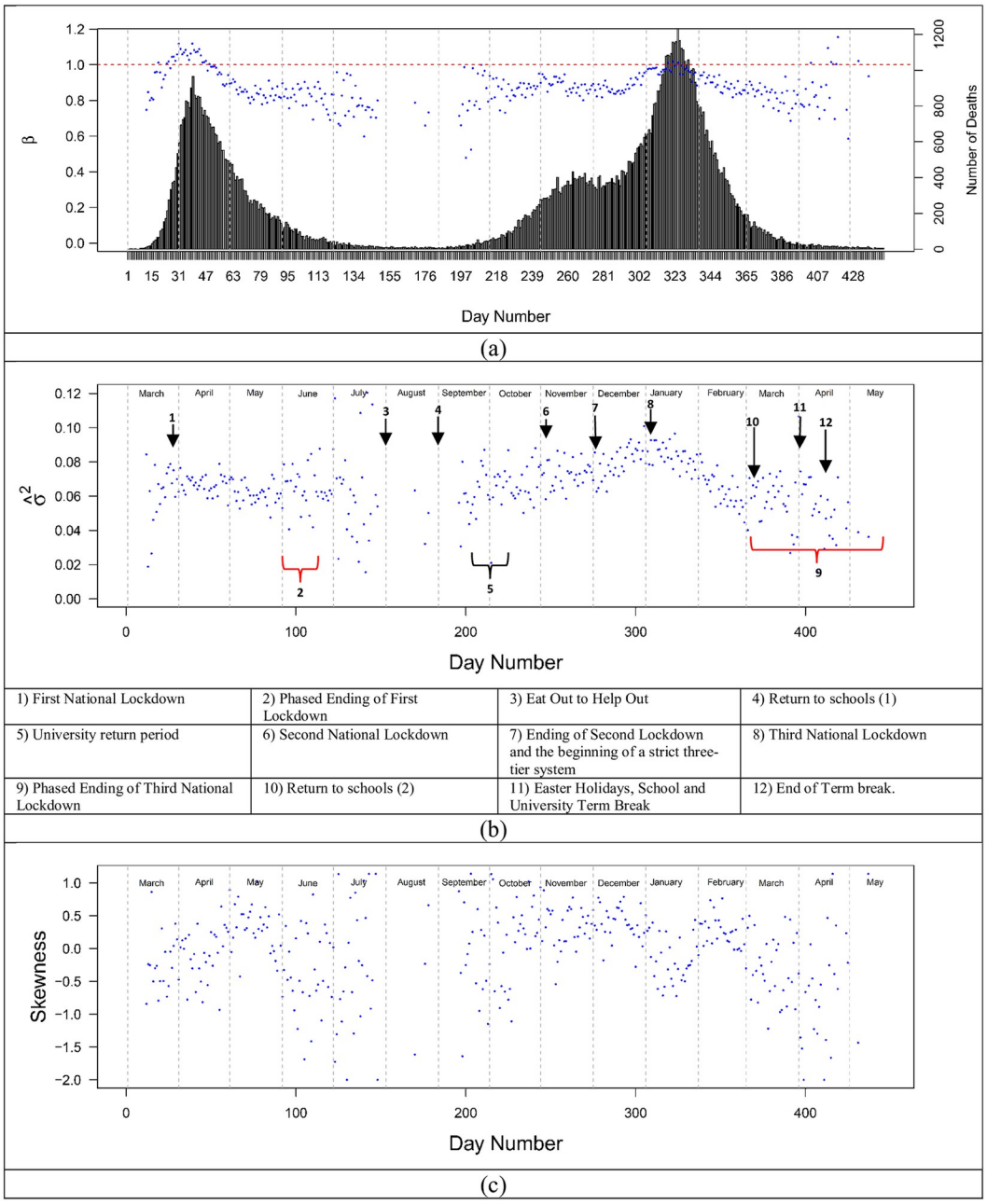
## Daily exponent, variance, and skewness for deaths

In contrast, daily exponents, variance, and skewness for COVID-19 deaths (Fig 5) were consistent and remained at a similar level throughout the pandemic. The analysis was restricted to days with 10 or more regions reporting deaths. For a short time at the beginning of the time series, regions exhibited $\beta > 1$ (super-linear scaling with greater urban impact), but this inverted such that $\beta < 1$ (sub-linear scaling with rural impact) circa 10/04/2020 where it remained with the exception of ~10 days scattered in January, April, and May 2021. In other words, COVID-19 deaths showed economies of scale with increasing population density; thus, for most of the pandemic rural regions were preferentially affected by deaths (e.g. $\beta < 1$). Variance and skewness from deaths had very little structure in comparison to cases and exhibited comparatively homogeneous behaviour throughout England. This was in stark contrast with the far greater structure in cases. This behaviour is consistent with the age demographics in England and Wales. Previous work has documented that populations dense regions serve as a magnet for young people while rural regions tend to have a greater proportion of elderly people [19]. The scaling exponents for deaths throughout are consistent with those seen for scaling of people 60 and above in England and Wales. This is overwhelmingly the demographic most likely to die from COVID-19.

## Age demographics

To explore age demographics further and its key part in the consistency of the scaling exponent observed in Fig 5 we included 18 age groups ranging from 0–4 years to 85+ years old and aligned this with regional boundaries defined in the death data (315 regions). Similar to previous work [19], the density scaling models gave reasonable fit to power laws (Fig 6a and S7 Fig). Young and middle age groups are fitted using a single power law fit and all other age groups are fitted using a double power-law fit. A segmented relationship (Fig 6a) indicates that certain age groups either accelerate or decline in urbanised regions. People aged 24–44 accelerate in high density regions whilst people aged 60 and over preferentially leave. Spearman correlation of residuals followed by hierarchical clustering (Fig 6b) shows that age categories break up into two main clusters separating younger people (0–49 years old) and older people (aged 50+). Between clusters there is mostly anti-correlation with some Spearman's rank correlation coefficient values reaching as strong as -0.64 (Aged 30–34 vs. Aged 75–79 (Fig 6c) and Aged 30–34 vs. Aged 80–84 (Fig 6d). On the other hand, positive correlation, mostly occurring within each cluster, have a Spearman's rank correlation coefficient reaching as strong as 0.94 (Aged 70–74 vs. Aged 75–79 (Fig 6e). Negative correlation between clusters means that regions where the density of older people exceeds that predicted by the scaling laws have fewer younger people and vice versa. A positive correlation within a cluster means that as one age group exceeds the expected density, so does the other. All Spearman's rank correlation coefficient values between all age categories are provided in S8 Fig.

When considering cumulative deaths over the study period, a segmented relationship is observed suggesting of a protective effect of high population density (Fig 7a). This is an artefact of age demographics. Restricting the population density to older age groups reduces the segmented relationship, and in some cases removes it completely. For example, when considering the aged 80–84 group the reduction and protection of COVID-19 death in urbanised regions seen for total population is no longer present (Fig 7) and a single power law model is now the preferred model.




| | | | |
|---|---|---|---|
| 1) First National Lockdown | 2) Phased Ending of First Lockdown | 3) Eat Out to Help Out | 4) Return to schools (1) |
| 5) University return period | 6) Second National Lockdown | 7) Ending of Second Lockdown and the beginning of a strict three-tier system | 8) Third National Lockdown |
| 9) Phased Ending of Third National Lockdown | 10) Return to schools (2) | 11) Easter Holidays, School and University Term Break | 12) End of Term break. |

**Fig 5. Daily time series of scaling exponent and residual variance and skewness for deaths 01/03/2020 and 11/01/2021.** (a) Time series of daily scaling exponent of COVID-19 deaths, (b) residual variance, and (c) residual skewness. The horizontal line in (a) indicates linear scaling. The bar chart indicates raw daily deaths. Arrows in (b) indicate key dates/time periods and red curly brackets represent phased endings to lockdowns. The second lockdown in Wales preceded England beginning on 20/10/2020.

https://doi.org/10.1371/journal.pone.0261725.g005

## Dispersion of COVID-19 case residuals over time

To better understand the distribution of residuals, we investigate the normal and generalized logistic distributions as candidate distributions using the LTLA data (Eq 5). The normal distribution is symmetric and has no skew. The GL distribution has three parameters which can

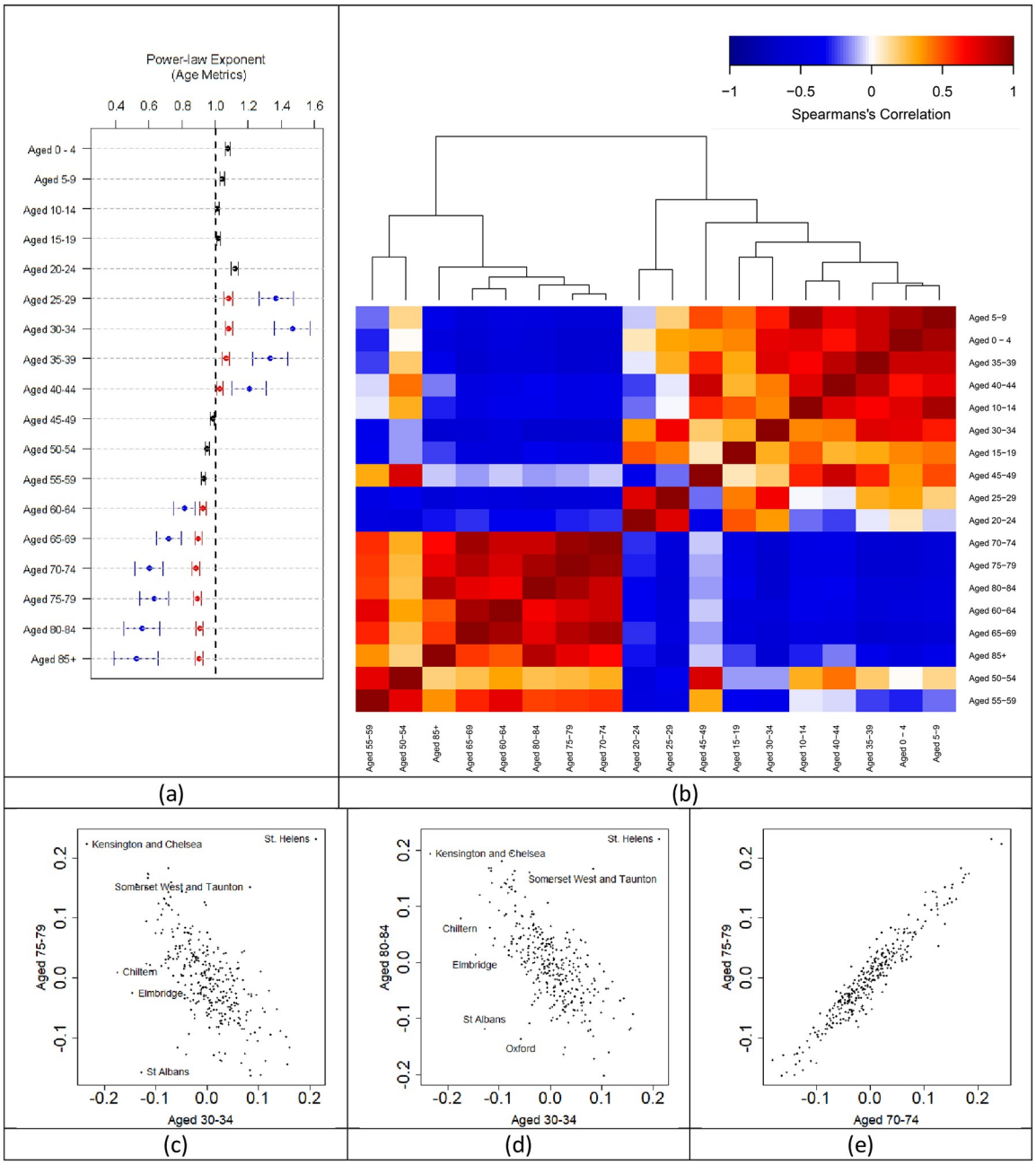



**Fig 6. Power law scaling exponents, hierarchical clustering and correlation for 18 categories of age.** (a) Black symbols indicate exponents for single power-law scaling. Red symbols indicate exponents below the critical density and blue symbols are for exponents above the critical density. Error bars represent the 95% confidence intervals for all exponents based on the standard errors of regression. The black dotted line represents linear scaling. (b) The colours in the heatmap refer to the strength of the correlation between residuals by evaluating the Spearman correlation coefficient. The red indicates positive correlation and blue indicates negative correlation. The darker the shade of red and blue signifies the strength of the correlation. Examples of residual relationships displayed in the heatmap include (c) Aged 75–79 vs. Aged 30–34, (d) Aged 80–84 vs. Aged 30–34 and (e) Aged 75–79 vs. Aged 70–74.

https://doi.org/10.1371/journal.pone.0261725.g006

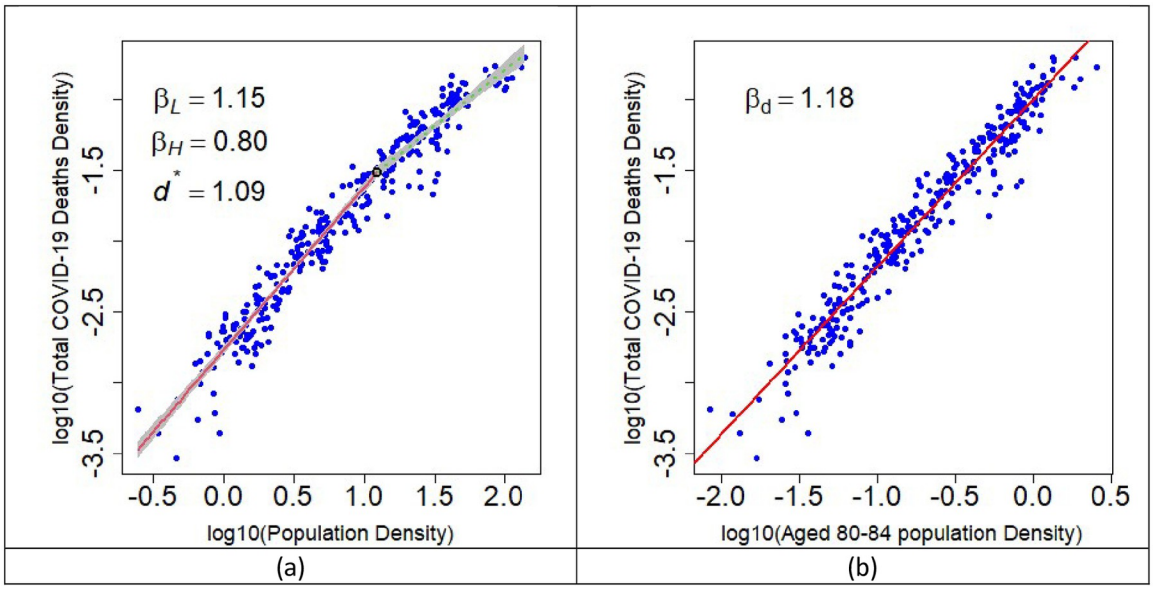



**Fig 7. Scaling relationship for COVID-19 death when looking at (a) Total Population and (b) restricted to the 80–84 age group.** In (a) a segmented relationship is better fitted where the black circle is the identified critical density. The red line is below and the green line is above the critical density. In (b) the segmented relationship disappears, and a single power-law is the preferred model. The decline in COVID-19 in urbanised regions is no longer viable when age range is restricted.

https://doi.org/10.1371/journal.pone.0261725.g007

accommodate a wider range of shapes including positive and negative skewing. When comparing normal and GL distributions as models for scaling law residuals the additional parameter needs to be accounted for. We used the Akaike (AIC) and Bayesian (BIC) information criteria to decide if normal or GL represented a better model for each day in the 10-month period. When selecting a model, lower AIC and BIC scores represent better fits. The differences between AIC and BIC scores obtained from fitting the two distributions to the residuals were for each day in the 15-month period (Fig 8). Positive values correspond to a generalised logistic distribution as the preferred model, whilst a negative value corresponds to a normal distribution as the preferred model. All daily histograms for cases and deaths can be found in S9 and S10 Figs in the supplementary material.

Although there is some noise in the differences, the contrast between cases and deaths is again clear. During the initial periods of the lockdowns (March, November and January) propagation of cases was associated with a GL distribution and negative skew whilst during less restrictive time frames (August, September, October and April) propagation is associated with a normal distribution. A variety of authors have noted the fat tails and/or positive skewing in of super-spreading events [6–10]. At LTLA scale, the number and size of individual spreading events place determine its position in a distribution. With a sufficient number of events, it will become a hotspot and appear on the extreme positive side of a positively skewed residual distribution. At times during the pandemic in England and Wales this was observed but was insufficient for the full period. As an example, modelling and simulation of propagation [7] using network science and a gamma distribution was attempted using varying parameter values to represent different proportions of "super-spreaders." This analysis indicated that the initial trajectory of exposed and infected people in a population accelerates quickly in networks where there are a high proportion of super-spreaders. However, a gamma distribution cannot be negatively skewed and the daily cases in the data here contain periods of negative skewing.

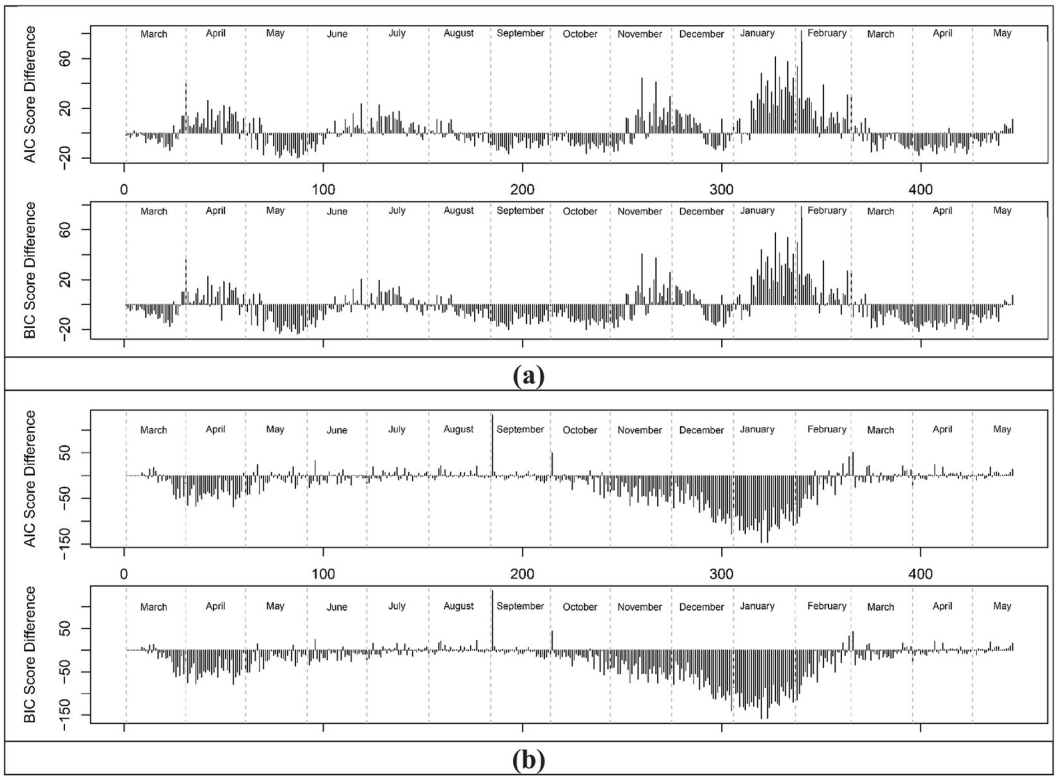



**Fig 8. AIC and BIC differences over time.** (a) COVID-19 cases and (b) COVID-19 death. Positive AIC/BIC indicates GL is a better fit and a negative AIC/BIC indicates normal is a better fit.

https://doi.org/10.1371/journal.pone.0261725.g008

"Super-spreading" events creating "hot-spots" are certainly important but the converse concepts of "super-isolation" and "cold-spots" better describe regions at the low end of a negatively skewed distribution and this phenomenon needs to be better appreciated and understood. Position relative to expectation was remarkably persistent (S6 Fig) and understanding the features of regions where a disease is not spreading or is consistently below expectation is needed. The contrasting behaviour of deaths is also of interest. Residuals more consistent with normally distributed residuals were the overwhelming feature of regional deaths during the 15-month period.

## Regional persistence of COVID-19 case residuals

To investigate the persistence of regional behaviour, the correlation between residuals was computed for all pairs of days and presented as a heatmap (Fig 9). This indicated the position of a particular region relative to the scaling laws was very persistent after the first 5 (Pearson correlation (Fig 9) to 9 days (Spearman rank correlation (S6 Fig). The near universal dark red appearance of the heatmap indicates that a region that was high relative to the population density scaling law early in the pandemic tended to remain there. While such correlation between close dates is to be expected (Fig 9c), persistence for 300–350 days (Fig 9a and 9b) or longer is remarkable. This persistence which survived 3 national lockdowns, multiple locally targeted measures, and an enormous expansion of testing needs explanation. In previous studies of the inter-relationships between indicators of health, wealth, well-being and age [7], we noted that mortality health outcomes are related to these other factors in complex ways. Based on this

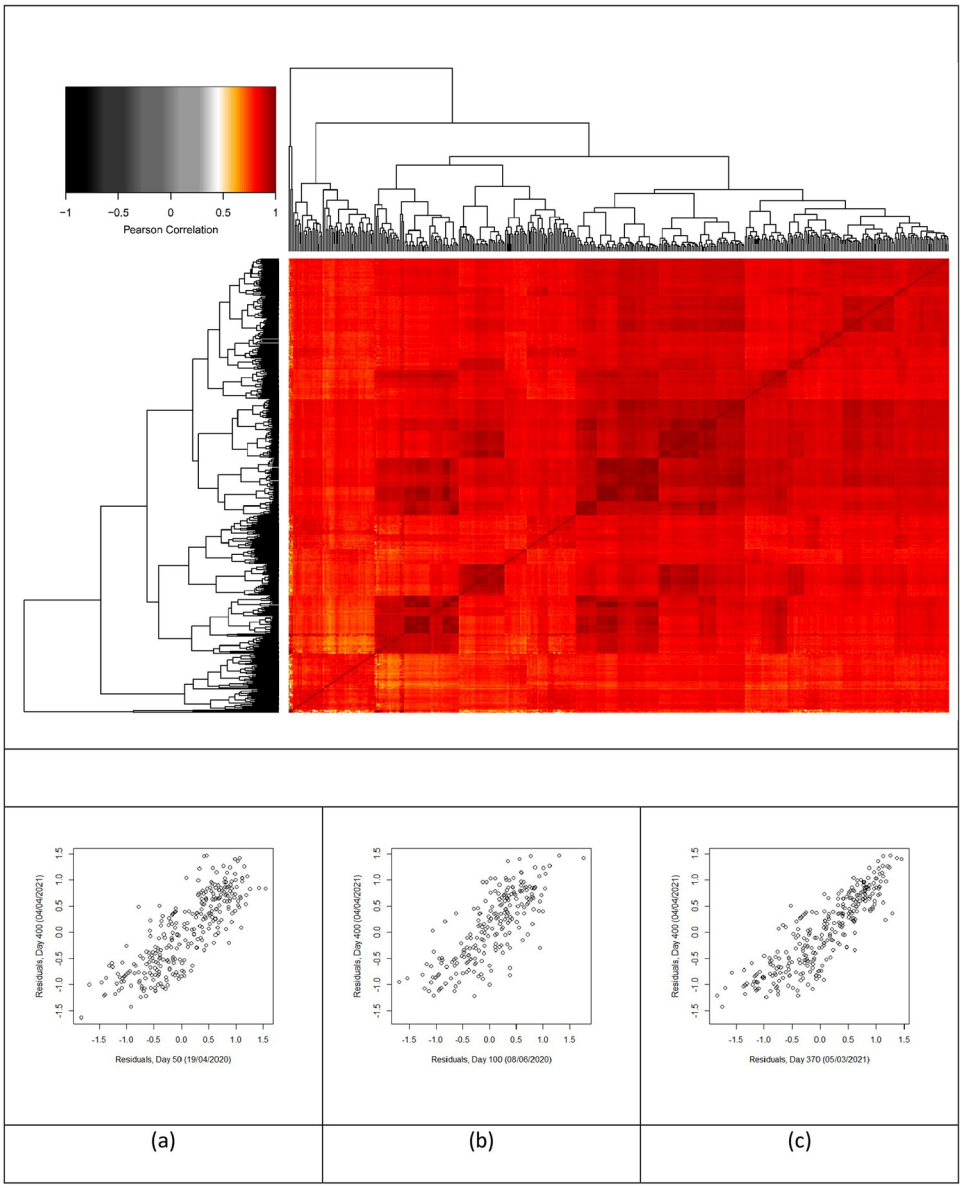



**Fig 9. Heatmap of Pearson's correlation coefficient with example paired dated.** Correlation of COVID-19 case residuals between all pairs of dates between 01/03/2020 and 20/05/2021 (main panel). Red indicates strong positive correlation. White and grayscale indicate low and negative correlation. The darker shade of colour is associated with a higher similarity between the two pairs. Example data sets where residuals from days 50 (a), 100 (b), and 370 (c) are correlated with those from day 400.

https://doi.org/10.1371/journal.pone.0261725.g009

earlier work, the socio-economic characteristics leading to a region's position relative to the scaling laws may have been in place before the pandemic and will remain beyond it. Further work is needed to test this hypothesis directly.

## Conclusions

This study has established that both regional *per capita* measures and scaling law residuals exhibit both positive and negative skewing. Positively skewed distributions have been widely

observed and used to model pandemic behaviour [7,12]. Such behaviour is important to indicate super-spreading and hot-spots, but insufficient to characterise the full sweep of the pandemic through multiple interventions.

Similarly, scaling law parameters are often thought to be constant or very slowly changing features of a process. In the case of COVID-19 cases, scaling parameters evolved over relatively short periods of time. For cases, the scaling law exponents reached a peak at the beginning of the first lockdown and gradually declined for approximately three months. Preferential propagation of COVID-19 cases switched between rural and urban regions several times during the first 150 days of the pandemic. Since then, urban regions have driven cases. COVID-19 mortality gave a more consistent picture of low population density regions preferentially and consistently affected with linear scaling only being approached during two periods in the 15-months studied. Those corresponded with the months of peak death in April 2020 and January 2021. The preferential impact of death on rural (low population density) regions was found to be due to the greater proportions of elderly people in these regions. Although this study used data from the ongoing pandemic, we are unaware of any scaling study in terms of the number of time points nor the documentation of such variability in $\beta$ in any field.

Variance relative to the population density scaling laws is a key descriptor of the distribution of regional cases. Lockdowns produce heterogeneity (higher variance) across regions while reducing cases. The re-opening of schools drove heterogeneity during a period of case growth indicative of locally important outbreaks. Country scale mixing such as occurred with the opening of universities and holiday periods promotes homogenisation (low variance). All key statistical metrics from regional death data were remarkably different from cases in the time period. This is consistent with regional age demographics in England and Wales. From a policy point of view these observations and patterns are particularly important, as they provide insight and expected indicative effects following implementation of health policies and allow better delivery of resources to areas needing them most acutely. In this case, it was rural communities.

Within this framework it is important to note that for the full 15 month period England and Wales had continuous community spread of SARS-COV-2. Excepting the very early period in March, there has been nothing that could be called a "surge." Within the 15-month period, the rise and fall of cases and deaths have been gradual as has the evolution of scaling metrics, variance structures, and distribution shapes.

Finally, regional behaviour relative to population density scaling laws was remarkably persistent throughout the pandemic. It is possible that the determinants of regional behaviour existed pre-pandemic and although government interventions have had an unambiguous impact on the rise and fall of cases and death, they have had little impact on whether a particular region is high or low in relative to nationwide population density scaling. Although residuals may appear randomly distributed around the power law, we find that this is not true. They are extensively correlated and reveal persistent structure.

## Supporting information

**S1 Fig. Daily histograms of LTLA COVID-19 cases per capita.** Black line represents the generalised logistic distribution and the red dashed line represents the normal distribution.
(PDF)

**S2 Fig. Daily LTLA Density scaling behaviour of COVID-19 cases (i.e. log(Case Density vs. log(Population Density)).** The blue dots are the empirical values. A red line represents the single exponent power-law fit.
(PDF)

**S3 Fig. Daily LTLA Density scaling plots of COVID-19 death (i.e. log(Death Density vs. log(Population Density)).** The blue dots are the empirical values (England). A red line represents the single exponent power-law fit.
(PDF)

**S4 Fig. Daily geoplots of LTLA COVID-19 case residuals.** Regions that are red are above expectation and blue is below. The darker the shade the further from the scaling law. The geoplots contain public sector information licensed under the Open Government Licence v3.0.
(PDF)

**S5 Fig. Daily geoplots of LTLA COVID-19 death residuals.** Regions that are red are above expectation and blue is below. The darker the shade the further from the scaling law. The geoplots contain public sector information licensed under the Open Government Licence v3.0.
(PDF)

**S6 Fig. Spearman's rank correlation coefficient.** Rank correlation of COVID-19 case residuals between all pairs of dates between 01/03/2020 and 20/05/2021. Red and blue colours refer to correlation values close to 1 and -1 respectively. The darker shade of colour is associated to a higher similarity between the two pairs.
(PDF)

**S7 Fig. Single and double power law scaling models identified using a Davies test.** Single power law models are represented with one single red line. Double power law models are represented with a red line, below the critical density and a green line above the critical density. The black circle is the position of the critical density.
(PDF)

**S8 Fig. Residual relationships between all 18 age categories.** The lower triangle are residual scatter plots, the diagonal are density histograms and the upper triangle are the Spearman's rank correlation coefficients.
(PDF)

**S9 Fig. Daily histograms of LTLA COVID-19 case residuals.** Black line represents the generalised logistic distribution and the red dashed line represents the normal distribution.
(PDF)

**S10 Fig. Daily histograms of LTLA COVID-19 case deaths.** Black line represents the generalised logistic distribution and the red dashed line represents the normal distribution.
(PDF)

**S1 Dataset. LTLA regions included in this study. 337** English and Welsch LTLA's that define regions in this study.
(XLSX)

**S2 Dataset. English and Welsch LTLA daily COVID-19 cases employed in this study.** Data covering the period from 01/03/2020 to 20/05/2021.
(XLSX)

**S3 Dataset. English LTLA daily COVID-19 death employed in this study.** Data covering the period from 01/03/2020 to 20/05/2021.
(XLSX)

**S4 Dataset. Total number of cases in England and Wales and deaths in England.** Data covering the period from 01/03/2020 to 20/05/2021.
(XLSX)

**S5 Dataset. Daily COVID-19 case residuals.** England and Wales.
(XLSX)

**S6 Dataset. Daily COVID-19 death residuals.** England.
(XLSX)

**S7 Dataset. 18 age categories.** Data for each region in England.
(XLSX)

**S8 Dataset. Age residuals.** England.
(CSV)

**S1 File.**
(ZIP)

## Acknowledgments

The authors are grateful to the UK Office of National Statistics and Public Health England for making these data available.

## Author Contributions

**Conceptualization:** Jack Sutton, Golnaz Shahtahmassebi, Haroldo V. Ribeiro, Quentin S. Hanley.

**Data curation:** Jack Sutton.

**Formal analysis:** Jack Sutton.

**Methodology:** Jack Sutton, Golnaz Shahtahmassebi, Haroldo V. Ribeiro, Quentin S. Hanley.

**Supervision:** Golnaz Shahtahmassebi, Quentin S. Hanley.

**Visualization:** Jack Sutton.

**Writing – original draft:** Jack Sutton, Golnaz Shahtahmassebi, Haroldo V. Ribeiro, Quentin S. Hanley.

**Writing – review & editing:** Jack Sutton, Golnaz Shahtahmassebi, Haroldo V. Ribeiro, Quentin S. Hanley.

## References

1. Morawska L, Cao J. Airborne transmission of SARS-CoV-2: The world should face the reality. Environ Int. 2020; 139: 105730. https://doi.org/10.1016/j.envint.2020.105730 PMID: 32294574
2. Anderson EL, Turnham P, Griffin JR, Clarke CC. Consideration of the Aerosol Transmission for COVID-19 and Public Health. Risk Analysis. 2020. pp. 902–907. https://doi.org/10.1111/risa.13500 PMID: 32356927
3. Prather KA, Wang CC, Schooley RT. Reducing transmission of SARS-CoV-2. Science (80-). 2020; 368: 1422–1424. https://doi.org/10.1126/science.abc6197 PMID: 32461212
4. Asadi S, Bouvier N, Wexler AS, Ristenpart WD. The coronavirus pandemic and aerosols: Does COVID-19 transmit via expiratory particles? Aerosol Sci Technol. 2020; 54: 635–638. https://doi.org/10.1080/02786826.2020.1749229 PMID: 32308568

5. Schläpfer M, Bettencourt LMA, Grauwin S, Raschke M, Claxton R, Smoreda Z, et al. The scaling of human interactions with city size. J R Soc Interface. 2014; 11: 20130789. https://doi.org/10.1098/rsif.2013.0789 PMID: 24990287

6. Kain MP, Childs ML, Becker AD, Mordecai EA. Chopping the tail: How preventing superspreading can help to maintain COVID-19 control. Epidemics. 2021; 34: 100430. https://doi.org/10.1016/j.epidem.2020.100430 PMID: 33360871

7. Reich O, Shalev G, Kalvari T. Modeling COVID-19 on a network: super-spreaders, testing and containment. medRxiv 2020043020081828. 2020. https://doi.org/10.1101/2020.04.30.20081828

8. Lau MSY, Grenfell B, Thomas M, Bryan M, Nelson K, Lopman B. Characterizing superspreading events and age-specific infectiousness of SARS-CoV-2 transmission in Georgia, USA. Proc Natl Acad Sci U S A. 2020; 117: 22430–22435. https://doi.org/10.1073/pnas.2011802117 PMID: 32820074

9. Fukui M, Furukawa C. Power Laws in Superspreading Events. medRxiv. 2020; 1–41. https://doi.org/https%3A//doi.org/10.1101/2020.10.21.20216895

10. Wong F, Collins JJ. Evidence that coronavirus superspreading is fat-tailed. Proc Natl Acad Sci U S A. 2020; 117: 29416–29418. https://doi.org/10.1073/pnas.2018490117 PMID: 33139561

11. Moreau VH. Forecast predictions for the COVID-19 pandemic in Brazil by statistical modeling using the Weibull distribution for daily new cases and deaths. Brazilian J Microbiol. 2020; 51: 1109–1115. https://doi.org/10.1007/s42770-020-00331-z PMID: 32809115

12. Roques L, Bonnefon O, Baudrot V, Soubeyrand S, Berestycki H. A parsimonious approach for spatial transmission and heterogeneity in the COVID-19 propagation: Modelling the COVID-19 propagation. R Soc Open Sci. 2020; 7. https://doi.org/10.1098/rsos.201382 PMID: 33489282

13. Li L, Yang Z, Dang Z, Meng C, Huang J, Meng H, et al. Propagation analysis and prediction of the COVID-19. Infect Dis Model. 2020; 5: 282–292. https://doi.org/10.1016/j.idm.2020.03.002 PMID: 32292868

14. Stier AJ, Berman MG, Bettencourt LMA. COVID-19 attack rate increases with city size. medRxiv. 2020; 1–23. https://doi.org/10.1101/2020.03.22.20041004

15. Ascani A, Faggian A, Montresor S. The geography of COVID-19 and the structure of local economies: The case of Italy. J Reg Sci. 2020; 1–35. https://doi.org/10.1111/jors.12510 PMID: 33362296

16. Cardoso B-HF, Gonçalves S. Universal scaling law for human-to-human transmission diseases. EPL (Europhysics Lett. 2021; 133: 58001.

17. Ribeiro H V., Sunahara AS, Sutton J, Perc M, Hanley QS. City size and the spreading of COVID-19 in Brazil. Jiang L-L, editor. PLoS One. 2020; 15: e0239699. https://doi.org/10.1371/journal.pone.0239699 PMID: 32966344

18. Ribeiro H V., Rybski D, Kropp JP. Effects of changing population or density on urban carbon dioxide emissions. Nat Commun. 2019; 10: 3204. https://doi.org/10.1038/s41467-019-11184-y PMID: 31324796

19. Sutton J, Shahtahmassebi G, Ribeiro H V., Hanley QS. Rural–urban scaling of age, mortality, crime and property reveals a loss of expected self-similar behaviour. Sci Rep. 2020; 10: 16863. https://doi.org/10.1038/s41598-020-74015-x PMID: 33033349

20. Ribeiro H V., Hanley QS, Lewis D. Unveiling relationships between crime and property in England and Wales via density scale-adjusted metrics and network tools. PLoS One. 2018; 13: e0192931. https://doi.org/10.1371/journal.pone.0192931 PMID: 29470499

21. Hanley QS, Lewis D, Ribeiro H V. Rural to urban population density scaling of crime and property transactions in english and welsh parliamentary constituencies. PLoS One. 2016; 11: e0149546. https://doi.org/10.1371/journal.pone.0149546 PMID: 26886219

22. Bettencourt LMA, Lobo J, Helbing D, Kühnert C, West GB. Growth, innovation, scaling, and the pace of life in cities. Proc Natl Acad Sci. 2007; 104: 7301–7306. https://doi.org/10.1073/pnas.0610172104 PMID: 17438298

23. Bokányi E, Szállási Z, Vattay G. Universal scaling laws in metro area election results. Braha D, editor. PLoS One. 2018; 13: e0192913. https://doi.org/10.1371/journal.pone.0192913 PMID: 29470518

24. Alves LGA, Ribeiro HV, Lenzi EK, Mendes RS. Distance to the scaling law: a useful approach for unveiling relationships between crime and urban metrics. PLoS One. 2013; 8: e69580. https://doi.org/10.1371/journal.pone.0069580 PMID: 23940525

25. Bettencourt LMA, Lobo J, Strumsky D, West GB. Urban scaling and its deviations: Revealing the structure of wealth, innovation and crime across cities. PLoS One. 2010; 5: e13541. https://doi.org/10.1371/journal.pone.0013541 PMID: 21085659

26. Finance O, Cottineau C. Are the absent always wrong? Dealing with zero values in urban scaling. Environ Plan B Urban Anal City Sci. 2018; 0: 1–15. https://doi.org/10.1177/2399808318785634

27. Team R. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2019. https://www.r-project.org/.

28. Pebesma E. Simple Features for R: Standardized Support for Spatial Vector Data. R J. 2018; 10: 439–446. https://doi.org/10.32614/RJ-2018-009

29. Hijmans RJ. raster: Geographic Data Analysis and Modeling. 2020. https://cran.r-project.org/package=raster.

30. Wickham H, François R, Henry L, Müller K. dplyr: A Grammar of Data Manipulation. 2019. https://cran.r-project.org/package=dplyr.

31. Bivand R, Nowosad J, Lovelace R. spData: Datasets for Spatial Analysis. 2020. https://cran.r-project.org/package=spData.

32. Tennekes M. tmap: Thematic Maps in R. J Stat Softw. 2018; 84: 1–39. https://doi.org/10.18637/jss.v084.i01 PMID: 30450020

33. Lovelace R, Cheshire J. Introduction to visualising spatial data in R. Natl Cent Res Methods Work Pap 08/14. 2017. https://doi.org/10.5281/zenodo.889551

34. Lovelace R, Cheshire J. Introduction to visualising spatial data in R Part I: Introduction. Tutorial. 2015. https://doi.org/10.5281/zenodo.889551

35. Cheshire J, Lovelace R. Introduction to Spatial Data and ggplot2. RPub. 2013.

36. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York; 2016. https://ggplot2.tidyverse.org.

37. Dragulescu AA, Arendt C. xlsx: An R package to Read, Write, Format Excel 2007 and Excel 97/2000/XP/2003 Files. 2018.

38. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A, et al. gplots: Various R Programming Tools for Plotting Data. 2020. https://cran.r-project.org/package=gplots.

39. Wickham H. httr: Tools for Working with URLs and HTTP. 2020. https://cran.r-project.org/package=httr.

40. Wickham H. The Split-Apply-Combine Strategy for Data Analysis. J Stat Softw. 2011; 40: 1–29.

41. Urbanek S. png: Read and write PNG images. 2013. https://cran.r-project.org/package=png.

42. Bivand R, Keitt T, Rowlingson B. rgdal: Bindings for the "Geospatial" Data Abstraction Library. 2021.

43. Bivand R, Rundel C. rgeos: Interface to Geometry Engine—Open Source ('GEOS'). 2020. https://cran.r-project.org/package=rgeos.

44. Grolemund G, Wickham H. Dates and Times Made Easy with Lubridate. J Stat Softw. 2011; 40: 1–25.

45. Delignette-Muller ML, Dutang C. An {R} Package for Fitting Distributions. J Stat Softw. 2015; 64: 1–34.

46. Wuertz D, Setz T, Chalabi Y, Boudt C, Chausse P, Miklovac M. Rmetrics—Autoregressive Conditional Heteroskedastic Modelling. 2020. https://cran.r-project.org/src/contrib/Archive/Rmetrics/.

47. Achim Zeileis, Windberger T. glogis: Fitting and Testing Generalized Logistic Distributions. R Packag version 10–1. 2018. https://cran.r-project.org/package=glogis.

48. Muggeo VMR. segmented: an R Package to Fit Regression Models with Broken-Line Relationships. R News. 2008; 8: 20–25.

49. Novomestky F, Komsta L. moments: Moments, cumulants, skewness, kurtosis and related tests. R package version 0.14. 2015. https://cran.r-project.org/package=moments.

50. Gross J, Ligges U. nortest: Tests for Normality. 2015. https://cran.r-project.org/package=nortest.

51. Meyer D, Buchta C. proxy: Distance and Similarity Measures. 2020. https://cran.r-project.org/package=proxy.

52. Neuwirth E. RColorBrewer: ColorBrewer Palettes. 2014. https://cran.r-project.org/package=RColorBrewer.

53. Revelle W. psych: Procedures for Psychological, Psychometric, and Personality Research. Evanston, Illinois; 2020. https://cran.r-project.org/package=psych.

54. Lemon J. plotrix: Various Plotting Functions. 2021. https://cran.r-project.org/package=plotrix.

55. Schlosser F, Maier BF, Jack O, Hinrichs D, Zachariae A, Brockmann D. COVID-19 lockdown induces disease-mitigating structural changes in mobility networks. Proc Natl Acad Sci. 2020; 117: 32883–32890. https://doi.org/10.1073/pnas.2012326117 PMID: 33273120

56. Melo HPM, Henriques J, Carvalho R, Verma T, da Cruz JP, Araújo NAM. Heterogeneous impact of a lockdown on inter-municipality mobility. Phys Rev Res. 2021; 3: 13032.