Allometric and Fluctuation Scaling in Health, Well-Being and Mortality

Jack Sutton

A thesis submitted in partial fulfilment of the requirements of Nottingham Trent University for the degree of Doctor of Philosophy

School of Science and Technology

Department of Physics and Mathematics

Nottingham Trent University

July 2022

Abstract

In the urban scaling hypothesis, it has been noted that cities exhibit self-similar behaviour. Much of the scaling literature focuses on cities, whilst not including rural regions. Initially, rural-urban population density scaling was investigated using a diverse set of indicators (crime, property, mortality and age) in England and Wales. These were fitted using either a single or segmented power-law (PL) model where preference was chosen using a Davies test and confirmed using both Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). In this study, it was found that most indicators exhibited a change point between rural-urban regions typically around 27 people per hectare. Mortality, for example, declined above the change point showing that urban regions have a 'protective' effect influenced by age demographics. Residuals obtained from the preferred PL model were analysed using methods such as hierarchical clustering and self-organising maps (SOM) displaying regional clusters, strong extensive correlations and disparities. The most interesting finding was that age demographics break the self-similarity behaviour that is a fundamental and underlying part of the urban scaling hypothesis. Scaling is usually cumulatively data over a large timeframe. Finer granularity of data is not easily accessible, although the COVID-19 pandemic was a unique opportunity to obtain and explore the scaling of daily data. It is thought that scaling exponent is slow changing and exhibited little fluctuation. However, it was found that COVID-19 cases revealed that the scaling exponents along with residual variance and skew varied with considerable complexity. Scaling exponents continually evolved and reversed where preference of propagation between rural-urban regions switched 6 times. Regional homogeneity occurred in periods with low variance where regions are located close to the PL. Contrary, regional heterogeneity occurred in periods with high variance where regions are located further away from the PL. Skew also exhibited both positive and

negative skew; both important features of propagation where the latter is not appreciated in the modelling of community propagation. Positive skew indicates a long tail of 'hotspots' and 'super-spreading' events whilst negative skew indicates a long tail of 'cold spots' and 'super-isolators'. In contrast, COVID-19 deaths exhibit near constant scale, variance and skew despite the extended studied timeframe, government intervention, different testing regimes and the national vaccination programme. This was also evident in the regions position relative to the PL where it remained either below or above the expectation throughout the pandemic. In the initial study of COVID-19, residual variance did not meet the conditions of standard linear regression. The variance expanded and contracted over time and residual distributions included both positive and negative skew, thus, normality and homoscedasticity assumptions of standard linear regression were not always met. This investigation stimulated the development of the generalised logistic distribution (GLD) within a Bayesian framework to model expectation and dispersion using Markov chain Monte Carlo (MCMC) methods. The advantage of the GLD is its flexibility when looking at skewed or otherwise non-normally distributed data. The GLD regression model and its key features are demonstrated using COVID-19 data. However, the proposed framework will benefit a range of systems with linear structure. The additional dispersion regression coefficients account for heteroscedasticity together with the parameters of the GLD to provide more realistic shapes (e.g., skew). The normal regression model assumes a normally distributed homoscedastic system, producing relatively large model bias, relative to the improved GLD regression model. Gelman-Rubin diagnostics and deviance information criterion (DIC) was included in the proposed framework showing good convergence and data were explained well by the fitted GLD regression model.

Acknowledgements

First and foremost, my utmost gratitude is to Dr Quentin Hanley, my director of study and Dr Golnaz Shahtahmassebi, my second supervisor, who have both inspired and supported me throughout the PhD program. Your invaluable advice, encouragement and feedback have been incredibly important, offering ideas and suggestions throughout, which not only benefited the research in this thesis, but it also developed me as a person. I shall never forget! I would also like to share my appreciation to special colleagues and friends: Prof Haroldo Ribeiro, Bruna Falgueras and Marcin Fiedziukiewicz: who have also supported me and offered excellent advice.

A special thanks to my partner, Charlotte Ireland. Their love, support and belief in me has been phenomenal throughout this journey and it has been a big part of the success of this thesis. Lastly, I would like to thank my incredible family for all their unconditional love and encouragement.

Author's Declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award. The PhD doctorate was gratefully financed with the aid of Nottingham Trent University Central Funding. Relevant scientific seminars and conferences were regularly attended which work was often presented.

Journal Articles:

1) 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

2) Sutton, J, Shahtahmassebi, G, Ribeiro, H.V. et al. Rural–urban scaling of age, mortality, crime and property reveals a loss of expected self-similar behaviour. Sci Rep 10, 16863 (2020). https://doi.org/10.1038/s41598-020-74015-x

3) Ribeiro HV, Sunahara AS, Sutton J, Perc M, Hanley QS (2020) City size and the spreading of COVID-19 in Brazil. PLoS ONE 15(9): e0239699. https://doi.org/10.1371/journal.pone.0239699

Conferences/Contributions:

1) UN75+1 (Online, November 2021). Jack Sutton, Quentin Hanley, Golnaz Shahtahmassebi and Haroldo V. Ribeiro. COVID-19 daily evolution of propagation: regional preference and persistence, variance structures, “super-spreaders” and “super-isolators”.

2) UN75+1 (Online, November 2021). Golnaz Shahtahmassebi, Jack Sutton and Quentin Hanley. Bayesian inference for risk comparison with application to COVID-19 mortality rate in the UK.

3) NetSci2019 (Vermont, USA, May 2019). Jack Sutton, Golnaz Shahtahmassebi, Haroldo V. Ribeiro and Quentin Hanley. Can we see two types of crime based on connections to mortality and property. Are Networks the Way Forward?

4) 2019 EUROCRIM (Ghent, Belgium, September 2019). Quentin Hanley, Jack Sutton, Golnaz Shahtahmassebi and Haroldo V. Ribeiro. What is Crime attached to: Crime, Mortality and Property after removing per capita bias.

Conferences and courses attended:

1) Rstudio::global2021 (Online, January 2021). RStudio Virtual conference.

2) Graphs and Networks (Online, July 2020). A one-day virtual conference hosted by the MGGG Redistricting Lab of Tisch College, Tufts University.

3) Academy for PhD Training in Statistics (2018-19):
   - University of Oxford, 2019. Design of Experiments and Studies. Flexible Regression.
   - University of Southampton, 2019. Applied Stochastic Processes. Statistical Modelling.

- Durham University, 2019. Computer Intensive Statistics. High-dimensional Statistics.
- University of Cambridge, 2018. Statistical Computing. Statistical Inference.

Word count for this thesis: 46138

Signed: 

Date: 17/05/2023

Contents

Abstract

Acknowledgements

Author's Declaration

Abbreviations

List of Latin Symbols

List of Greek Symbols

List of Mathematical Notation and Symbols

Abstract 3

Acknowledgements 5

Author's Declaration 6

Chapter 1: Overall Introduction 23

Chapter 2: Standard Power-law (PL) Methodology 30

2.1 Introduction 30

2.2 Theory 30

2.2.1 Scaling Framework 31

Chapter 3: Rural-urban Scaling of Age, Mortality, Crime and Property 35

3.1 Introduction 35

3.2 Theory 36

3.2.1 Bootstrapping 36

3.2.2 Networks 37

3.2.3 Self-Organising Maps 40
3.2.4 Similarity Measures 42
3.3 Materials and Methods 44
3.3.1 Datasets 44
3.3.2 Statistical Analysis 46
3.4 Results and Discussions 47
3.4.1 Overview of Regions 47
3.4.2 Rural-urban scaling 50
3.4.3 Critical Densities 53
3.4.4 Correlation and hierarchical clustering of residuals by category 55
3.4.5 Analysis of residuals by region 60
3.4.6 Self-organizing maps 61
3.5 Summary 64
3.5.1 Lognormal Method for Crime, Property, Mortality and Age 65
Chapter 4: Population Density and Spreading of COVID-19 in England and Wales 69
4.1 Introduction 69
4.2 Theory 71
4.2.1 Similarity Measures 71
4.2.2 Residual and Case Density Models 73
4.3 Materials and Methods 75
4.3.1 Datasets 75

4.3.2 Statistical Analysis 77
4.4 Results and Discussions 77
4.4.1 Overview of regions, cases and number of observations 78
4.4.2 Daily Progression of COVID-19 81
4.4.3 Daily Exponent, Variance and Skewness for Cases 82
4.4.4 Daily Exponent, Variance and Skewness for Deaths 89
4.4.5 Age Demographics 92
4.4.6 Dispersion of COVID-19 Case Residuals over Time 94
4.4.7 Regional Persistence of COVID-19 Case Residuals 96
4.5 Summary 98
4.5.1 Limitations 101
4.5.2 Lognormal Methods for Cases 102
4.5.3 Lognormal Methods for Deaths 104
Chapter 5: Generalised Logistic Regression Model 106
5.1 Introduction 106
5.1.1 Introducing the Generalized Logistic Distribution 109
5.2 Introducing the Generalized Logistic Distribution Regression Model 115
5.2.1 Linear Models 116
5.2.2 Generalised Linear Model 117
5.2.3 Proposed GLD Regression Model 119
5.3 Bayesian Framework 120

5.3.1 Bayesian Inference: Prior Distribution 121

5.3.2 Bayesian Inference: Posterior Distribution 121

5.3.3 Metropolis Hastings Algorithm 122

5.3.4 Implementation of the Metropolis Hasting Algorithm 124

5.3.5 Gelman Convergence Diagnostics 125

5.3.6 Deviance information criterion (DIC) 125

5.4 Bayesian GLD Regression Model Remarks 126

5.5 Summary and Conclusions 127

5.5.1 Possible Extensions and Wider Application 129

5.5.2 Limitations 131

Chapter 6: Application of the GLD Regression Model 132

6.1 Introduction 132

6.2 Dataset and R Packages 134

6.3 Application of the GLD regression Model 134

6.4 Implementation 136

6.5 Key Features 137

6.6. Model Checking 145

6.7 Summary 146

6.7.1 Further Limitations 148

Chapter 7: Overall Conclusions 149

7.1 Overall Summary 149

7.2 Other Contributions 153

7.3 Final Remarks 154

References 156

Appendix A: Age, Mortality, Crime and Property Study 174

A1. Similarity Measures 174

A2. Dataset. Data employed in this study. Data covering the period from 2013-2017 were captured on 20/03/2019. 175

A3. AIC and BIC values for single and double power-law models. 178

A4. A list of all low and high exponents, critical densities (if applicable) and intercepts. 180

A5. Density scaling against specific age groups 181

A6. Fits for metrics with restricted age groups. 182

A7. Mortality, Crime, Property and Age in Isolation Heatmap and Example Scatter plots. 182

A8. Age, Mortality, Crime and Property Heatmaps (Other Similarity Measures) 187

A9. Clustering. 192

Appendix B: COVID-19 Study 196

B1. Daily distributions of Per-Capita COVID-19 cases 196

B2. Daily scaling of COVID-19 cases 213

B3. Daily scaling of COVID-19 deaths 244

B4. Daily geoplots of COVID-19 case residuals 276

B5. Daily geoplots of COVID-19 death residuals 314

B6. Pairwise correlation between all days of COVID-19 case residuals 352

B7. Scaling of all 18 age categories 353

B8. Residual relationships between all 18 age categories 354

B9. Daily distributions of COVID-19 case residuals 355

B10. Daily distributions of COVID-19 death residuals 373

Appendix C: Examples of Linear Regression Models 391

C1. Null Model 391

C2. Simple Linear Regression 391

C3. Polynomial Regression 391

C4. Multiple Regression 391

Appendix D: Normal Regression Model 393

D1. Normal Regression Model with Varying Variance Model 393

Appendix E: Gelman and Rubin's Convergence Diagnostics and trace plots of the MCMC chains for the GL power-law model 394

E1. Daily Gelman and Rubin's R Statistics 394

E2. Daily Gelman and Rubin's R Statistics over 20000 iterations 397

E3. Daily parameter trace plots 440

E4. Daily posterior distributions 525

Appendix F: Daily Posterior Mean and SD 551

F1. Daily Posterior mean along with credible intervals and SD 551

Appendix G: Daily scaling using GL power-law 556

G1. Daily scaling of COVID-19 cases comparing normal (fix and varying) with the GL power-law model 556

Appendix H: R Code.....584

H1. GLD Regression Model.....584

H2. The M-H algorithm 585

H3. API to extract COVID-19 Data from PHE.....588

Abbreviations

TL	Taylor's Law
PL	Power Law
GLD	Generalised Logistic Distribution
MCMC	Markov Chain Monte Carlo
M-H	Metropolis Hastings
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
DIC	Deviance Information Criterion
IID	Independent and Identically Distributed
CCGs	Clinical Commissioning Groups
SOMs	Self-organising Maps
BMU	Best Matching Unit
DSAMs	Density Scale Adjusted Metrics
ONS	Office for National Statistics
ASB	Anti-Social Behaviour
CD&A	Criminal Damage and Arson
CSR&A	Colon, Sigmoid, Rectum and Anus
PHE	Public Health England
LTLAs	Lower Tier Local Authorities
PDF	Probability Density Function
CDF	Cumulative Density Function
MSOA	Middle Super Output Area
GLM	Generalised Linear Model
MOM	Method of Moments
MLE	Maximum Likelihood Estimate

List of Latin Symbols

X	An independent random variable
Y	A dependent random variable, also referred to as the indicator in a PL model
Y_0	Pre-exponential factor parameter in a single PL model
P	Population variable in a PL model
Y_D	Indicator density
P_D	Population density
A	An area of a region
d^*	A critical density change point
Y_1	Pre-exponential parameter above the density change point
B	The total number of bootstrap samples
V	Represents a set of nodes in a correlation network
E	Represents a set of edges in a correlation network
n^*	Indicates the number of nodes in a correlation network
m^*	Indicates the number of edges in a correlation network
A	Represents an adjacent matrix corresponding to a correlation network
k	Number of partitions in a correlation network
n_p	Total number of partitions in a correlation network
Q	Modularity Score for a correlation network
n_{SOM}	Represents the size of the input vector going into the SOM
m_{SOM}	Represents the number of SOM nodes
w	Weight of each BMU in the SOM
k_{clust}	Represents the number of clusters identified in the gap statistic
W_k	Pooled within-cluster sum of squares in the gap statistic equation

s_k	Simulation error term in the gap statistic equation
r	Represents a computed Pearson's correlation
p	The number of explanatory terms in a linear model
M'	Represents the total number of MCMC chains
t	Denotes the iteration step in the M-H algorithm
d	Indicates the dimension of the parameter vector space
k_{pen}	Denotes a penalty term in the AIC score
n	Denotes the sample size

List of Greek Symbols

β	Scaling exponent parameter in a PL model
β_d	Scaling exponent parameter in a density PL model
μ	Mean parameter corresponding to the normal distribution (also referred to as the location parameter)
σ^2_{norm}	Variance corresponding to the normal distribution (Scale Parameters)
ε	Represents residuals: the differences between the observed and predicted values.
β_L	Scaling exponent parameter below the density change point in a segmented PL model
β_H	Scaling exponent parameter above the density change point in a segmented PL model
θ_{corr}	Bootstrap sample correlation
τ	SOM learning rate parameter
θ	The location parameter corresponding to the GLD
σ	The scale parameter corresponding to the GLD
α	The shape parameter corresponding to the GLD
λ	The acceptance ratio corresponding to the M-H algorithm
μ'_R	The distribution rth moment where R denotes the number of moments
η	Represents the expectation linear predictor that links the linear model to the GLDs location parameter
η'	Represents the variance linear predictor that links the linear model to the GLDs scale parameter
$\boldsymbol{\beta}$	Represents a vector of mean regression coefficients
$\boldsymbol{\beta'}$	Represents a vector of variance regression coefficients

ψ_{par}	The parameter vector space
a_1	The shape parameter corresponding to the gamma distribution
b_1	The scale parameter corresponding to the gamma distribution
Σ	The variance-covariance matrix
β_0	Mean regression coefficient corresponding to the intercepts
β_1	Mean regression coefficient corresponding to the expectation
β_0'	Variance regression coefficient
β_1'	Variance regression coefficient corresponding to heteroscedasticity

List of Mathematical Symbols and Notation

$\mathbb{R}$	Real number set
$\mathbb{R}^+$	Positive real number set
$\in$	Element of symbol
$\sim$	Denotes connectivity. For example, in the correlation network nodes i and j are connected when $i \sim j$
$\subset$	Subset symbol. Subset has fewer elements than the set.
$\cap$	Intersection symbol. Elements that belong to both sets.
$\neq$	Not equal to symbol
$>$	Greater than symbol
$<$	Less than symbol
$\geq$	Greater than or equal to symbol
$\leq$	Less than or equal to symbol
$\forall$	Universal quantification symbol indicating relation to every member of the domain.
π	Represents an Irrational number approximately equal to 3.14159 (to 5.d.p)
∂	Denotes the derivative of a function
$(.)^T$	Indicates the transpose of a matrix
(X_{res}, Y_{res})	Represents a complete pair of residuals
$\rho(X_{res}, Y_{res})$	A function to denote Pearson's correlation coefficient between a pair of residuals
$S(rgX_{res}, rgY_{res})$	A function to denote Spearman's correlation between a pair of residuals
$K(X_{res}, Y_{res})$	A function to denote Kendall's correlation between a pair of residuals
$c(X_{res}, Y_{res})$	A function to denote cosine similarity between a pair of residuals
$J(X_{res}, Y_{res})$	A function to denote Jaccard similarity between a pair of residuals

$E(.)$	The expected value of a random variable
$Var(.)$	The variance value of a random variable
$Skew(.)$	The skew value of a random variable
$g(.)$	The link function in a GLM
Γ	The gamma function
$\psi(.)$	First derivative of the gamma function
$\psi'(.)$	Second derivative of the gamma function
$\psi''(.)$	Third derivative of the gamma function

Chapter 1: Overall Introduction

The work in urban scaling dates back to the early 1900s [1] and inspired others to explore how city sizes are distributed by rank order [2–4]. Inspired by some of this work, Nordbeck [5] comprehensively presents a generalisation of allometric growth (power-laws) which has formed the foundations of urban scaling [6]. The mathematics of allometric growth and ideas of urban scaling has become a feature of modern urban studies. During a similar time period, fluctuation scaling, first explored in agriculture [7], noted in a logarithmic presentation that a variance and area relationship forms a linear relationship, now known as the Taylor's Law (TL) [8,9]. The scaling relationship (TL) in logarithmic mean-variance plots has been found in a host of studies, beyond agriculture, linked to a range of different disciplines [10–14].

It has been observed that cities have agglomeration economies [15,16]. The historical framework [17,18] showcasing the evolution of agglomeration economies, indicates this was first identified over a century ago [19]. Since, this development of cities [20] has partly motivated the research leading to the work in urban scaling. Conventionally, in urban scaling, power laws (PL) are used to investigate the relationship between population and city indicators (e.g., GDP). The value of the PL scaling exponent indicates the type of scaling relationship. Studies have shown that super-linear scaling happens in desirable urban indicators, such as patents and total employment [21–23]. Similarly, undesirable urban indicators, such as homicide, general crime and AIDs cases, have also been shown to exhibit increasing returns with population scale [21,24–27]. It has also been previously observed that some urban indicators, such as number of petrol stations and road surfaces, exhibit sub-linear scaling indicative of economies of scale [21].

Much of the scaling literature has focused on urbanised systems and population linking them to a host of indicators [21–28]. Spatial coverage and the definition of regional boundaries is varied and in many scaling studies [21–23,25,28,29] there has been a missed opportunity to study rural environments. Although global urban population is large, rural land mass still dominates the world [30] and it is unclear whether rural-urban regions are fundamentally different. Temporal coverage is usually cumulatively over a large timeframe (monthly, yearly) since finer granularity of data are difficult to find. Thus, a finer evolution of scaling metric has not yet been well explored. A fundamental and underpinning part of the work in scaling is the self-similar behaviour of regions that is well approximated using PL models [31]. Population has long been used as the predictor in PL models, but more studies of crime and property proceeding this thesis have shown population density to be a better fit [26,27,32]. In either case (population or population density) the corresponding PL model parameters are estimated using simple linear regression techniques such as the least square method [33]. These techniques rely on assumptions of normality, homoscedasticity, independent errors and linearity [33–36] which overall are underappreciated assumptions in the scaling literature. This is an important consideration as it can influence the fitting of the estimated model parameters and if not considered carefully can cause model bias. This indicates a need for an alternative methodology that is more flexible and reliable.

Least square methods have been applied to estimate model parameters when fitting single and segmented PL models to log-transformed data in chapters 3 and 4. However, it is worth noting some limitations leading to bias in estimated model parameters caused when following a set of strict linear regression assumptions. These are important considerations not usually accounted for in the scaling field. These are as follows: 1) data are assumed independent, 2) variance around the mean are assumed Gaussian and independent to the mean, 3) the system is assumed homoscedastic (constant variance) and 4) the log-transformed data is assumed to be linear. Furthermore, statistical analysis such as R^2 values only provide insight into the amount of variation explained by the model and does not consider the statistical significance that the data was generated by the model. Leitão et al [37]

proposes a probabilistic framework where fluctuations are accounted for explicitly: a more robust approach to fitting estimated model parameters in scaling. Specifically, the models proposed have different assumed noise (Gaussian, lognormal, person) and are evaluated using a range of different datasets. Leitão et al [37] estimate the exponents using maximum likelihood estimations (MLEs), bootstrap (with replacement) confidence intervals, test whether the reported scaling is valid by computing p-values (p-value > 0.05 indicated that the model is not rejected), test for non-linear scaling (i.e. when $\beta \neq 1$) and use BIC scores to inform of the best performing model. Strikingly they found that most reported scaling using PLs were rejected. In some cases, the reported exponents between models was so extreme that scaling relationships changed entirely (e.g. the reported exponent for EU cinema using the lognormal model with fixed fluctuation was 1.46 (0.19) indicating super linear scaling, whilst with free fluctuation was 1.00 (0.30) indicating linear scaling [37]).

This thesis is presented in 7 chapters. Chapter 1 is a broad overview of the establishment of behaviour relating to urban growth and PL models to predict urban indicators. Many current scaling studies focus on urbanized regions whilst neglecting rural regions. The current methodology uses standard linear regression with normality, homoscedasticity, independence and linearity assumptions. The data explored in scaling is usually cumulative spanning across a large timeframe. Chapter 2 comprehensively presents the current mathematics of PL models along with the familiar logarithmic presentation of urban scaling laws using normality methods.

Chapter 3 expands on previous studies of crime and property, exploring additional indicators using population density PL models to better understand the inter-relationships between indicator residuals and to explore regional variation. The applied indicators included mortality (cancer, dementia, suicide, homicide etc) and age to determine whether these behaved in a similar way to crime and property and continue to explore the density scaling framework. The key findings were published in the journal Scientific Reports [38] and presented within this thesis in chapter 3. In reflection of the Leitão et al [37] study, the lognormal model with fixed and free variance was applied

to all 67 indicators of crime, property, mortality and age. These results are compared to least squared methods to highlight any differences, explore the effect of variance and to validate whether the scaling holds for all reported single and segmented relationships.

Chapter 4 is a study of COVID-19 using case and death metrics. Initially the goal was to consider population density as a proxy for infectious interaction. It became a unique opportunity to investigate a contagious transmissible disease during a worldwide pandemic and the first to analysis the daily evolution of scale along with other descriptive statistics such as variance and skew. It also piloted the use of the generalised logistic distribution (GLD) to model residuals. The normal distribution was not always a good fit whilst the GLD was able to capture realistic shapes including positive and negative skew. The latter is indicative of 'cold spots' and 'super-isolated' regions which is a feature of the pandemic that is underappreciated in the modelling of COVID-19. The key findings up to day 446 were published in PLoS ONE [39] and presented within this thesis in chapter 4- updated to the full 759 days of data availability. The analysis of COVID-19 case residuals in chapter 4 demonstrated that the current methodology underpinning PL models in scaling are unsatisfactory. In the current framework simple linear regression techniques such as the least squares method are applied to fit PL model parameters. It is assumed that residuals are linear, normally distributed, independent and homoscedastic. In the study of COVID-19 these assumptions were often violated and in large parts of the studied timeframe were better characterized using a GLD opposed to a normal fitting. The GLD is a three-parameter distribution consisting of: location, scale and shape parameters. This makes the GLD a flexible, versatile and applicable framework to model complex shapes including positive and negative skew. This motivated the development of an alternative methodology improving current methods. Similar to chapter 3, in view of the Leitão et al [37] study, the lognormal model with fixed and free fluctuations have been applied to the daily COVID-19 cases and deaths data to see how least square methods compare. Leitão et al [37] present an initial insight into the effect of variance to PLs when considering it as either fixed or free. Meanwhile, modelling variance in a unified framework has been further developed in chapters 5 and 6, where other distributions, other than the normal, have been considered in response to results found in the COVID-19 study.

Chapter 5 presents a generalised expectation and dispersion regression model where the response Y follows a GLD in a Bayesian framework. It begins by demonstrating why the GLD is a good candidate distribution and its wide applicability. A Bayesian approach was adopted to incorporate non-informative priors with mean zero and a large variance to express prior ignorance. This yielded a complicated non-analytically tractable joint posterior distribution. A Markov Chain Monte Carlo (MCMC) method was applied; specifically, the walking Metropolis Hasting (MH) algorithm was implemented to fit the proposed model parameters over 20,000 iterations disregarding the first fifty percent as 'burn-in' and using Gelman convergence diagnostics to monitor whether the chains depart from the initial values and converge. Deviance information criterion (DIC) is a Bayesian version of the Akaike information criterion (AIC) score [40] and within this framework, the GLD model will be compared with normal regression models and monitor goodness-of-fit.

Chapter 6 is the application of the developed Bayesian GLD regression model in chapter 5 to demonstrating the key features of the model using the COVID-19 data. Although the method was applied to PL models and data from the pandemic, it is general to a wide range of systems currently using linear regression. Expectation and dispersion parameters generally converge well across the entire 759 days. A significant benefit of the proposed model are the additional parameters such as a dispersion regression coefficient accounting for heteroscedasticity together with the additional parameters of the GLD to provide more realistic shapes (e.g., skew). Since the same data from the study of COVID-19 in chapter 4 was applied using normality methods a comparison between Bayesian GLD regression model, where possible, were compared with the previous standard normal regression model to examine the variations between the two models highlighting model bias created by the latter. Leitão et al [37], first proposed an alternative to PL scaling ,providing a more robust framework to fitting PL model parameters. The modelling is built around the Gaussian distribution where different variations of the model are considered. This includes fixed and free fluctuations along with a case when the exponent is restricted to 1 (assuming linearity). Chapter 5 expands on these contributions where the modelling process is not assumed Gaussian. Instead, modelling is built around

the GLD in reflections of observations seen in chapter 4. The main difference is that the proposal in chapters 5 and 6 allows for complex shapes such as positive and negative skew and recognises features of heteroscedasticity in the data. Furthermore, the modelling is developed within a Bayesian framework, opposed to Leitão et al methods who adapt to a frequentist approach. Considering the problem at hand, a Bayesian framework, allows for external information and the use of elicitation techniques. The rewards of adapting to a more flexible framework is showcased using COVID-19 data and are compared to previous methods.

Chapter 7 summarises the main findings of the thesis along with an outline of the novelty and additional knowledge gained during this PhD research. The new proposed methodology presented in this research has been shown to be a significant improvement to previous methods and can be applied to a wide range of applications. The limitations and future opportunities are also acknowledged within this chapter to ensure for the continuation and development of this research in future work.

The initial analysis and discussions of COVID-19 [38] (chapter 4) contributed to the published work on the study of city size and the spreading of COVID-19 in Brazil [41] using conventional population data. At the earlier stages of the pandemic, discussions led to the notion that population density is a proxy for interaction and opportunity for spreading. For example, in highly dense regions such as a city, more person-to-person interaction is expected and therefore a greater opportunity for spread compared to a rural region where there are fewer interactions and opportunity for spreading the disease is significantly reduced. In contrast to the snapshot presented of Brazilian cities using conventional population centres, chapter 4 reveals population density was not as simple proxy for interaction and instead exhibited considerable complexities throughout the pandemic. Instead, in the extended timeseries in the case of the UK rural-urban regions preference of spread switched 6 times throughout the 25-month studied timeframe. The exploration of population density as a proxy , contributed to the work, available in [41], and was initially motived by discussions involving myself, but the analysis was primarily conducted by others. Because of this, the discussions

leading to this work is mentioned, but an in-depth presentation of this work is not available within this thesis.

Chapter 2: Standard Power-law (PL) Methodology

2.1 Introduction

Initially, motivated by the observations of agglomerations, studies in urban scaling have been long linked with PL connecting urban indicators with population in the most urbanised environments (cities). However, recent studies include rural regions and PL models have since been adapted using densities opposed to discrete values [26,27]. This is an area normalising approach to rural-urban scaling and has been shown to be a better fit: two regions of equal population but different area might not be expected to have equal behaviour. The inclusion of rural regions in some metrics causes segmentation but PL with a change point can accommodate these relationships and with model preference based on AIC and BIC Information Criteria scores. The log transformation of the models forms a simple linear regression structure and model parameters are fitted using conventional least squares methods. The scaling exponent obtained from the model indicates the scaling behaviour and whether it is sub-linear, linear or super-linear. Residuals obtained from the preferred model are used to better understand correlation between metrics after removal of *per capita* bias and to explore regional variation.

2.2 Theory

In this section the rural-urban scaling framework is presented, first by introducing conventional approaches using discrete data. The data are then transformed by considering an area normalization

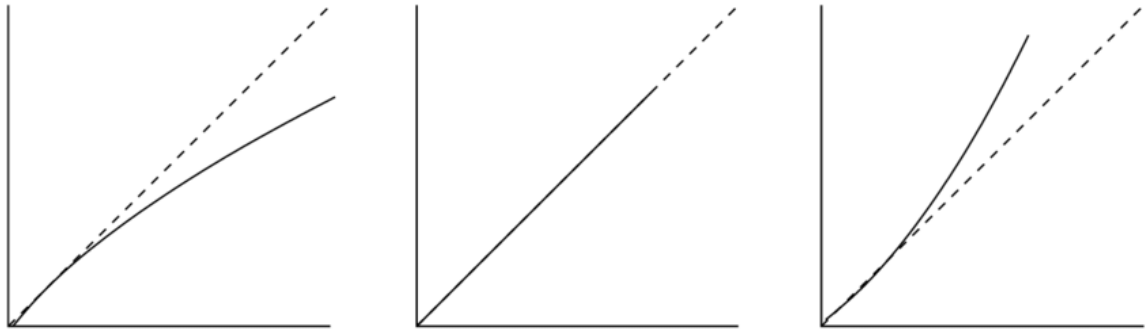
process that converts raw values into densities. Many of the methods in this section are adapted from studies of crime and property [26,27] and a host of other scale adjusted metrics [42] and were used in the analysis of crime, property, mortality and age data in chapter 3 and later on in the analysis of COVID-19 data in chapter 4.

2.2.1 Scaling Framework

The agglomeration economies effect has inspired the works of urban scaling. The ambitions to model city behaviour have led to a simple PL in the form of:

$$Y = Y_0 P^{\beta} 10^{\varepsilon} \qquad 2.1$$

where, Y is an indicator, Y_0 is a pre-exponential factor, P is the population, β is the scaling exponent and ε are residuals that are independent and identically distributed (IID) with common $N(0, \sigma^2_{norm})$ where 10 is a consequence to base 10 log transformation. The value of β indicates the type of scaling relationship and whether it is sublinear, linear or super-linear. When $\beta < 1$, the scaling relationship is sub-linear (Fig 2.1 (a)), when $\beta = 1$, the scaling relationship is linear (Fig 2.1 (b)) and when $\beta > 1$, the scaling relationship is super-linear (Fig 2.1(c)). The agglomeration economies are observed when $\beta > 1$ indicative of increased returns in larger populated regions. This is advantageous in indicators of innovation, but also includes a host of other, disadvantageous indicators, such as serious crimes [21]. However, there is some evidence [43] that some indicators within the industry sector reporting super linearity ($\beta > 1$) will not always be long lasting and depends on the level of maturity of the sector. After the initial period of innovation usually occurring in big cities, the exponent becomes linear indicating that the productivity has moved beyond the urbanised environment.



(a) (b) (c)

Fig 2.1. Scaling Relationships. (a) $\beta < 1$ illustrating sub-linear scaling, (b) $\beta = 1$ illustrating linear scaling and (c) $\beta > 1$ illustrating super-linear scaling. The dashed line is a complete linear relationship.

Population density scaling is an area normalised approach to scale adjusted metrics [42] which has been shown to better predict a host of indicators when using arbitrary urban and rural regions to define boundaries. This adaptation to equation 2.1 is now in the form given by:

$$Y_D = Y_0 {P_D}^{\beta} 10^{\varepsilon} \qquad 2.2$$

where $Y_D = Y/A$ and $P_D = P/A$ is the population and indicator density respectively where A are regional areas. The residuals, ε, have the same definition to equation 2.1. Similar to population scaling, the exponent, β, provides insight into the types of scaling relationships exhibited such as whether it is sub-linear, linear or super-linear (Fig 2.1). In the context of rural-urban scaling, in the case of sub-linear scaling, rural regions are preferentially affected by acceleration. In the case of super-linear scaling, urban regions are preferentially affected by acceleration. Linear scaling indicates constant return despite the density of a region. The log transformation of equation 2.2 is given by:

$$\log(Y_D) = \log(Y_0) + \beta \log(P_D) + \varepsilon \qquad 2.3$$

which is useful such that the least squares method can be applied to obtain PL model parameters. The estimated parameter $\hat{\beta}$ is obtained using the following [44]:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(\log(P_{D,i}) - \log(\bar{P}_D))(\log(Y_{D,i}) - \log(\bar{Y}_D))}{\sum_{i=1}^{n}(\log(P_{D,i}) - \log(\bar{P}_D))^2}$$

and the estimated parameter$\log(\hat{Y}_0)$ is obtained by:

$$\log(\hat{Y}_0) = \log(\bar{Y}_D) - \hat{\beta}\log(\bar{P}_D)$$

where $\log(\bar{P}_D)$ is the population density sample mean and $\log(\bar{Y}_D)$ is the indicator density sample mean in logarithmic space and n is the number of complete pairwise observations. It has been shown that empirically a density change point, d^*, between rural-urban regions appear consistently in the range of 10-70 people per hectare [26,27]. To account for this change point, equation 2.3, needs adjusting such that:

$$\log(Y_D) = \begin{cases} \log(Y_0) + \beta_L \log(P_D) + \varepsilon & P_D < d^* \\ \log(Y_1) + \beta_H \log(P_D) + \varepsilon & P_D \geq d^* \end{cases} \quad 2.4$$

where β_L and $\log(Y_0)$ are the exponent and intercept below the critical density, d^*, and β_H and $\log(Y_1)$ are the exponent and intercept above the critical density, d^*.

Then residuals, ε_i, are the difference between the observation and expected values and where model parameters ($\hat{\beta}$, $\hat{Y}_0$) minimise the variance $\sum \varepsilon_i^2$ such that:

$$\varepsilon_i = \log(Y_{D,i}) - \log(\hat{Y}_{D,i}) \quad 2.5$$

for $i = 1, \ldots, n$ and $\log(\hat{Y}_{D,i})$ is the estimate of $\log(Y_{D,i})$. Residuals are obtained from the preferred model, whether that being a single or segmented PL model. If the residual value ε_i is negative, then a region is below the expectation. Contrary, if the residual value $\varepsilon_i$is positive then the region is above the expectation.

This standard approach has limitations depending on the data set due to normality, homoscedasticity, linear and independent assumptions. Chapter 3 presents an example of the conventional analysis expanding the scope of population density analysis to include 67 indicators of analysis. Chapter 4 makes clear that while useful; the assumptions of classical least squares are extensively violated. Chapter 5 is a developed Bayesian GLD regression framework that improves the conventional methodologies that were applied in chapters 3 and 4. This developed approach is extremely useful in a scaling setting, but the proposal is more general to other disciplines. Overall, the developed methodology is a flexible approach and assumptions of standard linear regression are no longer necessary.

Chapter 3: Rural-urban Scaling of Age, Mortality, Crime and Property

3.1 Introduction

Human activity has become increasingly concentrated in urban regions. Although more people live in highly dense regions like cities; rural regions still make up most of the land mass [45–48]. There is an inherent bias when studying cities in the scaling literature as a consequence of neglecting rural regions. Thus, it is not clear whether urban systems are fundamentally different to rural systems. Recent studies [26,27] of crime and property revealed a critical population density between rural-urban regions around which some crime and property types accelerated (e.g. Bike Theft, Robbery, Flats) and others declined (e.g. Anti-social behaviour, Shop-lifting, Semi-detached) in urban regions. This critical density delineating rural-urban transitions exhibited typically in the range of 10-70 people per hectare, but the statistical mechanics underlying this behaviour is unclear.

The application of PL has improved the understanding of scaling. However, *per-capita* models remain in many policies and resource allocation. Policy makers need to understand that these models are bias and rural-urban regions are fundamentally different in a host of indicators. For example, the distribution of health care resource in the UK is conducted by the clinical commissioning groups (CCGs). The allocation is based on *per-capita* adjusted for mortality, market forces and a range of other factors based on nutrition, obesity, smoking, drug use etc [49–51]. The PL scaling framework is an opportunity to better understand the taxonomy of a variety of metrics. Regional clustering and allocation is also thought as a north-south divide appearing in discussions of economic and social issues [52–54].

The definition of rural varies. Examples of regions that are considered rural include one or some of: 1) not an urbanized environment [55], 2) areas with a low population, 3) indicators of rural life [56], 4) density/area mass and/or 5) heat island intensity [57]. In this thesis rural and urban regions are on a spectrum from very low population density to very high population density. The regions position relative to population density provides a quantitative metric along that spectrum.

This chapter investigates a range of mortality, crime, property and age indicators throughout England and Wales using unitary authorities, non-metropolitan districts, metropolitan boroughs and London boroughs to define regional boundaries. This will be explored to determine if mortality behaves similarly to previous work [26,27]. The key advance in this study is the introduction of 26 mortality indicators along with 18 age categories. This is a diverse dataset and the first to attempt to explore an extensive set of indicators, including 67 metrics in total, within the population density framework. The additional age categories are shown to break the self-similarity underlying the urban scaling hypothesis and a fundamental part of the acceleration and decline in urban regions. In addition, this scaling study is the first to introduce SOMs used to observe regional similarity and explore age demographics.

3.2 Theory

In this section the framework defined in section 2.2 is applied and residuals from the preferred single (equation 2.3) or segmented (equation 2.4) PL model are analysed using techniques such as bootstrapping (section 3.2.1), networks (section 3.2.2), SOMs (section 3.2.3) and various similarity measures (section 3.2.4). The work extends the studies of crime and property [26,27] by including mortality and age metrics to explore whether these behave in a similar way.

3.2.1 Bootstrapping

Bootstrapping is a class of Monte Carlo methods that randomly resample from a distribution with replacement to assign measures of accuracy to sample estimates [58–63]. These techniques can be

applied in a regression setting to assess correlation and uncertainties of pairwise data. Consider some sample residuals $X_{res\,i} = (x_{res\,1}, x_{res\,2}, \dots, x_{res\,n})$ and $Y_{res\,i} = (y_{res\,1}, y_{res\,2}, \dots, y_{res\,n})$ where n is the number of complete pairwise data. To compute an estimate of the correlation and the bootstrap estimate of the standard error of that correlation between pairwise residuals, $X_{res\,i}$ and $Y_{res\,i}$, the bootstrap algorithm is applied [58,63–65] and proceeds as:

1. For each bootstrap replicate, indexed $b = 1, \dots, B$:
   a) Generate pairwise samples between $x_{res}^{*(b)} = x_{res\,1}^{*}, \dots, x_{res\,n}^{*}$ and $y_{res}^{*(b)} = y_{res\,1}^{*}, \dots, y_{res\,n}^{*}$ with replacement from the observed samples $(x_{res\,1}, x_{res\,2}, \dots, x_{res\,n})$ and $(y_{res\,1}, y_{res\,2}, \dots, y_{res\,n})$
   b) Compute the b^{th} replicate $\hat{\theta}_{corr}^{(b)}$ from the b^{th} bootstrap sample where $\hat{\theta}_{corr}$ is the sample correlation R between(X_{res}, Y_{res}).

2. The bootstrap estimate of the standard error $se(R)$ is the sample standard deviation of the replicates $\hat{\theta}_{res}^{(1)}, \hat{\theta}_{res}^{(2)}, \dots, \hat{\theta}_{res}^{(B)} = R^{(1)}, R^{(2)}, \dots, R^{(B)}$.

Here, this algorithm will be used to bootstrap, with 2000 replications, all the pairwise residual correlations, between indicators of age, mortality, crime and property to identify and construct networks with Pearson's correlation coefficients that are significant at 99% confidence.

3.2.2 Networks

Networks have a wide range of application [66–70]. It can be used as a tool to visualise and understand large complex multivariate data. A network $N = (V, E)$ is a set of nodes and edges that connects them such that $V = v_1, v_2, \dots, v_{n^*}$ is a set of n^* nodes and $E = e_1, e_2, \dots, e_{m^*}$ is a set of m^* edges between nodes [71,72]. A simple undirected connected network containing no loops or multiple edges in the form of an adjacent matrix $A \in \mathbb{R}^{n^* \times n^*}$ is defined as:

$$a_{i,j} = \begin{cases} 1 & i \sim j, \\ 0 & otherwise \end{cases} \quad 3.1$$

where $i \sim j$ denotes that the nodes i and j are connected. An example of an adjacent matrix is given by:

$$A = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

which produces the following network:

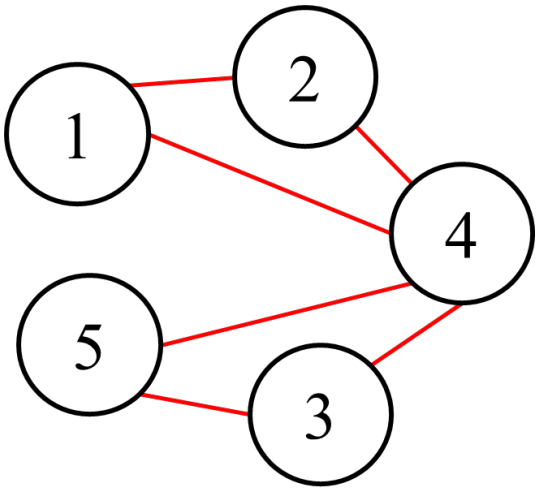


Fig 3.1. An example of a simple undirected network. Black circles represent nodes and red lines between nodes represent the edges that connects them.

3.2.2.1 Modularity

The modularity measure indicates community structure within a network [71,73,74]. This means that subsets of nodes are densely connected and sparsely connected to other communities in the network. There exist several algorithms [66,74,75] involving the partitioning of nodes V where a collection of sets $V_i \subset V$ such that $V_i \cap V_j = \emptyset$ and $\forall i \neq j$. The measure of modularity [66,75–77] is given by

$$Q = \sum_{k=1}^{n_p} \left[\frac{|E_k|}{m} - \frac{1}{4m^2} \left(\sum_{j \epsilon V_k} k_j \right)^2 \right] \quad 3.2$$

where $|E_k|$ is the number of links between nodes in the k_{th} partition of the network, V_k and n_p is the number of partitions. The value of Q is then optimised over all possible partitions. A network with high community structure (Fig 3.2) shows nodes forming dense intra-connections within clusters and clusters form sparse inter-connections with other clusters in the network.

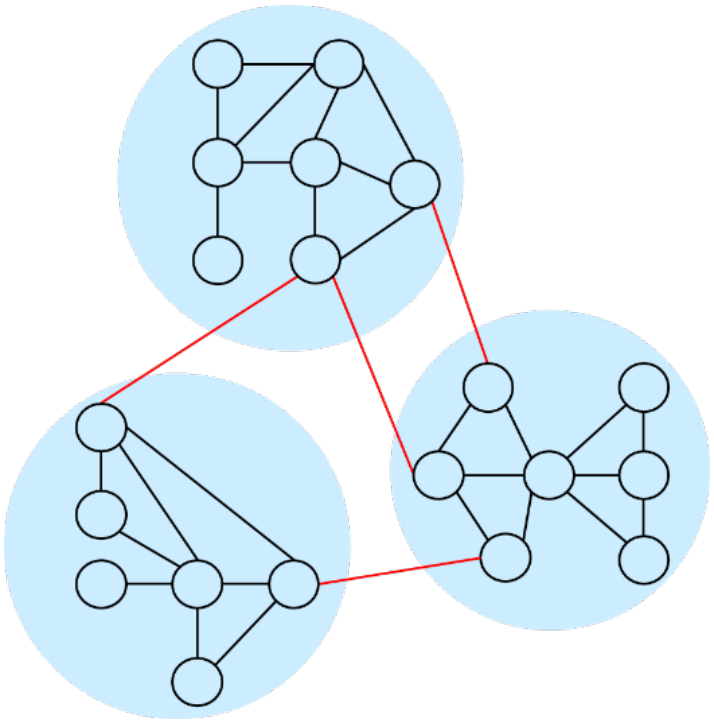


Fig 3.2. An illustration of community structure. Blue shaded areas represent individual communities. Black lines represent inter-connections (connections within communities) and red lines represent intra-connections (connections between communities).

In this network nodes represent indicators and edges between nodes represent significant correlation using bootstrapping methods (section 3.2.1). A current limitation of this network is a discontinuous representation of correlation as this will form two separate networks: one for all significant positive correlation and another for all significant negative correlation. The Pearson correlation was selected in correspondence to previous work on crime and property [26,27] although other similarity and correlation measures were explored (Appendix A1). In this thesis, the modularity score obtained using equation 3.2 is based on the Girvan-Newman algorithm [78]. To identify groups

of communities, the algorithm iteratively removes edges from the network with the highest betweenness centrality and recalculates until in breaks down the network into small hubs.

3.2.3 Self-Organising Maps

Self-Organising Maps (SOMs) [79,80] are introduced to analyse residuals for any potential regional clustering based on the value of the residuals. In general, this method allows high dimensional space to be represented in an array of pre-defined nodes, m_{SOM}. The arrangement of nodes can vary in size and geometrically (either hexagonal or rectangular) (Fig 3.3). These are both important predetermined aspects of the model. It has been proposed that these considerations depend on size and the dispersion of the input data [81] and ideally the resulting SOM should not have any empty nodes. The basic idea is each node in the SOM is examined to calculate the weight vectors that are most like the input vector. The winning node is commonly referred to as the best matching unit (BMU) [79]. This is performed iteratively and the resulting map is a low dimensional model providing information about how the input data is related [79,82]. The SOM algorithm [79] is as follows:

1. Randomize the node weight vectors in a map.

2. For each iteration t:

    a) Obtain the Euclidean distance between all the nodes in the map and calculate the similarity between each node and the input data. This is given by [83]:

$$D(j) = \sum_{i=1}^{n_{SOM}} \sum_{j=1}^{m_{SOM}} (x_{res\,ij} - w_{ij})^2$$

    where $i = 1, 2, \ldots, n_{SOM}$ is the size of the input vector and $j = 1, 2, \ldots, m_{SOM}$ is the number of nodes in the SOM.

    b) Then assign the input vector to the node j that has the minimum Euclidean distance.

c) Update the weight vectors of the 'winning' node and all the neighbourhood nodes such that [79]:

$$w_{ij}(t+1) = w_{ij}(t) + \tau(t)(x_{res}(t) - w_{ij}(t))$$

where in the context of this chapter the input vector $x_{res}(t)$ will be residuals for each region at iteration t. Thus, $w_{ij}(t)$ is the old weight and $w_{ij}(t+1)$ is the new weight. Finally, the parameter $\tau(t)$ in the model is known as the learning rate parameter, usually in the interval [0,1]. The value of $\tau(t)$ decreases with t. This will ensure that the model network (Fig 3.3) guarantees convergence. In the above, step 1 is the initialisation stage and steps in 2 is the learning phase. After the learning phase has completed, the resulting SOM will have positioned all input residuals into a node (Fig 3.3). The residuals positioned in the same node are similar and due to the algorithm above neighbouring SOM nodes will also have some degree of similarity. Clustering analysis will be applied to find all the neighbouring nodes in SOM that are also similar.

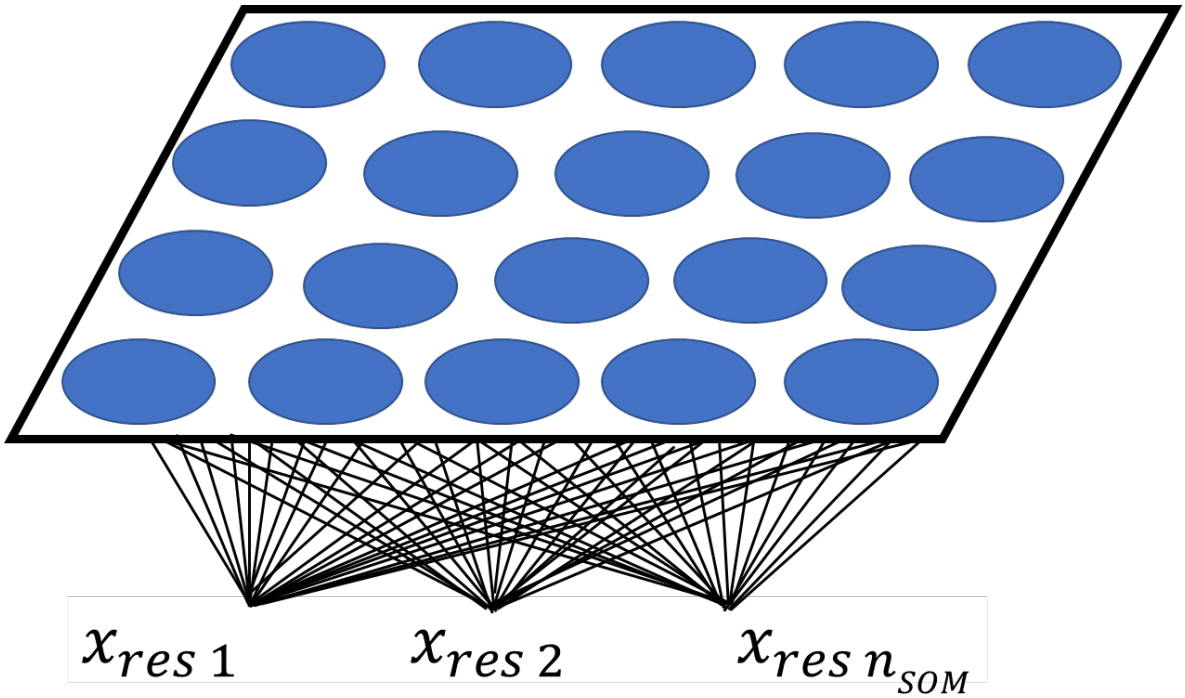


Fig 3.3. Self-Organising Map Architecture. Nodes are represented in blue and lines between the input vector and the nodes represent the iterative learning phase of the SOM algorithm.

Considering that the weights of the neighbouring nodes in the SOM are influenced by the BMU, a gap statistic was introduced to account for any clusters in the SOM [84]. The gap statistic [84] is given by:

$$Gap_n(k_{clust}) = E_n^*\{\log(W_{k_{clust}})\} - \log(W_{k_{clust}}) \quad 3.3$$

where E_n^* is the expectation and n is the sample size from the reference distribution. The reference distribution must meet one criterion. The distribution must have no structure and clusters. An obvious and the most common choice is the uniform distribution since this is a randomly distributed. The log $(W_{k_{clust}})$ is the pooled within-cluster sum of squares around the cluster mean where k_{clust} is the number of clusters. In each case $E_n^*\{\log(W_{k_{clust}})\}$ is estimated by an average of samples of $\log(w_{k_{clust}}^*)$. These are generated using Monte Carlo techniques where the sample $X_1^*, \ldots, X_n^*$ is drawn from the uniform distribution assuming that this is the chosen reference distribution. The value of the gap statistics indicates whether the clustering structure resembles the reference distribution. If it is large, then it does resemble the reference distribution. Contrary, if it is small then it does not resemble the reference distribution. After the completion of the algorithm, it can be found that the optimal number of clusters, denoted as k_{clust}, is found when $Gap(k_{clust}) \geq Gap(k_{clust} + 1) - s_{k_{clust}+1}$ where $s_{k_{clust}}$ is the simulation error in $E_n^*\{\log(W_{k_{clust}})\}$.

3.2.4 Similarity Measures

Similarity measures were explored to assess the similarity or dissimilarity between all pairs of indicator scale adjusted metrics. Residuals were obtained from all indicators and regions using the preferred single or segmented PL model. Suppose there are some pairs of residuals represented as $X_{res} = (x_{res\,1}, x_{res\,2}, \ldots, x_{res\,n})$ and $Y_{res} = (y_{res\,1}, y_{res\,2}, \ldots, y_{res\,n})$ where n is the number of

complete pairwise data. Then, the Pearson correlation between two sets of residuals is the covariance divided by the product of the standard deviations:

$$\rho(X_{res}, Y_{res}) = \frac{Cov(X_{res}, Y_{res})}{\sigma_{X_{res}} \sigma_{Y_{res}}} \quad 3.4$$

where $-1 \leq \rho(X_{res}, Y_{res}) \leq 1$. The correlation coefficient satisfies the relation $-1 \leq \rho(X_{res}, Y_{res}) \leq 1$ with the extremes indicating a perfect negative or positive linear relationship. If $\rho(X_{res}, Y_{res}) = 0$ there is no linear relationship; either the scatter of the residuals is random, or some form of non-linear relationship holds. A computationally convenient form [85] of equation 3.4 usually denoted as r is given by:

$$r = \frac{\sum_{i=1}^{n}(x_{res\,i} - \bar{x}_{res})(y_{res\,i} - \bar{y}_{res})}{\sqrt{\sum_{i=1}^{n}(x_{res\,i} - \bar{x}_{res})^2}\sqrt{\sum_{i=1}^{n}(y_{res\,i} - \bar{y}_{res})^2}} \quad 3.5$$

The correlation coefficient is one way of measuring how good a straight line fits the data. Other similarity measures exist and several of these were explored (e.g., Spearman, Kendall, Cosine and Jaccard). The Spearman correlation is the Pearson correlation coefficient between the rank variables. Thus, the residuals $X_{res\,i}, Y_{res\,i}$ are converted to $rg_{X_{res\,i}}, rg_{Y_{res\,i}}$ and the covariance is divided by the product of the standard deviations:

$$S(rg_{X_{res}}, rg_{Y_{res}}) = \frac{Cov(rg_{X_{res}}, rg_{Y_{res}})}{\sigma_{rg_{X_{res}}} \sigma_{rg_{Y_{res}}}} \quad 3.6$$

The Kendall correlation between residuals is the difference between concordant pairs and discordant pairs divided by the binomial coefficient for the number of ways to choose two items from n items:

$$K(X_{res}, Y_{res}) = \frac{n_c - n_d}{\binom{n}{2}} \quad 3.7$$

The cosine similarity between residuals is the dot product of the two vectors divided by the product of the two vectors' magnitude:

$$c(X_{res}, Y_{res}) = \cos(\theta) = \frac{\boldsymbol{X_{res}} \cdot \boldsymbol{Y_{res}}}{\|\boldsymbol{X_{res}}\| \|\boldsymbol{Y_{res}}\|} \quad 3.8$$

The Jaccard similarity which is a measure of dissimilarity between residuals is the size of the intersection divided by the size of the union:

$$J(X_{res}, Y_{res}) = \frac{|X_{res} \cap Y_{res}|}{|X_{res} \cup Y_{res}|} = \frac{|X_{res} \cap Y_{res}|}{|X_{res}| + |Y_{res}| - |X_{res} \cap Y_{res}|} \quad 3.9$$

3.3 Materials and Methods

3.3.1 Datasets

The data obtained in this chapter includes a diverse range of indicators. It includes 26 mortality categories and 18 age ranges obtained from NOMIS (https://www.nomisweb.co.uk) website. This is a database service run by the University of Durham on behalf of the UK Office of National Statistics (ONS). It also included 14 crime types and 9 property types obtained by UKCrimeStats (https://www.ukcrimestats.com) website. This data is formatted and aligned by UKCrimeStats to regions using data from the UK Home Office and Land Registry and geographic shape files obtained from the Ordinance Survey Boundary Line dataset. UKCrimeStats also provided population and land area. The collective data (67 indicators) explored is aligned to 348 England and Welch regions

consisting of 239 non-metropolitan districts, 36 metropolitan boroughs, 32 London boroughs and 58 unitary authorities.

It is worth noting some data processing operated by NOMIS effecting the mortality and age data. To anonymise regions reporting low or missing values of mortality there are some conditions that slightly distort the data. If there are regions with mortality, ≤ 2 then the following regions will report a 0. Similarly, if regions have 3 or 4 mortalities then these are reported as 5. In addition, age is model adjusted for a particular year based on the most recent census. In this case, during this study this was the 2011 UK census covering the period between 2013-17 captured on 20/03/2019. The collection of data explored in this chapter is available in Table 3.1.

Crime Types		
Anti-Social Behaviour (ASB)	Bike Theft	Burglary
Criminal Damage and Arson (CD & A)	Drugs	Order
Other Crime	Other Theft	Robbery
Shoplifting	Theft from Person	Vehicle Crime
Violent Crime	Weapons	
Property Types		
Detached	Flats	Freehold
Leasehold	New	Old
Semi-detached	Terraced	Other Property
Mortality Types		
Accidents	Bladder Cancer	Brain Cancer
Colon, sigmoid, rectum and anus Cancer (CSR&A)	Gallbladder Cancer	Kidney Cancer
Larynx Cancer	Liver Cancer	Oesophagus Cancer
Ovary Cancer	Pancreas Cancer	Prostate Cancer
Stomach Cancer	Lung Cancer	Uterus Cancer
Bone Cancer	Breast Cancer	Lymphoid Cancer
Skin Cancer	Cardiac arrhythmias	Cardiomyopathy
Dementia	Diabetes	Suicide
Homicide	Other Mortality	
Age Categories		
Aged 0-4	Aged 5-9	Aged 10-14
Aged 15-19	Aged 20-24	Aged 25-29
Aged 30-34	Aged 35-39	Aged 40-44
Aged 45-49	Aged 50-54	Aged 55-59
Aged 60-64	Aged 65-69	Aged 70-74
Aged 75-79	Aged 80-84	Aged 85+

Table 3.1. Comprehensive list of indicators studied. Sixty-seven indicators were studied: 14 indicators of crime, 9 indicators of property, 26 indicators of mortality and 18 indicators of age. Table available in Ref. [38].

3.3.2 Statistical Analysis

The data were analysed using the statistical software R version (3.6.2) [86] with the Segmented (0.5-3.0) [87–90], proxy (0.4-2.3) [91], boot (1.3-2.4) [65,92], kohonen (3.0.1) [93,94], factoextra (1.0.6) [95], moments (0.14) [96], gplots (3.0.3) [97], ggplots (3.3.1) [98], car (3.0-8) [99], nortest (1.0-4) [100], RColorbrewer (1.1-2) [101], NbClust (3.0) [102], tidyverse (1.3.0) [103], cowplot (1.0.0) [104], psych (1.9.12.31) [105], sf (0.8-1) [106], raster (3.0-12) [107], dplyr (0.8.3) [108], spData (0.3.3) [109], tmap (2.3-2) [110], leaflet (2.0.3) [111], mapview (2.7.0) [112], shiny (1.4.0.2) [113] and png (0.1-7)

[114] packages. These R packages are identical to the ones used in Ref. [38]. The network were constructed using Gephi version (0.9.2) [115].

3.4 Results and Discussions

3.4.1 Overview of Regions

In total, 348 regions in England and Wales were explored consisting of 239 non-metropolitan districts, 36 metropolitan boroughs, 32 London boroughs and 58 unitary authorities ranging from 289 ha (City of London, England) up to 518,037 (Powys, Wales) and have populations between 2158 (Isles of Scilly, England) to 1,070,912 (Birmingham, England). Population densities vary from 0.25 people per hectare (p/ha) (Eden, Cumbria, England) up to 139 p/ha (Islington, England). These regions are the same regions described in Ref. [38]. The wide range or rural and urban regions ensured that all environment types were explored in this study. An overview of area (a), population (b) and population density (c) are provided in figure 3.4.

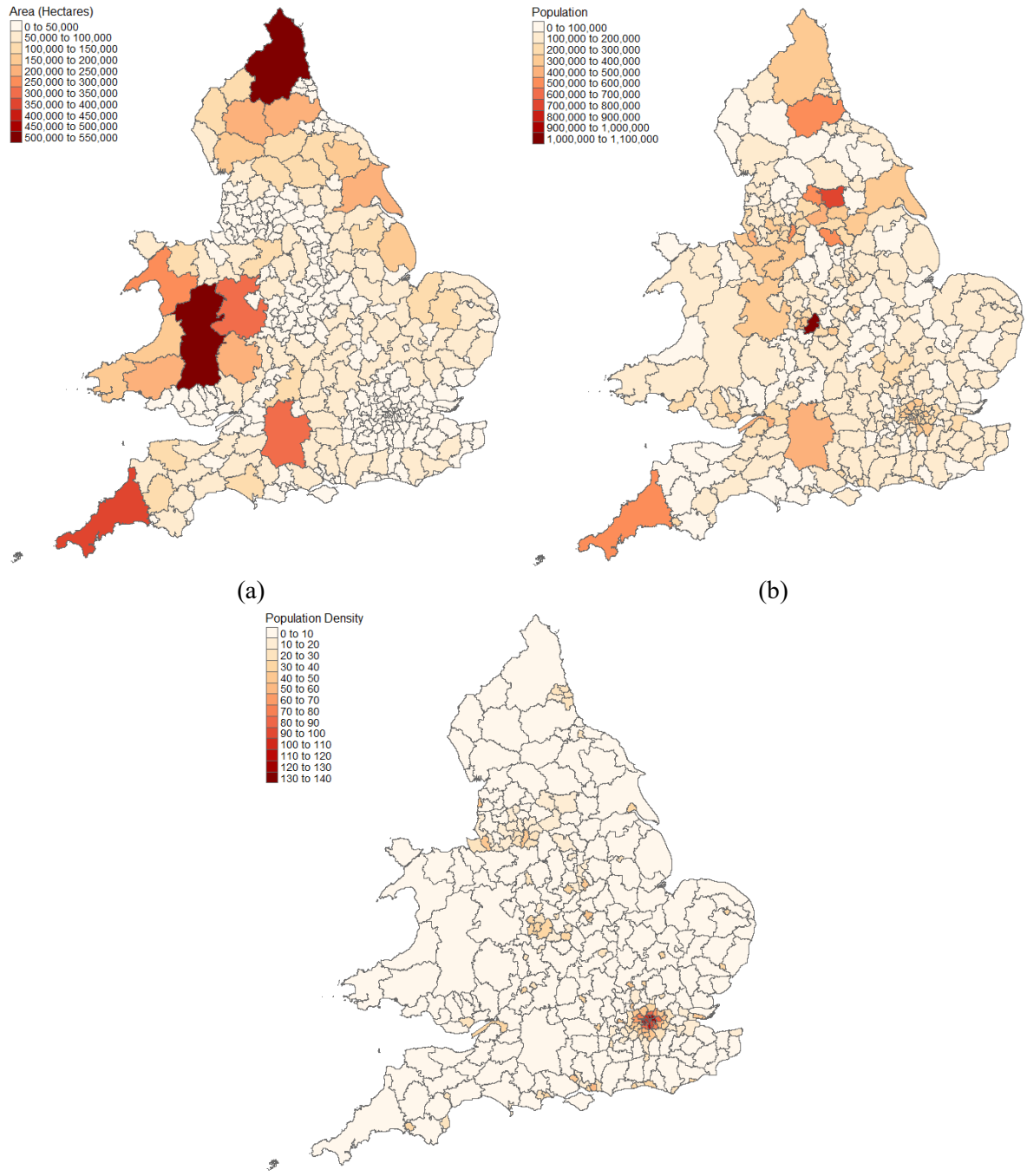


Fig 3.4. Map of England and Wales consisting of unitary authorities, non-metropolitan districts, metropolitan boroughs and London boroughs. The shades of red indicate (a) Area (Hectares), (b) Population and (c) Population Density. The darker the red the higher the value.

Crime, property, mortality and age metrics were fitted to either the single (equation 2.3) or segmented (equation 2.4) PLs depending on Davies test along with AIC and BIC scores (Fig 3.5, Appendices A2-A4). All regions distributed closely around the expectation of PLs except the City of London which deviates from the PL in 23 separate metrics. The deviation is some cases is so extreme

(e.g., Fig 3.5) this particular region was interrogated in isolation to better understand why the City of London is an extreme outlier. It is the smallest region explored in this chapter with an area of 289-hectare and a resident population of 7355. Strikingly, it has a much higher day time population which can exceed as much as 350,000 people. It is worth noting that in some scaling studies it has been documented that PLs change depending on the meaning of population such as whether it is defined as a resident or floating population [116]. In the scaling plots, the City of London is an extreme outlier for most crime metrics where it tended to positively deviate from the PL. On the other hand, the City of London is an extreme outlier for dementia where it negatively deviates from the PL. The negative deviation in dementia in the City of London can be explained by the age demographic since it has a much lower older population. However, the magnitude of deviation is almost a factor of 10 below the PL making this region exceptional. Future studies of dementia risk should consider these findings and investigate the City of London in more detail to better understand the factors involved in the reduction of dementia mortality.

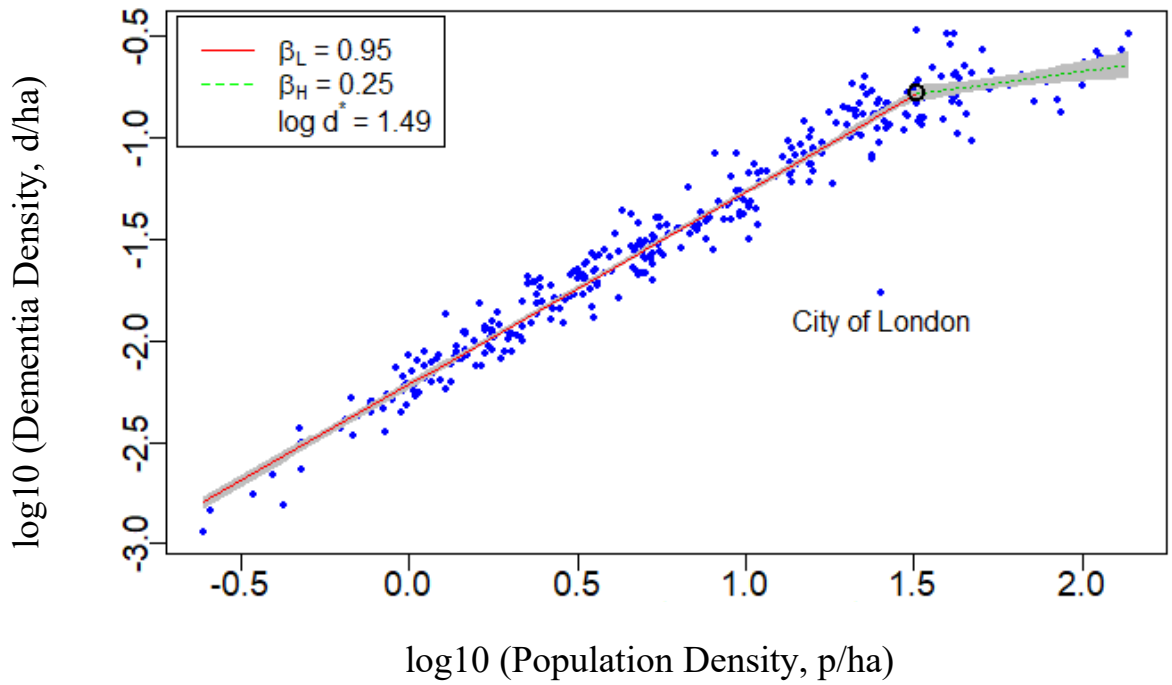


Fig 3.5. Scaling plot between dementia and Alzheimer's disease and total population densities. The red solid line and the dotted green line represent the scaling relationship before and after the change point respectively. The black circle is the change point. The city of London is shown to have exceptionally low incidence of Dementia. This figure was published in Ref. [38].

3.4.2 Rural-urban scaling

Crime, property, mortality and age scaling exponents obtained from either single (grey) or segmented PL models (red and blue) appear in figure 3.6. In the case of segmentation, some metrics decline in highly dense regions (e.g., shop lifting) and others accelerate (e.g., Drugs).

In general, crime and property exponents (Figs 3.6(a, b); Appendix A2-4) are similar to previous work [38]. In 8 metrics a single PL (equation 2.3) was the preferred model: ASB, Burglary, Vehicle Crime, Violent Crime, Other Crime, Bike Theft, Weapons and Order. Remaining crimes exhibited a segmented relationship and thus the double PL (equation 2.4) was the preferred model: Drugs, Other Theft, CD&A, Shoplifting, Theft from the Person and Robbery. Drugs, Other Theft, Theft from the Person and Robbery accelerated whilst Shoplifting and CD&A declined above a critical population density. The heterogeneity of behaviour in crime challenges the crime opportunity [117,118] and the situational actions theories [119,120]. Crimes that follow a single PL suggest a uniformly increasing opportunities or criminogenic settings with population density. A clearer understanding of both acceleration and decline above a critical density needs further attention to understand what these opportunities and criminogenic setting might represent. For example, this will help understand the conditions within high density regions that seemingly present better opportunities for crimes such as theft from person and to a slightly lesser extent robbery. The notion of designed environments [121] is careful planning usually deployed in cities aimed at reducing crime. A subset of crimes where preference is towards a single PL model (e.g., Burglary) challenges the designed environment concept since it is shown that certain crime increase with scale.

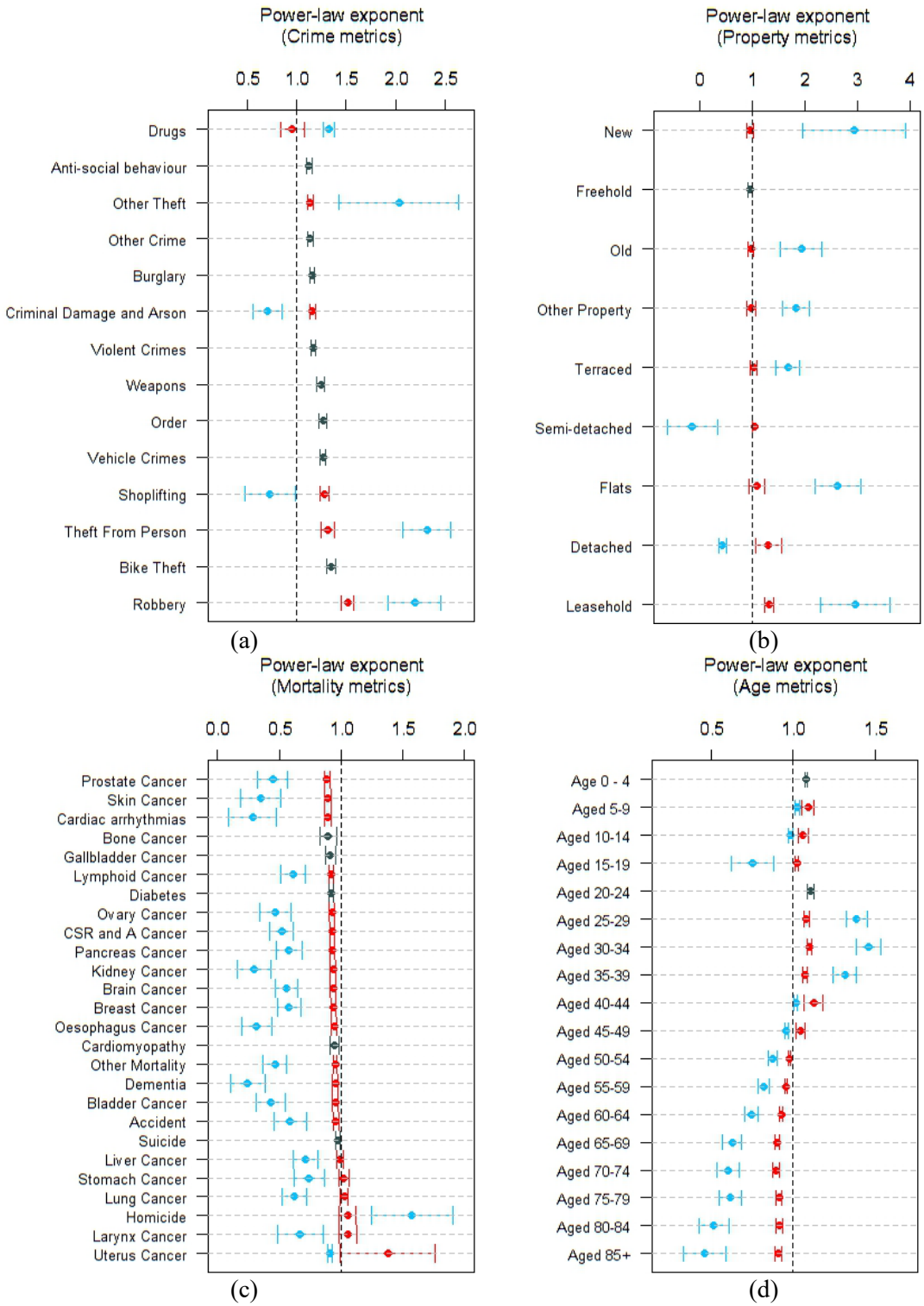


Fig 3.6. Allometric scaling exponents for crime (a), property transactions (b), mortality (c) and age (d) using density metrics. The black symbol represents the exponent obtained from a single PL model. Red and blue symbols represent exponents obtained from a segmenting PL model where red is the exponent below the change point and blue is the exponent above the change point. Error bars represent the 95% confidence intervals for β, β_L, β_H based on the standard errors of regression ($1.96 *$ standard error). The vertical black dotted line indicates linear scaling. This figure was published in Ref. [38].

Almost all mortality metrics (Figs. 3.6(c); Appendix A2-4) exhibited linear to sub-linear scaling except for 7 different cancer types (liver, stomach, lung, larynx and uterine cancer), homicide, diabetes, cardiomyopathy and suicide. Linear scaling occurred in bone cancer, gallbladder cancer, diabetes, cardiomyopathy and suicide. Homicide and uterus cancer was the only mortality type to exhibit linear to super-linear scaling. To better understand the improved mortality in high density regions the scaling of the age groups is investigated to see whether this produces the 'protective' effect.

Examination of the 18 different age groups (ranging from 0-4 to 85+) shows that population density has a profound influence on age demographics (Fig 3.6(d); Appendix A2-4). The ordered age groups exbibits significant structure revealing that young people aged 25-39 accelerate in highly dense regions whilst people aged 45 and above preferentially leave. The strength of the super linear behaviour above the critical population density seen in young people ($\beta_H = 1.46$ for people aged 30-34) could be a leading factor that possibly explains almost all reported super-linear economic indicators [21,22,122]. This can be seen when observing robbery and restricting total population to the 30-34 age group (Fig 3.7; Appendices A5-A6). The acceleration in highly dense regions disappears and a critical population density is no longer present. Instead, robbery is defined as a single PL with a high degree of super linear behaviour throughout all rural-urban regions ($\beta = 1.42$). Furthermore, this effect can be seen in mortality where exponents can be seen to significantly change when targeting certain age groups opposed to the general population. For example, kidney cancer and dementia when observing for the total population, show that highly dense regions exhibit sub-linear scaling (Fig 3.5 and 3.6; Appendix A2-A4). Other variation of restricting age groups opposed to conventional total population were explored and available in Appendices A5-A6. The findings suggest that age groups break the self-similarity behaviour underpinning the urban scaling hypothesis and that age group do not exhibit the same scaling behaviour.

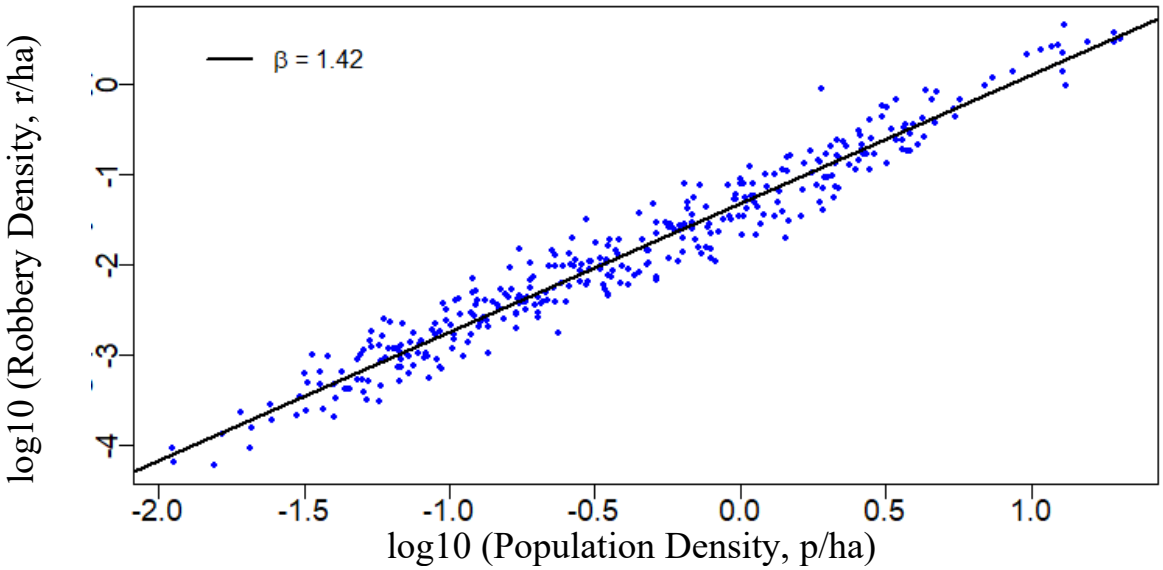


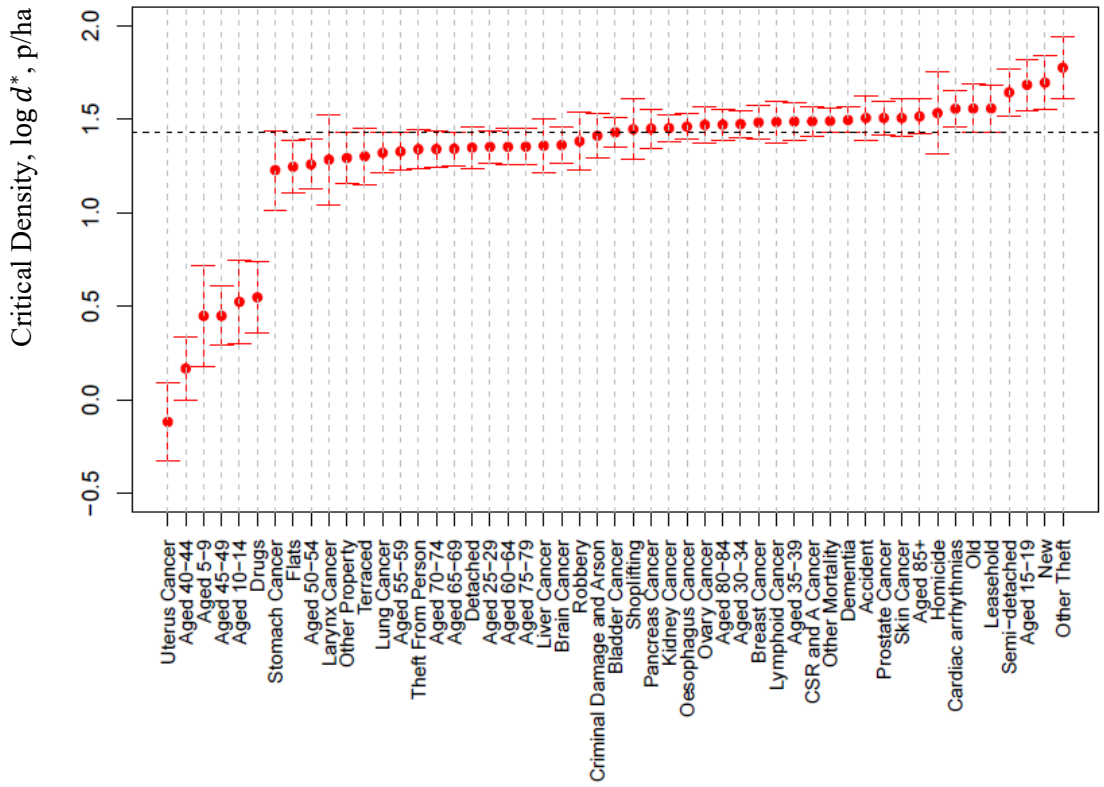
Fig 3.7. Scaling plot between Robbery and the age group 30 to 34 years old. Conventional scaling use total population. The acceleration seen in the total population in high density previous seen for Robbery is no longer present. This figure was published in Ref. [38].

These results are significant from a policy point of view and suggest that some current resource allocation is not fit for purpose and needs reconsideration. As noted previously, formula, applied to the UK National Health Service funding is done through a *per-capita* model [49,123,124] which evidently is not always appropriate for many health related metrics. The *per-capita* approach lacks an appreciation of the deceleration of care costs in urban regions where the age demographic is much lower. Thus, there is not an appreciation of the diseconomies of scale in low population regions (rural). This rural-urban divide exhibited in mortality is an underappreciated feature in the scaling literature.

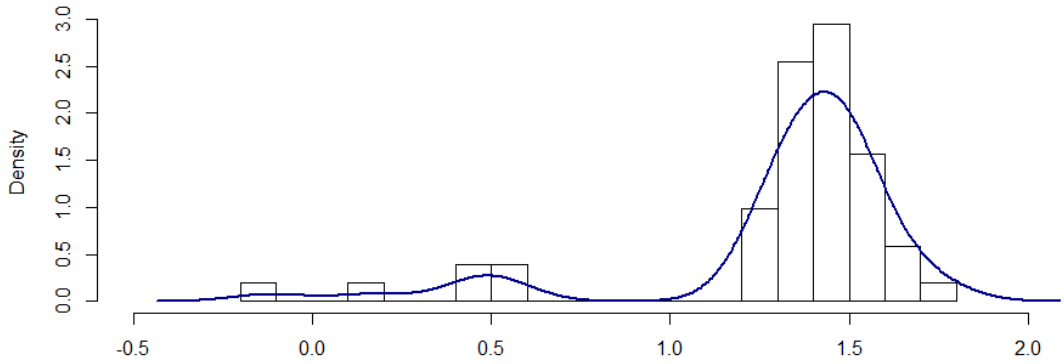
3.4.3 Critical Densities

A segmented PL (equation 2.4) is the preferred model in fifty-one of the metrics examined in this study (6 crime, 8 property, 21 mortality and 16 age) out of a possible sixty-seven. The critical population density distributed around a median of 27 p/ha (Fig 3.8) similar to a previous study [26] where it was documented that the average critical population density for 19 crimes and property was 30 p/ha. Strikingly, almost all critical population densities fit a single distribution (Fig 3.8(b)) despite the diverse range of metrics including crime, property, mortality and age. The few metrics that do not

conform to the single distribution are four different age groups (aged 5-9, aged 10-14, aged 40-44 and aged 45-49) along with uterus cancer and drugs which are located left to the main distribution. Interestingly, the age range 40-49 is the boundary between young people that are super-linearly attracted to highly dense regions and older people that exhibit sub-linear behaviour and thus preferentially seem to leave. This provides a reason why 40- 44 and 45-49 age groups have a significantly different critical population density compared most metrics. The difference in the remaining two age groups (aged 5-9 and aged 10-14), mortality (uterus cancer) crime types (drugs) is still an unexplained observation. In addition, there is no clear explanation of the remaining 45 metrics with a critical population density that forms a unimodal distribution and why it pivots consistently around the same position (27 p/ha). To understand the persisting change point and the mechanics that cause it, percolation theory may offer some insight [125]. These techniques have been applied to the clustering of regions [126] to better understand regional variation. Overall, it is unclear whether there is a unifying statistical mechanics that predicts human behaviour (crimes), health (mortality) economics (property transaction values) and age demographics at a critical density.



(a)



(b)

Fig 3.8. The critical densities (change point) in models with segmenting behaviour. In panel (a) critical densities are the ranked ordered values from smallest to biggest. The horizontal dashed black line is the median. In panel (b) the histogram of critical densities form a bimodal distribution. This figure was published in Ref. [38].

3.4.4 Correlation and hierarchical clustering of residuals by category

Residuals using equation 2.5 were obtained from the preferred model (either equations 2.3 for single or equation 2.4 for segmented relationships) for each metric. This is a simple difference between the

actual observed data and the expectation given in the PL model. All pairs of residuals for all 67 metrics were assessed for correlation and clusters (Appendix A7). The correlation using Pearson's correlation coefficient and hierarchical clustering of residuals (Fig 3.9) separated metrics into three main clusters.

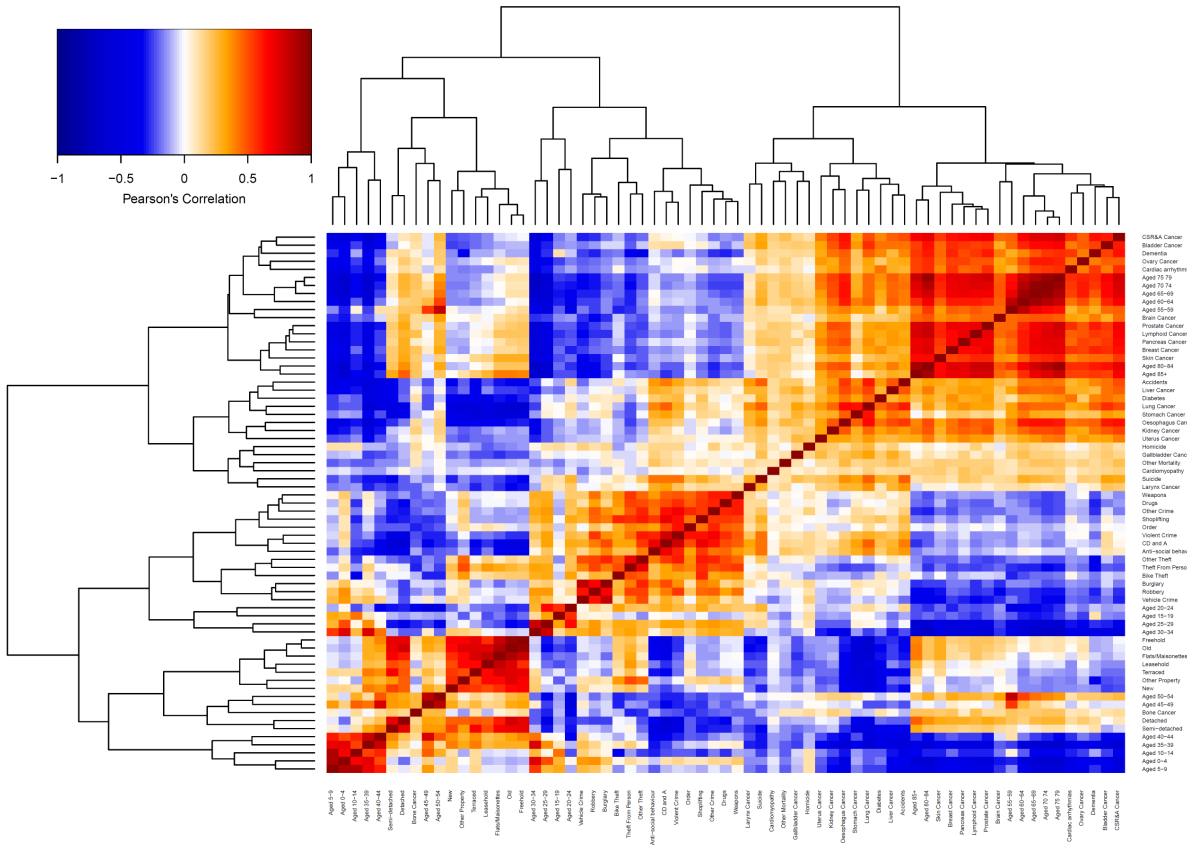


Fig 3.9. Indicator clustering and heatmap between all pairs of indicators. The colour in the heatmap represents the Pearson's correlation coefficient ($\rho_{i,j}$) between all pairs (i and j) of indicators. The red represents positive correlation and blue represents negative correlation. The darker the colour the stronger the correlation. Hierarchical cluster performed and results of these are in the upper and left panels. This figure was published in Ref. [38].

The biggest cluster consisted of mortality and old age groups (Fig 3.9; Appendix A7). Correlations were above 0.5 with values reaching as high 0.72 (e.g., Fig A7.2 in Appendix A7: Lymphoid Cancer vs. Prostate Cancer) with the exception of bone cancer, larynx cancer and homicide. Mortality, however, separated into two clusters. A subset of mortalities (CSR&A cancer, bladder cancer, dementia, ovary cancer, cardiac arrhythmias, brain cancer, prostate cancer, lymphoid

cancer, pancreas cancer, breast cancer and skin cancer) clustered with older people (above 55). This subset of mortalities and older people also tended to be positively correlated with most property types. The second subset of mortalities did not cluster with any other type of indicator. It consisted of Accidents, Liver cancer, diabetes, lung cancer, stomach cancer, oesophagus cancer, kidney cancer, uterus cancer, homicide, suicide and larynx cancer. This subset of mortalities has some degree of preventable cause of death where most are connected to certain lifestyle choices (smoking, obesity, alcohol etc). Contrary to the other group of mortalities, this group of preventable deaths exhibited anti-correlation with property (Fig A7.7 in appendix A7: Flat Residuals vs. Lung Cancer Residuals) meaning that above expectation in property value resulting in below expectation in preventable deaths and vice versa. The protective effect of the property transaction value above the scaling expectation extends to a wide range of mortalities including homicide and suicide. These conclusions were generally reinforced using a range of other correlation measures (Appendix A1: Kendall and Spearman's rank correlation coefficient and appendix A8. Other similarity measures were explored (Appendix A1: Cosine and e-jaccard similarity) which were less informative.

The second cluster consisting of crimes and younger people (aged 15-34) also exhibited positive correlation with Pearson's correlation reaching as high as 0.83 (e.g., Appendix A7: Burglary vs. Vehicle Crime). The third cluster consisting of property types along with the very young and the middle-aged groups also exhibited very strong correlation such as Freehold vs. Old properties (Fig A7.5 in Appendix A7) which had a Pearson's correlation coefficient = 0.95. However, this cluster, in comparison to others, exhibited far more anti-correlation. This occurred between property types and very young age groups (aged 0-14).

The pairwise structure in the heatmap is a limitation as it considers and displays all correlation despite whether it is statistically significant or not. To account for this limitation, the bootstrapped (section 3.2.1) Pearson's correlation coefficients between all pairs of residuals with 2000 replications were used to identify correlation significant at 99% confidence (section 3.2.2). Correlation meeting this condition are presented as two networks: one network for all positive

connections and another for all negative connections (see section 3.2.2 for methods). The resulting positive network (Fig 3.10) has 66 nodes with 784 significant connections connecting them (out of 2211 possible connections). The network excludes bone cancer which was the only metric to have no statistically significant correlation of residuals with any other metric. The optimal modularity (see section 3.2.2.1) applying the Girvan-Newman algorithm [78] to detect community structure was 0.472. This algorithm partitioned the network into 3 hubs. The hubs are similar to the clusters identified in the heatmap (Fig 3.9). These are elderly and mortality; children, middle-aged people and property; young people and crime. Suicide and larynx cancer are an exception as this subset of mortalities clustered with the young adults and crime. This result was also confirmed in the clustering analysis (Fig 3.9). Studies of larynx cancer have found links with alcohol [127] and smoking [128]. The associations between Suicide and Larynx cancer along with positive correlation with several anti-social crimes and preventable deaths (e.g. ASB, CD&A, violence, accidents, diabetes, liver and lung cancers) suggests that intervention and prevention strategies are needed to focus on health care that will help with mental health [129,130] and community safety [131].

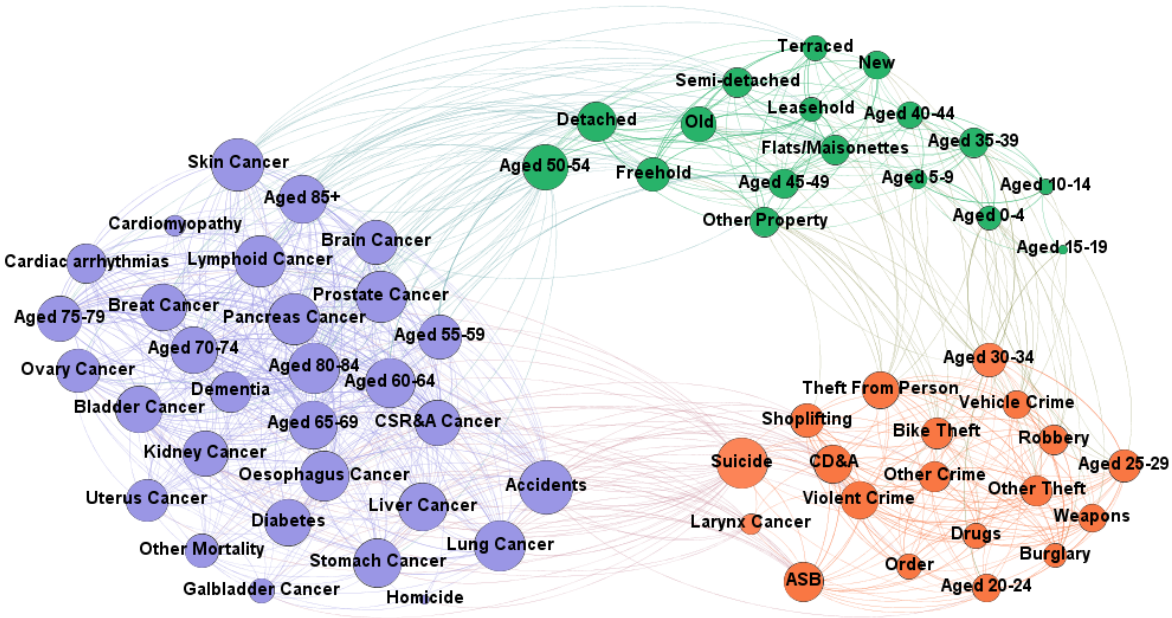


Fig 3.10. A complex network representing significant positive correlation between all residuals. Nodes represent indicators of crime, property, mortality and age. Edges between nodes represent significant positive correlation. The colours of the node represent hubs which is maximized using modularity methods. Edge thicknesses are proportional to the correlation. Node sizes are proportional to their degrees. This figure was published in Ref. [38].

A network of significant anti-correlation (Fig 3.11) has 67 nodes with 604 significant anti-correlation connections (out of 2211 possible connections). The network includes all metrics which means that all metrics have at least one statistically significant connection. The optimal modularity score (0.234) was reduced in comparison to the positive network (Fig 3.10) which partitioned the network into 3 communities. The orange community in figure 3.11 includes all property types along with several mortalities: oesophagus cancer, liver cancer, accidents, diabetes, lung cancer, stomach cancer, homicide, gallbladder cancer, suicide and larynx cancer. This result was also seen in the heatmap (Fig 3.9). It reveals that if property exceeds the scaling law expectation, then deaths such as homicide, accidents and suicide are below the scaling law expectation. Specifically, higher transaction of property occurring in regions are less affected by some deaths. Many of these deaths (e.g., diabetes, suicide, lung cancer) have been linked to certain lifestyle choices and associations with environmental factors such as smoking and poor mental health. This suggests a need for targeted approach so that the most affected regions such as those with a deflated property economy are supported with additional resources (i.e., mental health support).

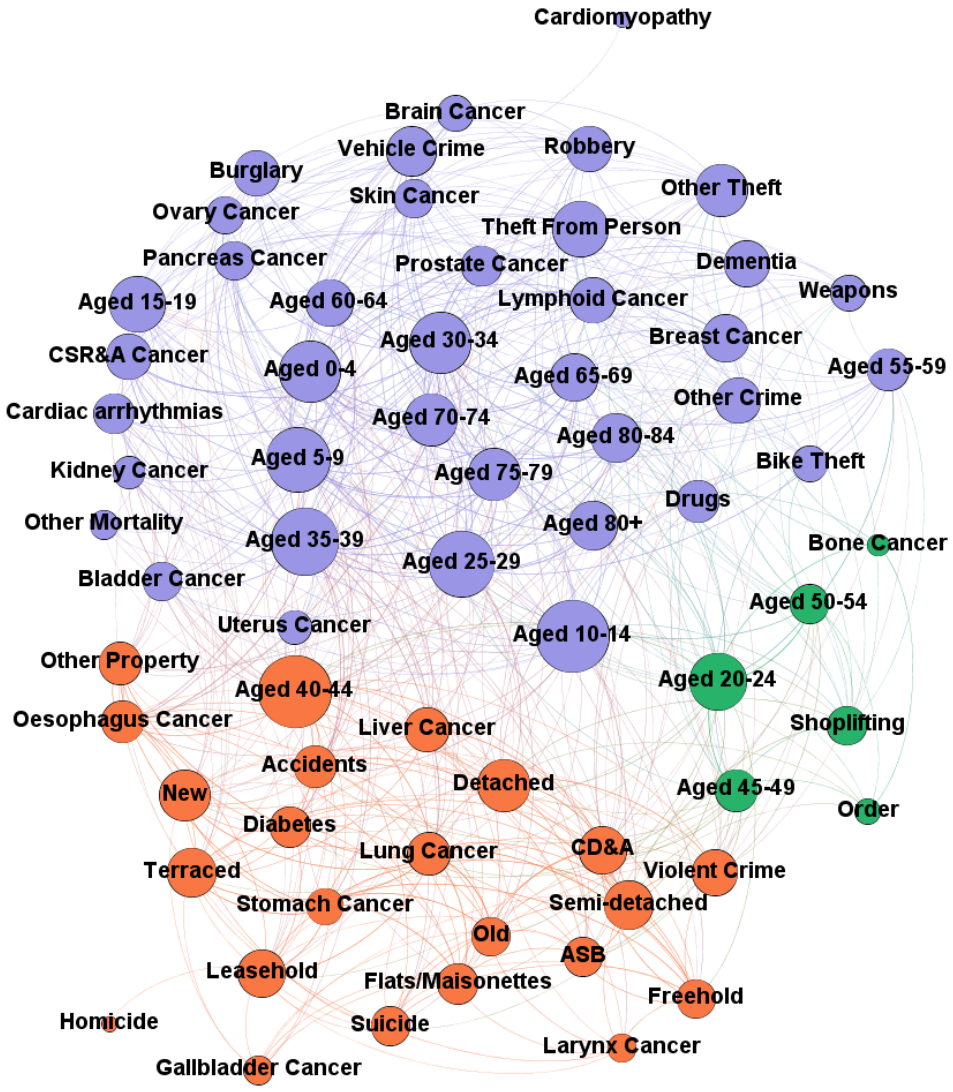


Fig 3.11. A complex network representing significant negative correlation between all residuals. Nodes represent indicators of crime, property, mortality and age. Edges between nodes represent significant negative correlation. The colours of the node represent hubs which is maximized using modularity methods. Edge thicknesses are proportional to the correlation. Node sizes are proportional to their degrees. This figure was published in Ref. [38].

3.4.5 Analysis of residuals by region

To further understand regional differences, the indicator residual matrix was transposed to obtain a regional residual matrix. This presented as a regional heatmap (Fig 3.12) opposed to the indicator heatmap (Fig 3.9) to access correlation and hierarchical clustering between all pairs of regions. This broadly clustered the regions into two main groups with universal anti-correlation at the two extreme ends. If a region in one group has excess mortality and crime above the PL, then a region in the other

group is below the PL. A geomap of the two main clusters is available in appendix A9 which divided England and Wales into two large whole clusters with some exceptions. Generally, with a few exceptions, the map shows that North England, Wales and the Midlands regions cluster together and Southern England regions cluster together.

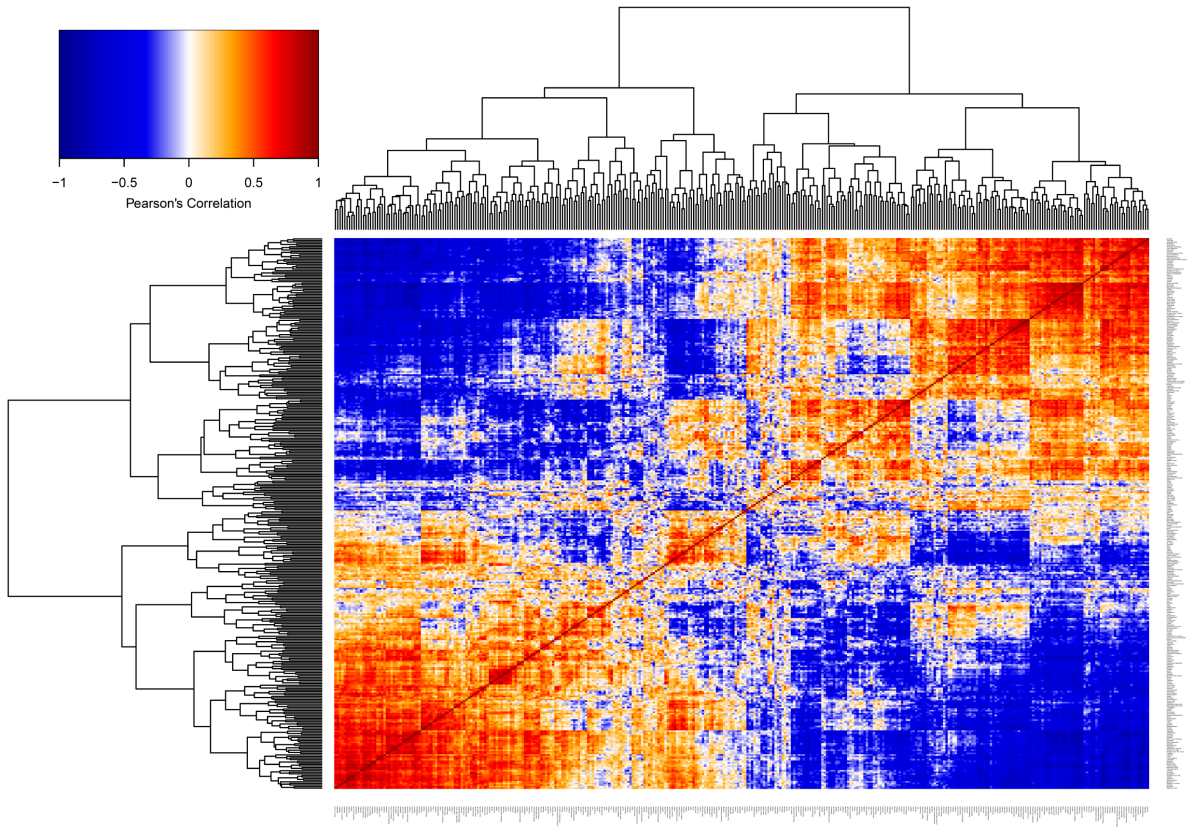


Fig 3.12. Regional heatmap and clustering. Format and colour coding are the same as figure 3.9. The hierarchical cluster divides England Wales into two separate divisions. A geomap of the two divisions is available in Appendix A9. This figure was published in Ref. [38].

3.4.6 Self-organizing maps

The heterogeneity in the clusters presented in the heatmap (Fig 3.12) does not provide sufficient understanding of the specific types of indicators that are most effecting the divide and the role of age demographics is also not clear. To explore this further, a SOM was constructed (section 3.2.3). After completion, the 348 regions distributed onto an 8 by 8 hexagonal self-organising map (Appendix A9). As described in the methods, nodes in the self-organising map are influenced by similar weights in neighbouring nodes. To account for this, clustering algorithms were sampled including hierarchical clustering, total within sum of squares and the gap statistic to expose any potential clusters in the

SOM (Appendix A9). Broadly, the clustering algorithms agreed in separating the SOM into 4 clusters. The results from the gap statistic indicated 4 clusters containing 2, 95, 190 and 61 regions (coloured orange, red, blue and green respectively (Fig 3.13)).

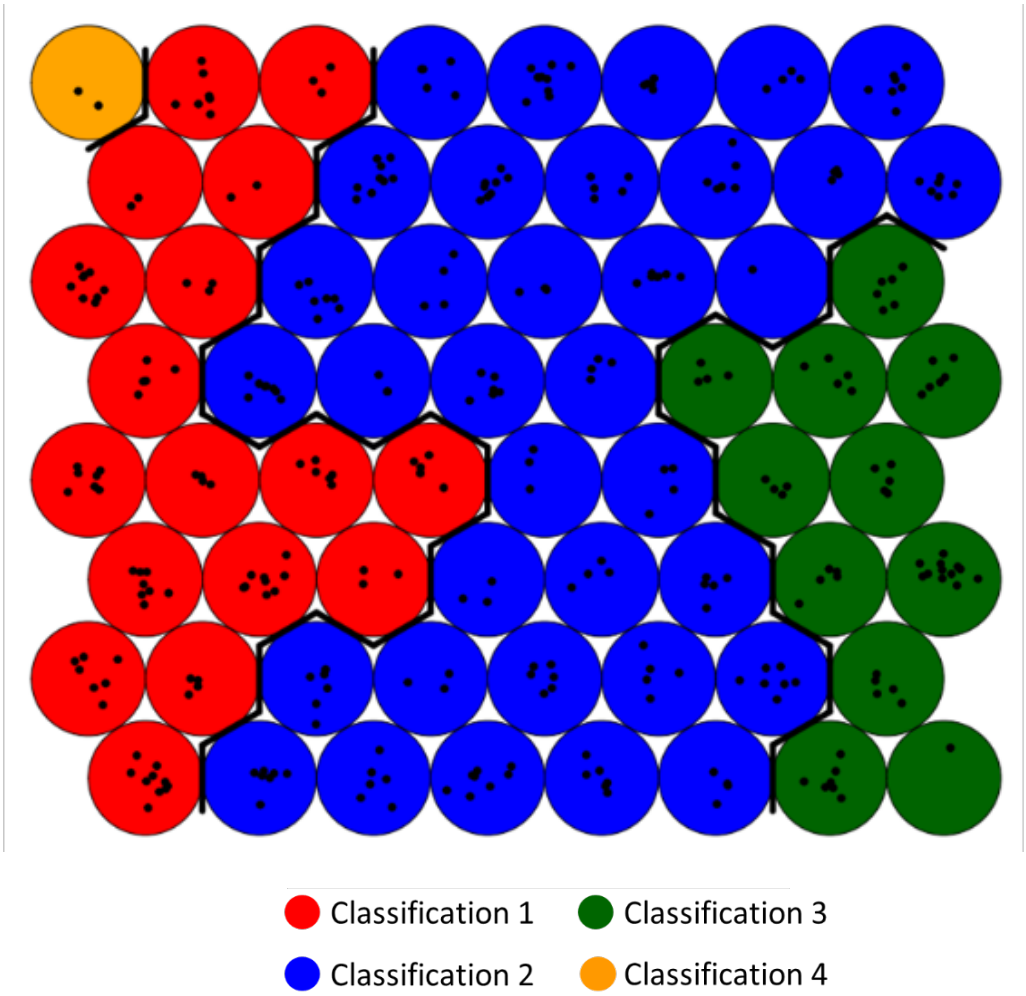
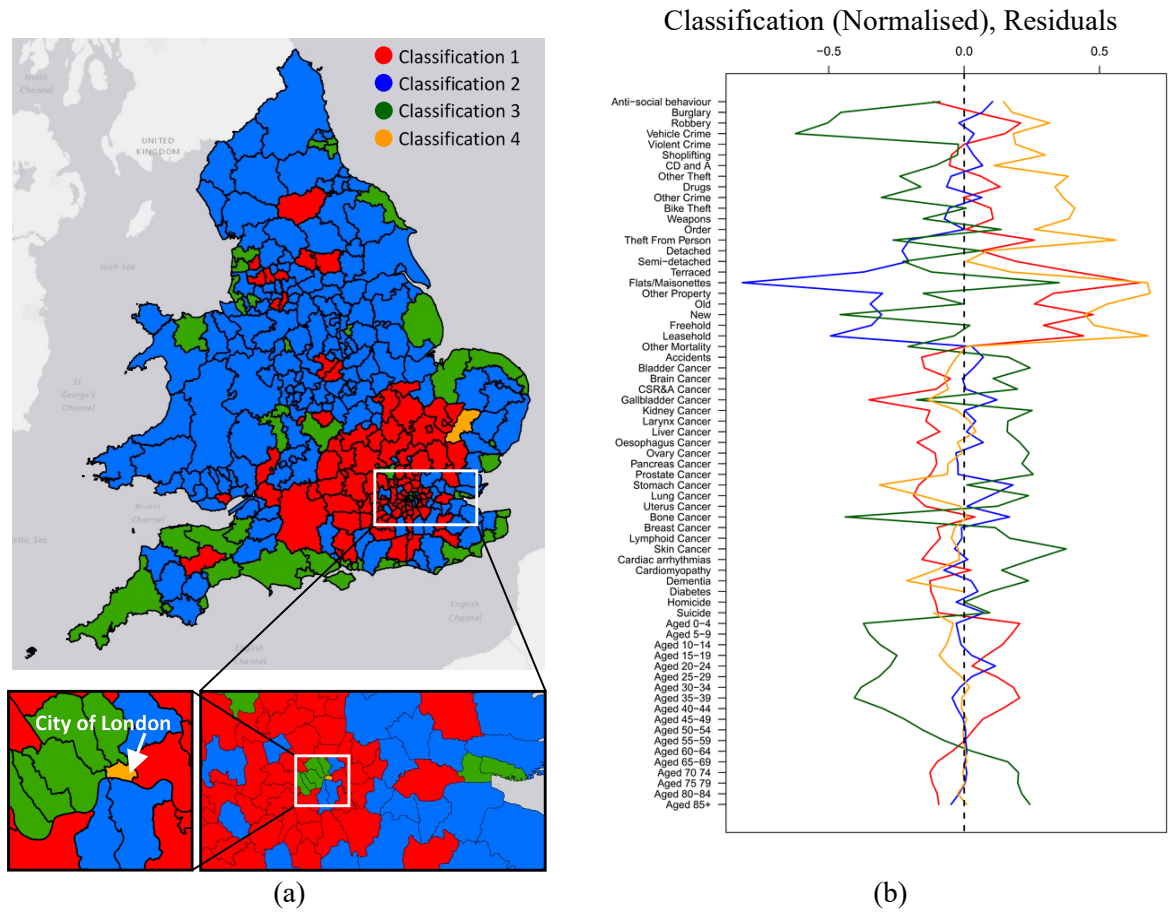


Fig 3.13. Self-organising map (SOM) structure. The SOM had a hexagonal 8 by 8 structure with 4 clusters chosen using the gap statistic. The nodes are orange, red, blue and green which represent clusters 1, 2, 3 and 4 respectively. The black dots within the nodes represent all 348 Unitary Authorities, non-metropolitan districts, metropolitan boroughs and London boroughs. If they are close, they are similar and if they are far apart, they are dissimilar in terms of the residuals. Figure available in Ref. [38].

The clusters in the SOM were identified and represented in a geomap (Fig 3.14; Appendix A9) to better understand regional heterogeneity in England and Wales. The 4 clusters, also represented in Ref. [38] broadly represent: (i) 61 mostly coastal areas (green) with some urban inland exceptions consisting of St. Helens, Stoke-on-Trent, Wyre Forest, Malvern Hills, Strafford-on-Avon,

Dacorum, Ipswich, Kensington and Chelsea, Hammersmith and Fulham, City of Westminster, Islington and Camden; (ii) 2 regions (orange) consisting of City of London and St. Edmundsbury; (iii) 95 regions (red) mostly within the south of England (exceptions are: Richmondshire, Leeds, Bradford, Preston, Chorley, Blackburn with Darwen, Rosendale, Trafford and Manchester); and (iv) 190 more rural regions (blue) primarily in the north of England and Wales.



(a) (b)

Fig 3.14. Map of England and Wales. The colour represents the 4 different clusters identified by the self-organising map. (a) The 348 regions are defined by unitary authorities, non-metropolitan districts, metropolitan boroughs and London boroughs. (b) The mean residuals for each of the clusters are normalized for comparison purposes. The un-normalized versions are available in appendix A9. This figure was published in Ref. [38].

The regions City of London and St. Edmundsbury form the yellow cluster and are incomparable to other regions explored in this study. The average of the residuals in these two regions are extremely low for a large selection of mortalities (exceptions include larynx and liver cancer). They also exhibited extremely high property value and crime types. Neither of the regions had any homicide within the

studied timeframe hence the lack of representation in figure 3.14. To interrogate these exceptional regions further, correlation of indicators between the City of London and St. Edmundsbury were explored. The similarity is shown to be related to high crime and property residuals and low mortality residuals (see Fig A9.5 in Appendix A9). It is worth noting that the methodology of the SOM model is not related to the correlation. It is instead related to the similarity of the residual vectors between regions. The green cluster, representing mostly coastal regions, were characterised by an excess of mortality and an elderly demographics (people aged 60 +). Contrary, the red cluster, representing mostly South of England regions were characterised by younger age demographic and excess property value. Finally, the blue cluster, forming a large part of England and Wales were characterised average and close to the PL expectations for age, crime and mortality indicators whilst property indicators are below the PL expectation.

3.5 Summary

The key novelty in this chapter was exploration of morality, similarity and SOMs. This results in a fundamental challenge to the urban scaling hypothesis. It has previously been thought that urban regions exhibit self-similar behaviour. This study challenges this concept underpinning the urban scaling hypothesis and sets out the problems with *per-capita* models which is the cornerstone of many health-related policies. In addition, conventional urban scaling considers that the people that make up the general population, despite age, are treated equally in the acceleration of life in cities. This study contradicts this hypothesis and is shown that the acceleration depends on the age demographic in urban regions. Although many indicators accelerate in urban regions there are many others that decline. In either case of acceleration or decline, this occurs consistently around a critical population density in the range of 25-30 people per hectare. The position of the critical population density also defines most of the age demographic behaviour between rural-urban regions. Young people accelerate in urbanise environments whilst the elderly preferentially decline. Many scaling studies [21,26,42,132,133] show fundamental bias when applying linear *per-capita* measures. This is also true in the diverse set of non-transmissible disease and age demographics explored in this study. A

widely discussed UK phenomenon is a north-south divide attributed to various inequalities and significant differences relative to a regions north-south location. Although this behaviour is not solely a north-south problem and epidemiologist researching this excess death need to be aware of the limitations and bias of *per-capita* models. The northern regions, forming a large part of the blue regions contributed to the biggest cluster within the SOM. The analysis of these residuals in this group indicated that they did not have exceptional mortality. A *per-capita* framework is skewed in favour of population dense regions. However, in this study cumulative mortalities exhibited economies of scale beyond a critical population density that define the most urbanised regions. This behaviour relates to the scaling of different age groups and the scaling behaviour across all population densities along with the urbanised conditions that provide protection (e.g., dementia).

The consistent and robust critical population density demonstrates that studying rural regions along with urban regions remains a missed opportunity. The existence and consistent position of the critical population density defining a segmented relationship is still an unexplained and important phenomenon that needs further exploration. Analysis of residuals exhibited extensive inter-correlation in a diverse set of metrics relating to crime, property, mortality and age. Residuals appear to distribute around a PL randomly. This is not true and instead they exhibit extensive correlation and persistent structure. The analysis of residuals also provided a clear and more definitive picture of UK variation and an understanding of regional clustering.

3.5.1 Lognormal Method for Crime, Property, Mortality and Age

There are many limitations discussed in scaling when applying traditional linear regression methods. Amongst others, this includes following a set of strict assumptions that if not validated will lead to bias in estimated model parameters [134]. Data in this chapter are indicator densities and log-transformed, thus, it was opted that comparisons of exponents in this chapter are compared to methods proposed by Leitão et al [37] using the lognormal model with fixed ($\omega = 2$) and free ($\omega \, \epsilon [1,3]$)

fluctuations where ω denotes the fluctuation. The lognormal model was applied to all 67 indicators and compared to the least squared method for all reported single (β) and segmented (β_L, β_H) PL model parameters. A summary of these are available in Table 3.2 along with computed bootstrapped confidence intervals and p-values (indicating statistical significance of the scaling) using methods described by Leitão et al [37] and using corresponding Python code [135]. The bold text and asterisk in Table 3.2 indicates a p-value > 0.05 signifying that that the applied model is not rejected. The choice of the fluctuations is a conscious decision such that a fixed fluctuation $\omega = 2$ is generally accepted in scaling whilst the general form $\omega \epsilon [1,3]$ aims to capture different effects such as the variability in human activity and imprecisions in the collection of data [12,37].

Generally, exponents obtained from either least square or Leitão methods agree for crime, mortality and age (except for Theft From Person, Kidney Cancer and Dementia) indicators. This was found to be true if confidence intervals computed in the least squared method overlap with point estimate exponents generated by the Leitão method. The largest discrepancies between methods are found in property indicators (except for Freehold, Terraced and Other Property) and in PL exponents above the critical density (d^*) corresponding to regions with a higher population density. The likely reason for the differences in upper PL (β_H) is that there are not many orders of magnitude of data above the critical density (d^*) representing higher population densities.

To test for statistical significance of the exponents, p-values obtained using Leitão methods for all single (β) and double (β_L, β_H) PL models were computed. Almost all lognormal with fixed and free fluctuations agree when reported as statistically significant. There are 12 mortality (out of a possible 26), 3 crime (out of a possible 14), 5 age categories (out of a possible 18) and 0 property indicators reporting statistical significance (p-value > 0.05) in both lower and upper PL models. Although there are many incidences in segmented models where either one of the lower or upper PL model is statistically significant. This exercise highlights bias in many reported exponents and a need

to consider fluctuations explicitly in future studies when using PL models. Some of these issues highlighted here are addressed later in chapters 5-6. However, the interesting aspect of the consistent critical density still holds and requires further investigation to understand the statistical mechanics that causes this phenomenon.

Indicator	**Model**					
	Least squared method		Leitão lognormal model			
			Fixed fluctuations $\omega = 2$		Free fluctuations $\omega \epsilon[1,3]$	
	β, β_L (Single or Low Density)	β_H (High Density)	β, β_L (Single or Low Density)	β_H (High Density)	β, β_L (Single or Low Density)	β_H (High Density)
Crime Type						
ASB	1.13 (0.03)	-	1.13 (0.03)	-	1.13 (0.03)	-
Bike Theft	1.35 (0.05)	-	1.35 (0.05)	-	1.37 (0.06)	-
Burglary	1.16 (0.02)	-	1.16 (0.02)	-	1.16 (0.03)	-
CD & A	1.16 (0.03)	0.71 (0.15)	**1.16 (0.03)***	**0.72 (0.09)***	**1.17 (0.03)***	**0.74 (0.12) ***
Drugs	0.96 (0.12)	1.33 (0.06)	**0.93 (0.13)***	**1.32 (0.06)***	**0.91 (0.13) ***	1.32 (0.06)
Order	1.27 (0.04)	-	**1.27 (0.03)***	-	1.25 (0.03)	-
Other Crime	1.14 (0.03)	-	1.14 (0.02)		1.16 (0.04)	-
Other Theft	1.14 (0.03)	2.04 (0.61)	1.14 (0.03)	2.00 (0.27)	1.15 (0.03)	2.00 (0.42)
Robbery	1.51 (0.06)	2.20 (0.27)	**1.52 (0.06)***	2.00 (0.04)	**1.56 (0.07) ***	2.00 (0.09)
Shoplifting	1.28 (0.04)	0.73 (0.25)	1.28 (0.05)	**0.73 (0.24)***	1.28 (0.06)	**0.73 (0.24) ***
Theft From Person	1.31 (0.06)	2.32 (0.25)	**1.31 (0.06)***	2.00 (0.09)	**1.32 (0.06) ***	2.0 (0.05)
Vehicle Crime	1.27 (0.03)	-	**1.27 (0.02)***	-	**1.26 (0.02) ***	-
Violent Crimes	1.17 (0.02)	-	1.17 (0.02)	-	1.17 (0.02)	-
Weapons	1.24 (0.03)	-	1.24 (0.04)	-	1.24 (0.05)	-
Property Types						
Detached	0.80 (0.08)	-0.27 (0.29)	0.80 (0.50)	0.5 (0.0)	0.61 (0.29)	0.52 (0.10)
Flats	1.08 (0.15)	2.62 (0.44)	1.08 (0.14)	**2.00 (0.17) ***	1.35 (0.35)	2.00 (0.00)
Freehold	0.95 (0.04)	-	0.95 (0.62)	-	0.83 (0.16)	-
Leasehold	1.31 (0.09)	2.94 (0.66)	1.29 (0.09)	2.00 (0.16)	1.38 (0.28)	2.00 (0.00)
New	0.95 (0.06)	2.92 (0.98)	0.95 (0.07)	**2.0 (0.03) ***	0.92 (0.15)	**2.00 (0.09) ***
Old	0.97 (0.05)	1.92 (0.39)	0.97 (0.74)	**1.91 (0.51) ***	0.82 (0.22)	**1.95 (0.30) ***
Semi-detached	1.03 (0.04)	-0.14 (0.49)	1.03 (0.43)	0.5 (0.12)	1.04 (0.16)	0.50 (0.47)
Terraced	1.01 (0.07)	1.67 (0.22)	1.01 (0.11)	**1.67 (0.24) ***	1.02 (0.20)	**1.70 (0.22) ***
Other Property	0.98 (0.08)	1.83 (0.26)	**0.98 (0.07)**	1.85 (0.29)	**1.01 (0.12) ***	1.77 (0.39)
Mortality Type						
Accidents	0.96 (0.03)	0.59 (0.14)	**0.96 (0.02) ***	**0.62 (0.13) ***	**0.96 (0.02) ***	**0.67 (0.15) ***
Bladder Cancer	0.95 (0.02)	0.43 (0.12)	**0.95 (0.02) ***	**0.5 (0.02) ***	**0.95 (0.02) ***	**0.5 (0.01) ***
Brain Cancer	0.94 (0.03)	0.56 (0.09)	**0.94 (0.02) ***	**0.56 (0.09) ***	**0.94 (0.02) ***	**0.56 (0.09) ***
CSR and A	0.93 (0.02)	0.52 (0.10)	**0.93 (0.02) ***	**0.52 (0.07)**	**0.93 (0.02) ***	**0.54 (0.08) ***
Gallbladder Cancer	0.91 (0.04)	-	0.91 (0.04)	-	0.92 (0.05)	-
Kidney Cancer	0.93 (0.02)	0.30 (0.14)	0.93 (0.03)	0.5 (0.00)	0.94 (0.02)	0.5 (0.01)
Larynx Cancer	1.06 (0.07)	0.67 (0.19)	**1.06 (0.07) ***	**0.71 (0.08) ***	**1.06 (0.07) ***	**0.66 (0.21) ***
Liver Cancer	0.99 (0.02)	0.71 (0.10)	**0.99 (0.03) ***	**0.71 (0.08) ***	**0.99 (0.03) ***	**0.71 (0.08) ***
Oesophagus Cancer	0.94 (0.02)	0.32 (0.13)	**0.94 (0.02) ***	0.5 (0.00)	0.95 (0.02)	0.5 (0.00)
Ovary Cancer	0.93 (0.03)	0.47 (0.13)	**0.93 (0.02) ***	**0.50 (0.07) ***	**0.93 (0.02) ***	**0.5 (0.04) ***
Pancreas Cancer	0.93 (0.02)	0.58 (0.10)	0.93 (0.02)	**0.58 (0.13) ***	0.93 (0.02)	**0.58 (0.13) ***

Prostate Cancer	0.89 (0.02)	0.45 (0.12)	**0.89 (0.02) ***	**0.50 (0.08) ***	**0.89 (0.02) ***	**0.50 (0.07) ***
Stomach Cancer	1.02 (0.04)	0.74 (0.12)	1.02 (0.04)	0.74 (0.12)	1.02 (0.04)	0.74 (0.10)
Lung Cancer	1.03 (0.03)	0.62 (0.10)	**1.03 (0.02) ***	0.62 (0.09)	**1.02 (0.03) ***	0.63 (0.11)
Uterus Cancer	1.38 (0.38)	0.91 (0.02)	**1.38 (0.42) ***	0.91 (0.02)	**1.40 (0.61) ***	0.91 (0.02)
Bone Cancer	0.89 (0.06)	-	**0.89 (0.07) ***	-	0.92 (0.11)	
Breast Cancer	0.94 (0.02)	0.58 (0.10)	**0.94 (0.02) ***	**0.58 (0.12) ***	**0.94 (0.02) ***	**0.58 (0.12) ***
Lymphoid Cancer	0.92 (0.02)	0.61 (0.10)	0.92 (0.02)	**0.61 (0.10) ***	0.92 (0.02)	**0.61 (0.10) ***
Skin Cancer	0.89 (0.03)	0.35 (0.17)	0.89 (0.03)	**0.5 (0.02) ***	0.90 (0.03)	**0.5 (0.04) ***
Cardiac arrhythmias	0.89 (0.02)	0.29 (0.20)	**0.89 (0.03) ***	**0.5 (0.05) ***	**0.90 (0.03) ***	**0.50 (0.03) ***
Cardiomyopathy	0.95 (0.04)	-	0.95 (0.04)	-	0.93 (0.04)	-
Dementia	0.95 (0.02)	0.25 (0.15)	0.95 (0.03)	0.50 (0.01)	0.97 (0.03)	0.5 (0.00)
Diabetes	0.92 (0.02)	-	**0.92 (0.02) ***	-	**0.92 (0.02) ***	-
Suicide	0.97 (0.01)	-	0.97 (0.02)	-	**0.98 (0.02) ***	-
Homicide	1.05 (0.06)	1.57 (0.33)	**1.06 (0.08) ***	**1.60 (0.31) ***	**1.06 (0.07) ***	**1.60 (0.37) ***
Other Mortality	0.95 (0.02)	0.47 (0.10)	**0.95 (0.02) ***	0.50 (0.02)	**0.95 (0.02) ***	0.5 (0.03)
Age Categories						
Aged 0-4	1.08 (0.01)	-	**1.08 (0.01)**	-	1.08 (0.01)	-
Aged 5-9	1.09 (0.03)	1.02 (0.01)	**1.09 (0.03)**	1.02 (0.02)	1.09 (0.03)	**1.02 (0.01)**
Aged 10-14	1.06 (0.03)	0.98 (0.01)	**1.06 (0.02)**	0.98 (0.02)	**1.06 (0.02)**	0.98 (0.02)
Aged 15-19	1.02 (0.01)	0.75 (0.13)	**1.06 (0.02)**	**0.75 (0 17)**	1.02 (0.01)	**0.74 (0.18)**
Aged 20-24	1.11 (0.02)	-	1.11 (0.02)	-	1.10 (0.02)	-
Aged 25-29	1.08 (0.02)	1.39 (0.07)	1.08 (0.02)	**1.39 (0.09)**	1.08 (0.02)	**1.39 (0.10)**
Aged 30-34	1.10 (0.02)	1.46 (0.08)	**1.10 (0.01)**	1.46 (0.07)	**1.10 (0.02)**	**1.46 (0.08)**
Aged 35-39	1.07 (0.01)	1.32 (0.08)	**1.07 (0.01)**	**1.32 (0.06)**	**1.07 (0.01)**	**1.29 (0.07)**
Aged 40-44	1.13 (0.06)	1.02 (0.01)	**1.13 (0.07)**	**1.02 (0.01)**	**1.14 (0.09)**	**1.02 (0.01)**
Aged 45-49	1.05 (0.03)	0.96 (0.01)	1.05 (0.03)	0.96 (0.01)	1.05 (0.03)	0.96 (0.01)
Aged 50-54	0.98 (0.01)	0.88 (0.03)	0.98 (0.01)	0.88 (0.03)	0.98 (0.01)	0.87 (0.04)
Aged 55-59	0.96 (0.01)	0.82 (0.03)	0.96 (0.01)	**0.82 (0.04)**	0.96 (0.01)	**0.82 (0.04)**
Aged 60-64	0.93 (0.01)	0.75 (0.04)	0.93 (0.01)	**0.75 (0.06)**	0.93 (0.01)	**0.75 (0.05)**
Aged 65-69	0.90 (0.01)	0.63 (0.06)	0.90 (0.02)	**0.63 (0.08)**	0.90 (0.02)	**0.63 (0.09)**
Aged 70-74	0.90 (0.02)	0.61 (0.07)	0.90 (0.02)	**0.63 (0.09)**	0.90 (0.01)	**0.63 (0.08)**
Aged 75-79	0.92 (0.02)	0.62 (0.07)	0.91 (0.02)	**0.61 (0.09)**	0.91 (0.02)	**0.61 (0.09)**
Aged 80-84	0.92 (0.02)	0.52 (0.09)	0.92 (0.02)	**0.52 (0.07)**	0.92 (0.02)	**0.52 (0.07)**
Aged 85 +	0.91 (0.02)	0.46 (0.13)	0.91 (0.02)	**0.50 (0.06)**	0.91 (0.02)	**0.5 (0.07)**

Table 3.2. Summary of exponents (to 2 dp) using the least squared and Leitão methods [37] **reported for single (β) and double PL (β_L, β_H) models.** The first two columns are exponents, also documented in Fig 3.6, using the least squared method. The last four columns correspond to the Leitão methods using the lognormal model with assumed fixed ($\omega = 2$) and free ($\omega \epsilon [1,3]$) fluctuations where ω denotes the fluctuation. Methods to obtain confidence intervals using bootstrapping techniques and computed *p*-values are described in the Leitão et al paper [37]. *p*-values greater than 0.05 are indicated using an asterisk and bold font. This indicates that the null hypothesis that the data was generated by the model is accepted. The python code used to generate exponents, confidence intervals and p-values in this table using Leitão methods are available in ref [135].

Chapter 4: Population Density and Spreading of COVID-19 in England and Wales

4.1 Introduction

SARS-CoV-2 has been an opportunity to explore the scaling behaviour of a very transmissible disease. SARS-CoV-2 was first identified in China in late 2019 and was later declared a worldwide pandemic on 13 March 2020. It is thought that the main transmittance is through aerosols between people that can be either symptomatic or asymptomatic [136–139]. To deal with the spread of the disease many countries across the world deployed mitigating strategies such as distancing measures such as meeting size restrictions to reduce the number of infectious interactions along with various other strategies such as the use of face coverings. Many governments have based these mitigating decisions on cases, deaths and various healthcare (hospital admittance, patience on mechanical ventilation bed etc) data which at the very beginning of the pandemic was very sparse. At this stage it was hypothesized that population density might be a proxy for infectious interactions. A high population density would mean that there is more opportunity for spread and consequently more opportunity to being infected. For example, someone who lives in a city where there is a high population density is expected to have a greater number of interactions compared with someone who lives in a rural setting [140].

Separately the propagation of COVID-19 has been modelled using the Weibull [141], Poisson [142], Gamma [143] and Normal [144] distributions. The Weibull (Fig 4.1 (a)), Poisson (Fig 4.1 (b))

and Gamma (Fig 4.1 (c)) distributions only allow for positive skew whilst the normal (Fig 4.1 (d)) distribution is entirely symmetrical. Their use may be inherently biased by the inability to show negative skew [41,145–147].

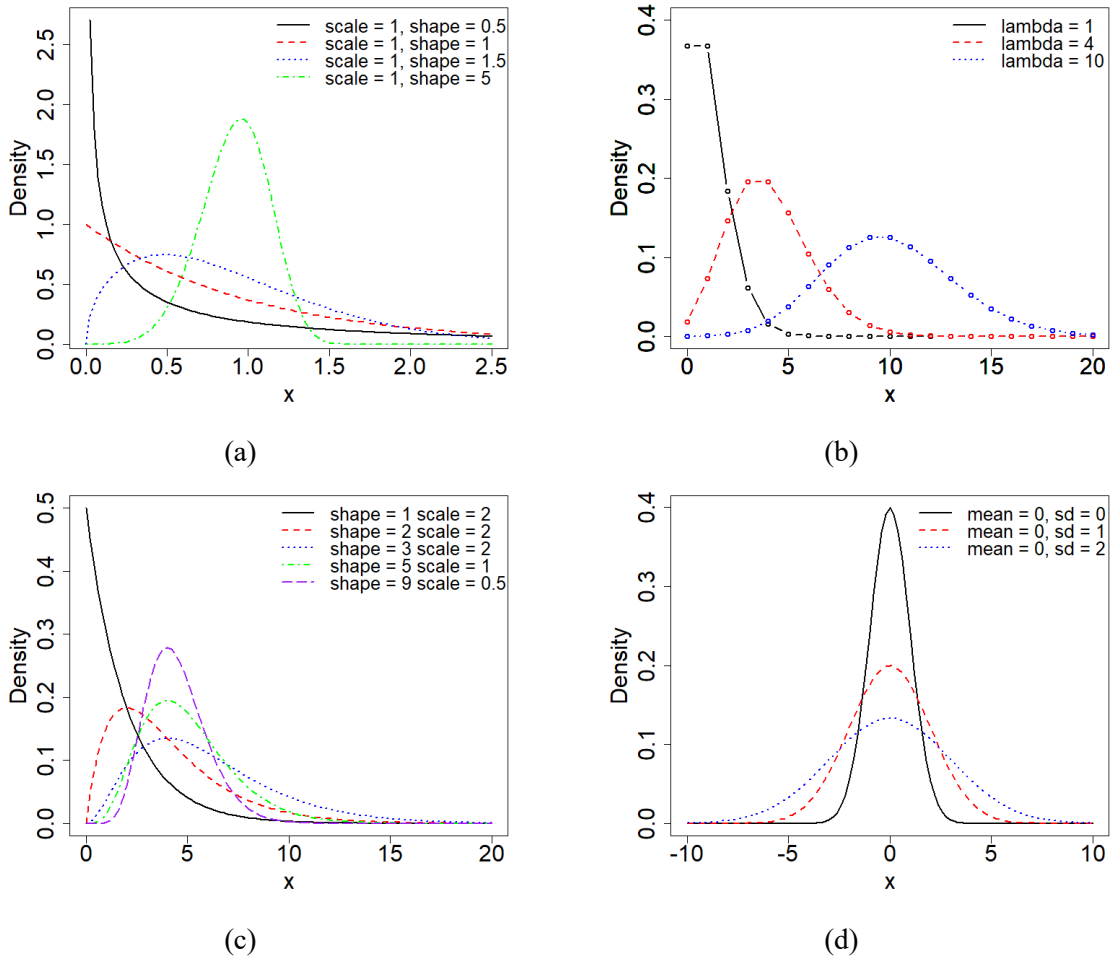


Fig 4.1. Illustrations of the (a) Weibull, (b) Poisson, (c) Gamma and (d) Normal distribution that are used to model the propagation of COVID-19 under various model parameter conditions.

Previous studies of population density using a variety of indicators (health, crime *etc*) have a limited number of time points that are usually aggregated over a month or year [21,26,27,38,133,147,148] whilst finer granularity (e.g. daily, weekly) data are much less common and very difficult to access. Due to the COVID-19 pandemic, Public Health England (https://coronavirus.data.gov.uk/) (PHE) provided a dataset that is updated and made available to the public on a daily basis. This is the first opportunity to explore scaling, spanning across an extended timeframe, of a contagious disease during a worldwide pandemic. The data available has been

paramount to the policy and government interventions in attempts to control the spread of the disease, lessen the burden on health care resources and to reduce excess deaths. Free public access to this data has helped shape public perceptions of the disease and assisted in data-driven informed decisions.

Here the daily scaling of COVID-19 cases and deaths was investigated using population density. England and Wales Lower Tier Local Authorities (LTLAs) were used as regional boundaries to document explore scaling behaviour along with residual variance and skew to explore how residuals that represent regions distribute over an extended time series. This is to provide an understanding how regions distribute during different times of the pandemic and present a general overview of the impact that policy and changing human behaviour has on propagation. This will also confirm whether or not if the Weibull, Poisson, Gamma or Normal distributions are sufficient to model disease. In recognition that age is an important consideration in the pandemic age categories ranging from 0-4 years old to 85+ years old were included to explore whether age demographics were similar to previous scaling of general non-transmissible disease [38].

4.2 Theory

In chapter 2, single (equation 2.3) and segmented (equation 2.4) PL models were defined. Residuals obtained using these models are a simple difference between the observed and estimated data (equation 2.5) and can be considered to be scale adjusted. Throughout the pandemic, cases and mortality are well described by a single PL for each day. As noted in chapter 2, some age indicators are better defined by a segmented fit. The following section defines some similarity measures and distributions used to the analysis of the residuals.

4.2.1 Similarity Measures

There exist several different measures to access the correlation and similarity between complete paired data (e.g., Kendall, Pearson, Spearman) (Appendix A1). The type of similarity measure is an important consideration and can depend on the structure of the data and whether they contain outliers.

For example, Kendall rank correlation coefficient [149] is used to measure the ordinal association between two measured quantities, Pearson's correlation assesses linearity, whilst Spearman correlation assesses monotonic relationships: whether that is linear or not [150]. To demonstrate the different results between Pearson's and Spearman's correlation coefficient, data are simulated to document the correlation of similarity. It is found that linear data (4.2(a)) and random data (4.2(b)) report a similar Spearman's and Pearson's correlation coefficient. Spearman's correlation coefficient is reported as 1 in data with a perfectly monotonically relationship, indicating a perfect fit with slightly reduced Pearson's coefficient (4.2(c)). In data with outliers (Figure 4.2 (d)) Spearman's correlation coefficient is slightly elevated compared to Pearson correlation. This is due to the ordering of the data where Spearman's correlation limits the outlier to the value of its rank, thus, it is less sensitive to outliers.

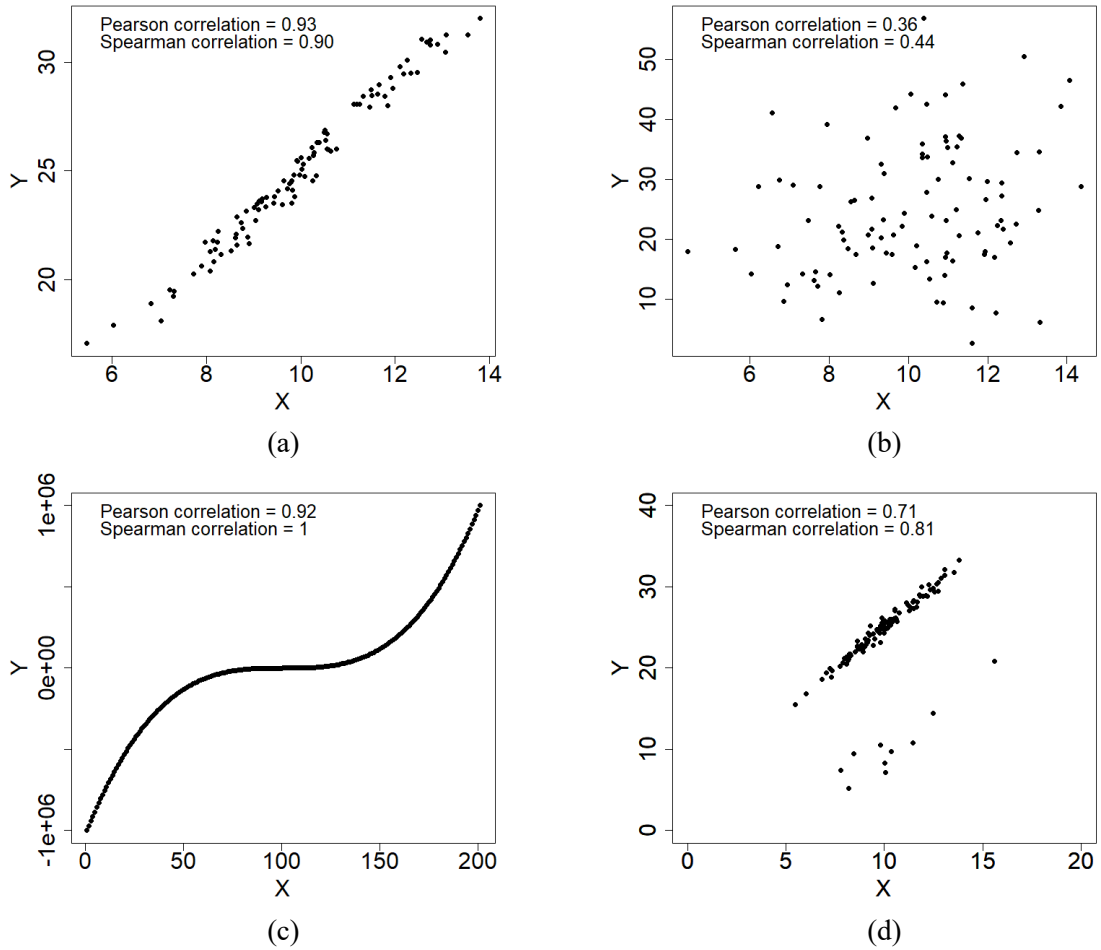


Fig 4.2. Pairwise simulated data. (a) linear Data with small amounts of noise, (b) linear data with large amounts of noise, (c) perfect monotonic data and (d) linear data with a few outliers.

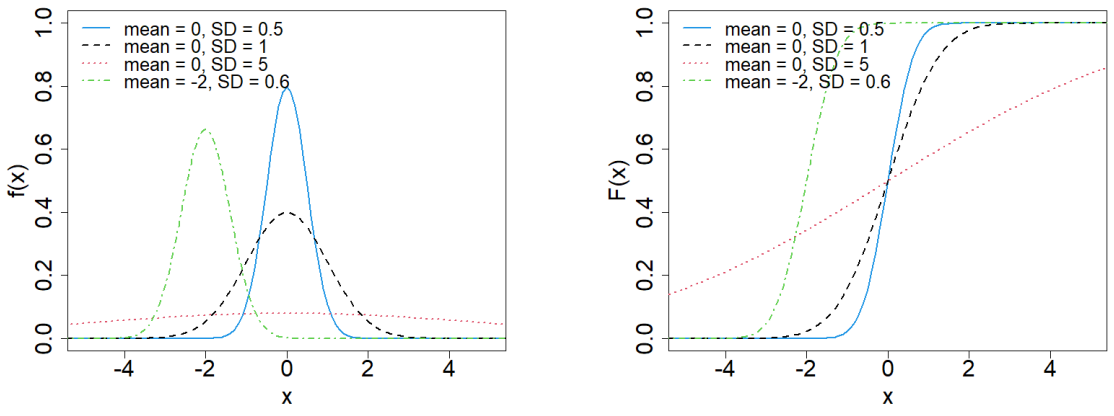
After obtaining residual from the preferred model (equations 2.3 and 2.4), Pearson's correlation and Spearman's rank correlation coefficient was computed between pairs of residuals. Similar to methods in chapter 3 residuals are represented as $X_{res\ i} = (x_{res\ 1}, x_{res\ 2}, \dots, x_{res\ n})$ and $Y_{res\ i} = (y_{res\ 1}, y_{res\ 2}, \dots, y_{res\ n})$ for each complete set of n pairs (X_{res}, Y_{res}) of residuals (Appendix A1).

4.2.2 Residual and Case Density Models

The distribution of residuals was fitted using either a normal (equation 4.3) or generalised logistic (GL) (equation 4.4) distribution. The normal distribution [151] is given by:

$$N(x; \mu, \sigma_{norm}) = \frac{1}{\sigma_{norm}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma_{norm}}\right)^2} \qquad 4.3$$

where μ and σ_{norm} are mean and standard deviation respectively such that $\mu \in \mathbb{R}$ is the location parameter and $\sigma^2_{norm} \in \mathbb{R}_{>0}$ is the squared scale parameter and as illustrated in figure 4.3, is a perfect symmetrical distribution.



(a) (b)

Fig 4.3. The normal distribution. (a) the probability distribution function (PDF) and (b) cumulative distribution function (CDF) with varying mean μ and SD σ parameters.

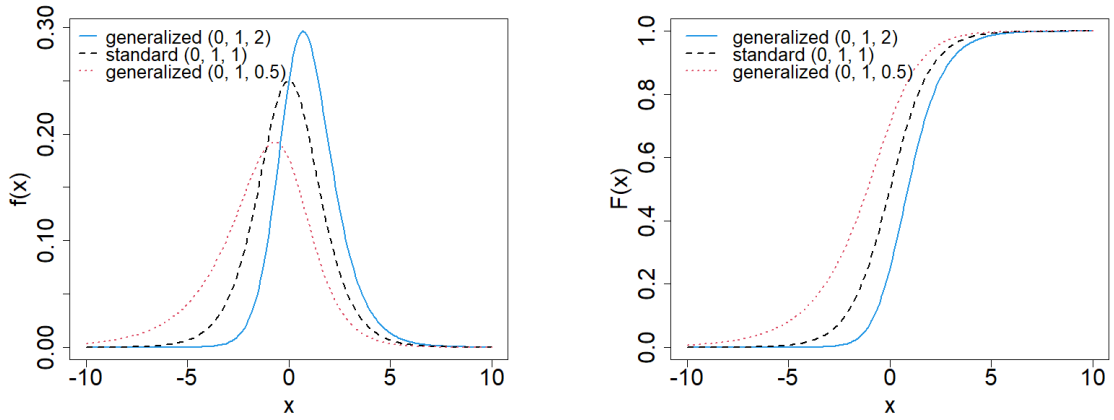
The type I generalised logistic distribution (GLD) [152,153] is given by:

$$GL(x;\theta,\sigma,\alpha) = \frac{\alpha}{\sigma}\frac{e^{-\frac{x-\theta}{\sigma}}}{\left\{1+e^{-\frac{x-\theta}{\sigma}}\right\}^{\alpha+1}} \quad 4.4$$

where θ, σ and α are the location, shape and scale parameters such that $\theta \in \mathbb{R}$, $\alpha > 0$, $\sigma > 0$ and $-\infty < x < +\infty$. Like the normal distribution, the generalised logistic (GL) distribution can be symmetrical when $(\theta, \sigma, \alpha) = (0,1,1)$. This parameterization reduces equation 4.4 to the standard logistic distribution in the form:

$$GL(x;0,1,1) = \frac{\alpha e^{-x}}{(1+e^{-x})^{\alpha+1}} \quad 4.5$$

The type I GL distribution (equation 4.4) is a very flexible, versatile and widely applicable distribution. It can also exhibit negative and positive skew (Fig 4.4) by adjusting the shape parameter, α. If $\alpha < 1$, the distribution is left-skewed and, if $\alpha > 1$, the distribution is right-skewed. Thus, the type I GL distribution is a good candidate distribution since it has that capacity to model data with a range of different shapes under a single mathematical framework.



(a) (b)

Fig 4.4. The generalised logistic distribution. (a) the probability distribution function (PDF) and (b) cumulative distribution function (CDF) with location ($\theta = 0$), scale ($\sigma = 1$) and varying shape ($\alpha = 2, 1, 0.5$) parameters.

The type I GL distribution has been shown to have nice statistical properties [152–155] corresponding to the Expectation $E(X)$, variance $Var(X)$ and $Skew(X)$ given by

$$E(X) = \theta + \sigma(\psi(\alpha) - \psi(1)) \quad 4.6$$

$$Var(X) = \sigma^2\left(\pi^2/6 + \psi^{'}(\alpha)\right) \quad 4.7$$

$$Skew(X) = \frac{\psi^{''}(\alpha) - \psi^{''}(1)}{(\psi^{'}(\alpha) + \psi^{'}(1))^{\frac{3}{2}}} \quad 4.8$$

respectively where $\psi(.)$ is the $\psi'(.)$ and $\psi''(.)$ are the first three derivatives of the function $log\Gamma(.)$ where Γ denotes the gamma function.

4.3 Materials and Methods

4.3.1 Datasets

The datasets explored in this chapter are similar to Ref. [39] representing the first 446 days of the pandemic. In this chapter, the data has since been updated to include an additional 313 days. Public Health England (PHE) (https://coronavirus.data.gov.uk/) provided a host of datasets relating to COVID-19. It has throughout the pandemic evolved with several updates and provided additional data (Cases, Deaths, Hospital Admissions etc.) In this chapter, English and Welsh data are used on the number of COVID-19 cases but only English deaths for LTLAs. For the death data, English regions were included since Wales have a different methodological approach in collecting death data. In the data set, English deaths that were obtained are people who had a positive test result for COVID-19 and die within 28 days. Alike to the range of datasets available there are also a range of time and spatial scales from both PHE and the UK ONS. PHE and ONS provided data at middle super output area (MSOA) (7,210 regions), but this data are updated weekly compared to PHE LTLA (337 regions) data which was updated daily. ONS also survey the prevalence of COVID-19, however, like their MSOA data, are only updated weekly. After taking all this information into consideration, the PHE daily data by LTLA was selected to define regional boundaries as a trade-off between temporal and spatial coverage as well as allowing for the most up-to-date coverage.

The England and Wales population estimates along with the 19 age categories of age ranging from 0-4 years old to 85+ years old were obtained using the 2011 census. At the time of writing this was the most up-to-date census data. Population and age estimates were obtained from NOMIS (https://www.nomisweb.co.uk) along with regional land area. Shape files were obtained from either 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). LTLAs for COVID-19 cases (in England and Wales), COVID-19 morality (England alone), LTLA population, LTLA area were aligned in a daily time series covering the period from 01/03/2020 to 29/03/2022. All data in this study are publicly available under Crown Copyright.

The COVID-19 data were obtained from PHE. Throughout the pandemic the format and reporting of COVID-19 changed several times. On some days regions with missing data have been reported as 0 and in other periods of time it has been reported as NULL. In each circumstance these have been treated as 0 COVID-19 cases or deaths for those regions. The presence and dealing with 0 values have been a long-lasting issue within scaling [156] due to the log transformation of the data which is necessary in the scaling methods to obtain model parameters. The exclusion of 0 values using these methods means that regions that do not report any COVID-19 cases or deaths on a particular day are excluded in the analysis. The limitations of zero-inflated values have been discussed at length [134,156]. This can be detrimental in PL estimated parameters, thus, similar to chapter 3, Leitão methods are applied to compare exponents and verify whether the scaling is not rejected. This analysis will be a particularly interesting in periods of times where there are lots or regions reporting 0 cases or deaths. In extreme cases of zero-inflated regions in PL models, it has been shown that scaling exponents are misleading. Consequently, regions may be shown as over-performing or under-performing for a population density [134]. At the beginning of the pandemic absences are frequent. PHE also decided the collate some regions together that are considered to have a small population. This included combining City of London (a small 289-hectare region within the greater London metropolitan area with a small resident population [38,39]) with Hackney and the Isles of Scilly with Cornwall. Daily analysis is restricted to the information provided at the time. Specifically, testing

regimes drastically changed throughout the pandemic and death reporting was limited. For example, at the beginning of the pandemic, testing was limited due to little availability. Thus, the mechanisms were not in place to identify as many cases of COVID-19 in comparison to when mass testing was introduced. Death reported by PHE earlier in the pandemic were restricted to those occurring within 28 days of a positive test. The above data processing is inherent in the data set. The R code to obtain the PHE data are available in appendix H3.

4.3.2 Statistical Analysis

The data were analysed using the statistical software R version (3.6.2) [157] with the sf (0.9-1) [158], raster (3.0-12) [107], dplyr (0.8.5) [108], spData (0.3.5) [109], tmap (2.3-2) [110], ggplot2 (3.3.0) [98,159–161], xlsx (0.5.7) [162], gplots (3.0.4) [97] , httr (1.4.2) [163], plyr (1.8.5) [164], png (0.1-7) [114], rgdal (1.5-19) [165], rgeos (0.5-5) [166], lubridate (1.7.9.2) [167], fitdistrplus (1.1-3) [168], fgarch (3042.83.2) [169], glogis (1.0-1) [170], segmented (1.3-1) [87], moments (0.14) [171], nortest (1.0-4) [100], proxy (0.4-24) [172], RColorBrewer (1.1-2) [101], psych (2.0.12) [173], car (3.0-10) and plotrix (3.7-8) [174] packages. The R packages used in this study are also documented in the publication in Ref. [39].

4.4 Results and Discussions

In the following section, the findings from the analysis of COVID-19 case and death metrics using PL models are reported following the methodology described in chapter 2. The daily exponents are presented along with residual variance and skew along with results from age demographics. Separately, residuals are examined to observe daily dispersion using normal and GL residual models. This chapter concludes with an examination of regional persistence exhibited throughout the pandemic.

4.4.1 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 (Fig 4.5(a)) from 1213 ha (Kensington and Chelsea) up to 518,037 ha (Powys) and have populations (Fig 4.5(b)) between 37,340 (Rutland) up to 1,070,912 (Birmingham). Population densities (Fig 4.5(c)) vary from 0.25 people per hectare (p/ha) (Eden) to 138 p/ha (Islington). Although the timeframe of the data has been updated since the publication of this work, the regions used are identical [39].

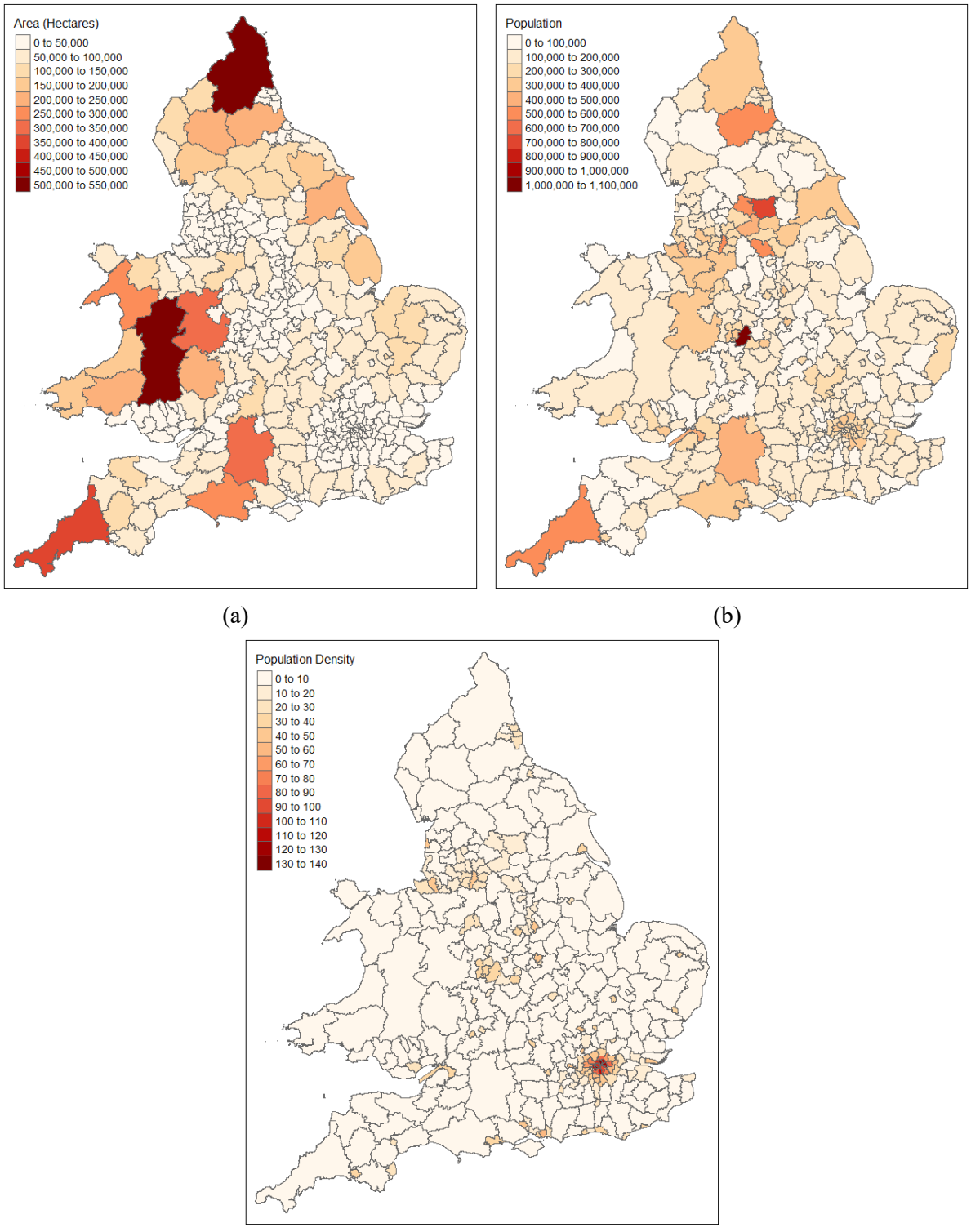
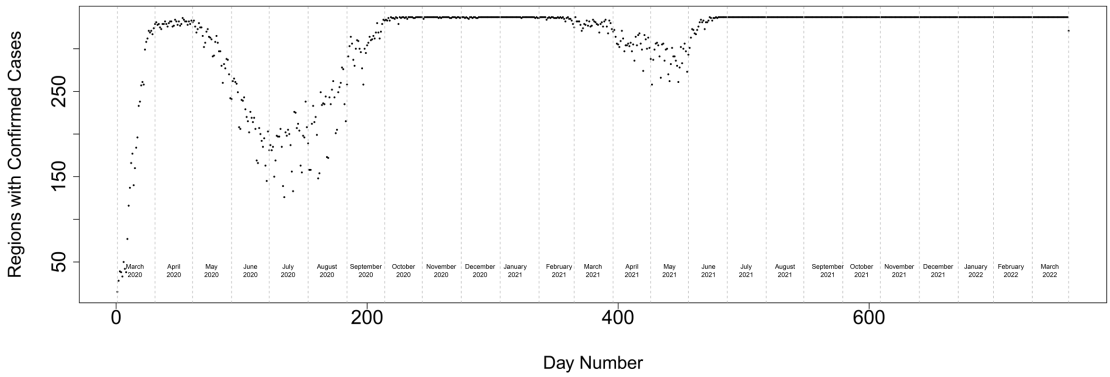


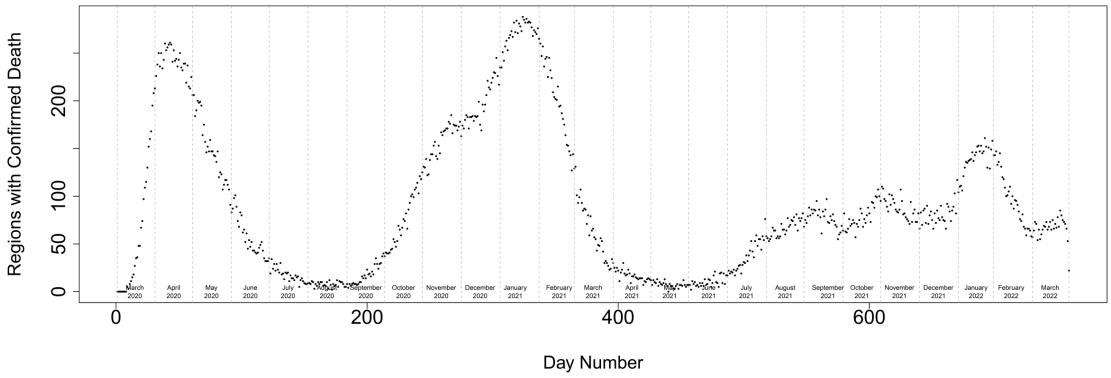
Fig 4.5. Map of England and Wales using LTLAs to define boundaries. The shade of red indicates the strength of the (a) Area (Hectares), (b) Population and (c) Population Density. The darker the red the higher the value.

COVID-19 cases and deaths were recorded daily and not all LTLA regions reported cases or deaths on each day. The variability in the observations tracks the general progress of the pandemic

(Fig 4.6). It can be observed that fewer regions reported cases is predominant at the very beginning of the pandemic, summer 2020 and spring 2021. Regions reporting no deaths are also predominant at the very beginning of the pandemic, summer 2020 and spring 2021; thereafter stays at a much lower level for the remaining time studied. Some of the behaviour is likely due to the variability of testing which was at a minimal capacity at the very beginning of the pandemic and significantly increased over the 25-month period. In recognition of this, daily scaling, variance and skew reflect the processes on a day-to-day basis and were not in correspondence with testing and the number of observations.



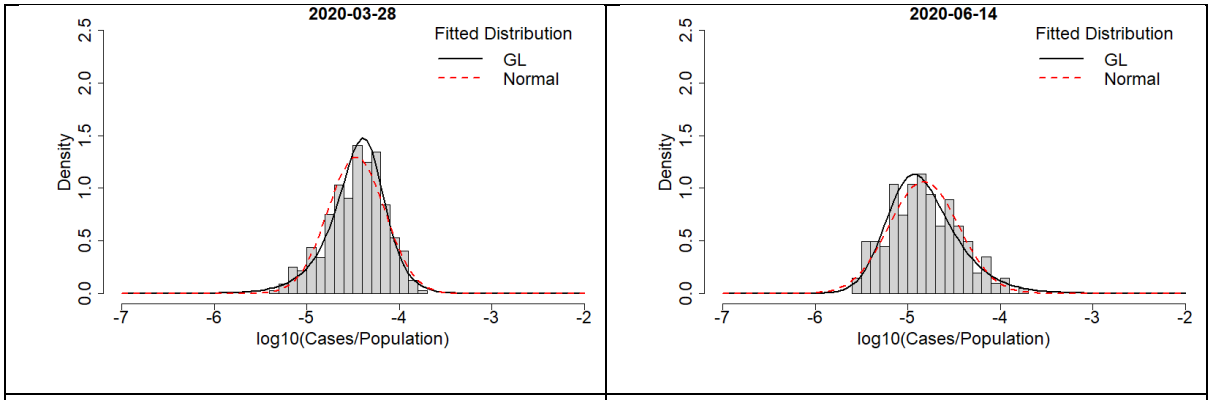
(a)



(b)

Fig 4.6. Time series of the number of LTLA regions reporting (a) cases or (b) deaths. The timeframe is between 01/03/2020 and 29/03/2022. Figure updated from Ref. [39] **with additional data.**

Per-capita cases presented as histograms (Fig 4.7) exhibited variable shapes with some periods of the pandemic showing negative skew (Fig 4.7(a)) whilst other periods included positive skew (Fig 4.7(b)). All daily *per-capita* case histograms can be found in Appendix B1.



(a) (b)

Fig 4.7. Histogram of *per-capita* cases. Red dashed line represents the fitted normal distribution and the solid black line represents the fitted GLD on (a) 28/03/2020 and (b) 14/06/2020. The GLD was able to capture both negative and positive skew. This figure was published in Ref. [39].

4.4.2 Daily Progression of COVID-19

Throughout the pandemic, daily scaling of COVID-19 cases and deaths were well approximated using a single PL model. Plots of log(COVID-19 cases) vs. log(Population Density) (Fig 4.8 a-d) show the distribution of the observed data around the single PL and the variability of residuals obtained using equation 2.5 at points during the pandemic. For example, the variability of residuals in September 2020 (Fig 4.8(b)) is far greater than it was in the December 2020 holidays (Fig 4.8(c)). The low variance periods indicate homogeneity of cases across the regions whilst high variance periods were indicative of more heterogeneity of cases across regions. All daily scaling plots and corresponding geomaps for case residuals (specifically in Appendices B2 and B4 respectively) and death residuals (specifically in Appendices B3 and B5 respectively) can be found in Appendix B.

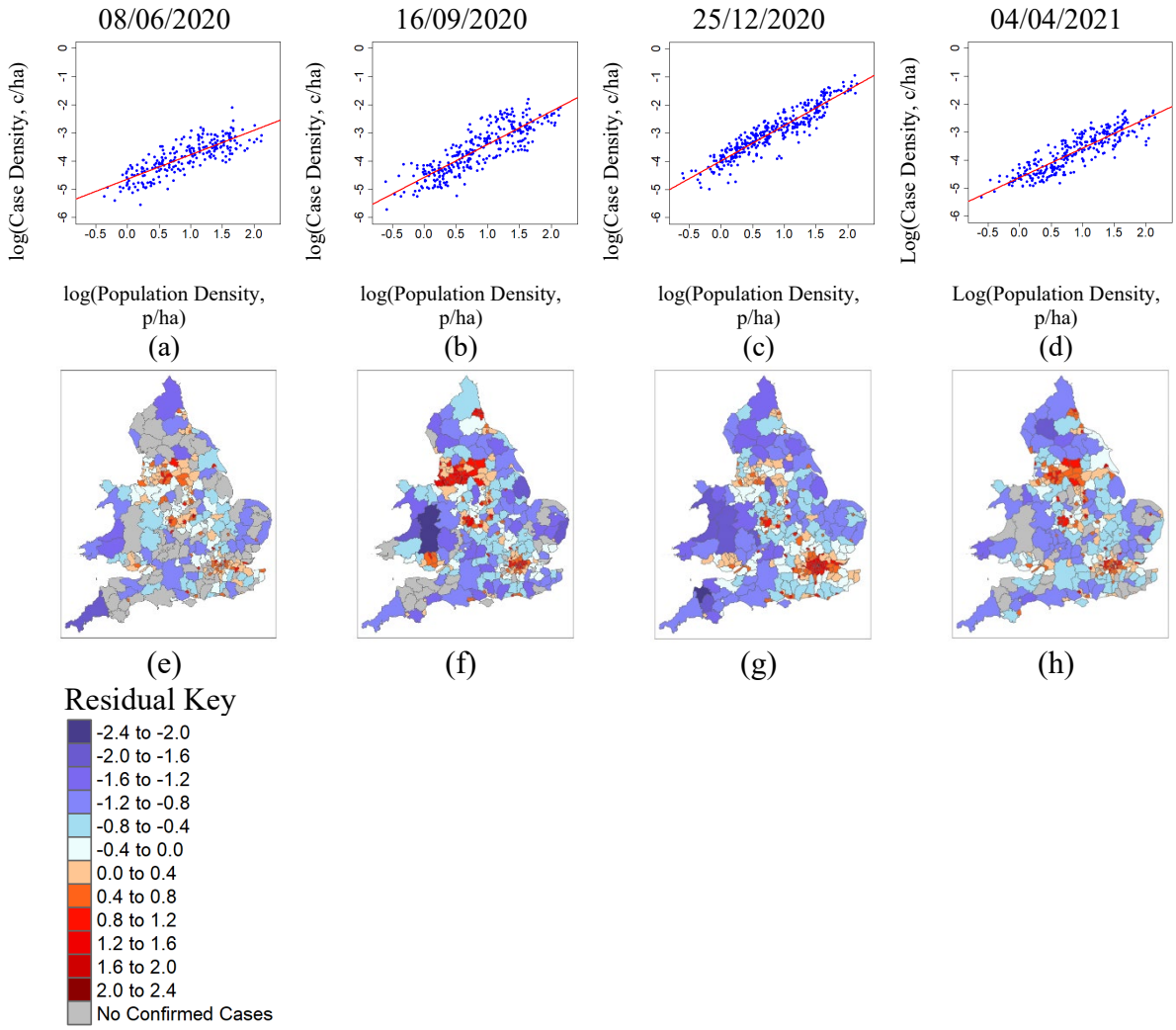
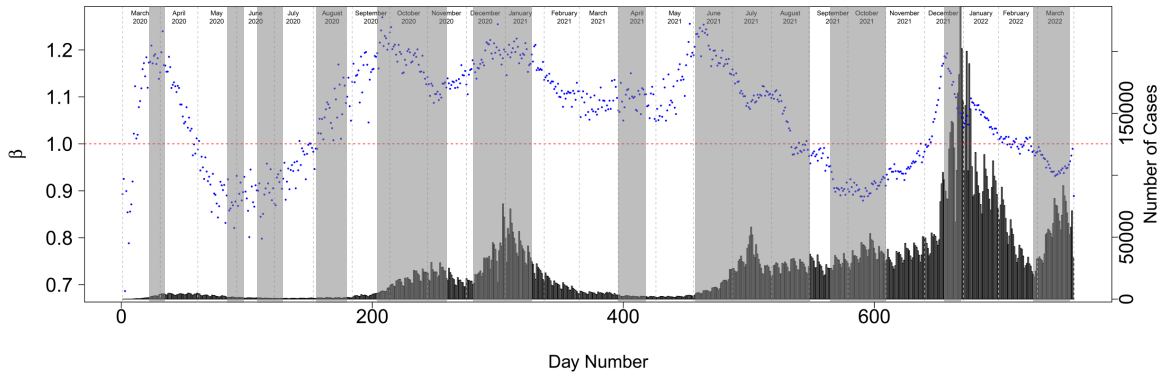


Fig 4.8. Plots of $log(Case\ Density) vs. log(Population\ Density)$ with corresponding geomap of residuals on a selection of days using LTLAs. These are recorded on the (a and e) 08/06/2020. (b and f) 16/09/2020. (c and g) 25/12/2020. (d and h) 04/04/2021. Regions that are red are above the PL and regions that are blue are below the PL. Higher deviations from the PL are indicated by the darker shades. This figure was published in Ref. [39].

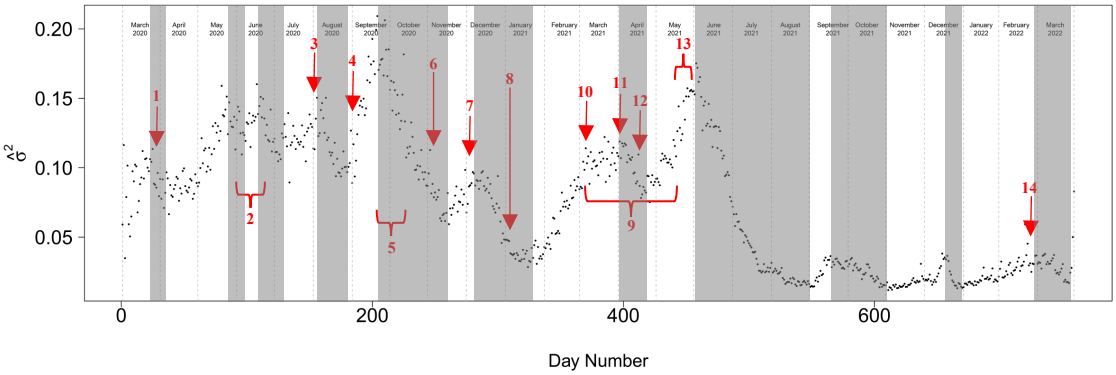
4.4.3 Daily Exponent, Variance and Skewness for Cases

The daily scaling exponent, (Fig 4.9(a)) residual variance (Fig 4.9(b)) and skew (Fig 4.9(c)) metrics were analysed using the LTLA data for cases. The exponent obtained in equation 2.3 via the least squares method using COVID-19 cases and population densities, were examined to explore the scaling behaviour. If $\beta < 1$, scaling behaviour is sub-linear and less populated regions (rural) were preferentially affected by the propagation of COVID-19. If $\beta = 1$, the scaling is linear and all regions despite whether they are rural or urban were proportionately affected. Lastly, if $\beta > 1$ the scaling is super-linear and cases accelerated with population density. Initially, at the very beginning of the pandemic, the scaling exponent reached a high level and peaked near to the start of the first national

lockdown (announced on 23/03/2020) in England and Wales. Thereafter, the scaling exponent gradually declined until restrictions were eased towards the end of May and early June. Throughout the pandemic peaks occurred when $\beta > 1$ indicating super-linear scaling, but this scaling relationship was not universal and preference for cases between rural vs. urban reversed six times. These transitions (Fig 4.9(a)) are seen when the trajectory of the exponent crossed the red horizonal line at $\beta = 1$ in early March 2020, late April 2020, late July 2020, late August 2021, early December 2021 and late February 2022. Overall, during the studied period, the exponent 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.

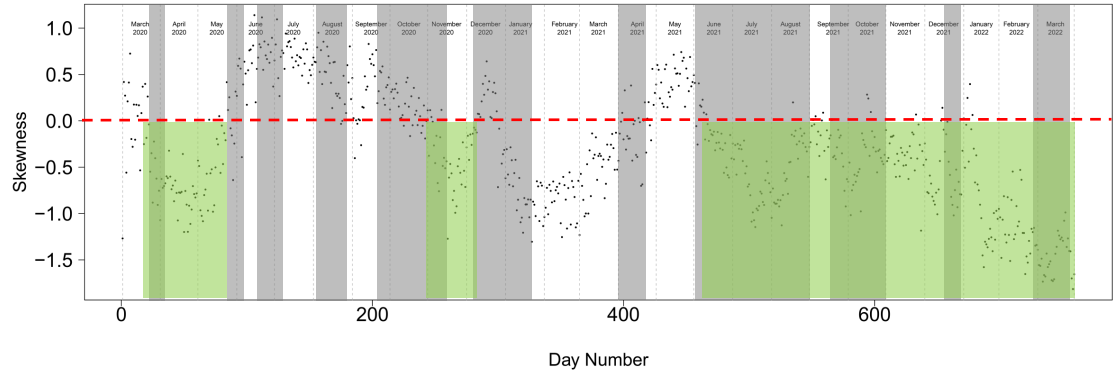


(a)



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.
13) Delta becomes the dominating Variant	14) No longer required to report a positive test nor isolate in any way with COVID-19		

(b)



(c)

Fig 4.9. Daily time series of (a) scaling exponents and (b) residual variance and (c) skewness for cases between 01/03/2020 and 29/03/2022. In (a) blue dots represent daily exponents, the red horizontal dashed line indicates linear scaling and the bar chart is the number of raw daily cases. In (b)

black dots represent daily variance. Red arrows and curly brackets mark key events listed in the lower panel. In (c) black dots represent skewness. The horizontal red dashed line indicates no skew. Above the line indicates positive skew whilst below the line indicates negative skew. The green shade indicates all the periods of negatively skew residuals. The grey shading in all the plots indicates periods of homogenisation. The national restrictions in Wales preceded England beginning on 20/10/2020. Figure updated from Ref. [39] with additional data.

Daily residual variance indicates the variation in the residuals to the scaling PL model (equation 2.3). High residual variance indicates a large amount of variation in all the LTLA regions and conversely low residual variance indicates small variations in all the LTLA regions. Overall, the residual variance throughout the pandemic (Fig 4.9(b)) has a clear structure and changed by over a factor of 10. Some of the structure in the residual variance can be explained by certain regional events.

The variance exhibited at the very beginning of the pandemic remained relatively constant. However, in late April and later stages of the first national lockdown residual variance increased. An increase in residual variance indicates a heterogeneous presentation of regions whilst low residual variance reduced the distance from the PL and represents a homogeneous presentation of regions. For example, periods of time with restrictions (e.g., national lockdown) tended to increase the distance of regions to the PL whilst released restrictions tended to reduce the distance of regions to the PL. More specifically, regional lockdowns that occurred in the summer 2020 (late June) in greater Manchester and Leicester increased the variance and regional heterogeneity. This also occurred in the January 2021 lockdown where variance increased and continues to increase for 2 months peaking near late March. After the relaxation on restrictions residual variance tended to decrease. Although the re-opening of schools is an exception to this general behaviour seen across the pandemic. Instead, this event homogenised and reduced the distance of regions to the PL. The reopening of schools in England a Wales is a little staggered across regions, but in general, schools re-opened in early September. In this period of re-opening, residual variance doubled in a relatively short period of time which saw an increase from approximately 0.1 to 0.2. This rapid increase is due to an intra-regional spreading (within regions) which did not lead to an inter-regional spreading between regions. During

the time of re-opening schools and fewer restrictions an expected ‘surge’ in cases was not seen and residual variance continues to increase starting in late August and throughout September. This increasing trend stopped approximately towards the end of September and began to decrease to a very low level, reaching a minimum at the beginning of November corresponding to second national lockdown (05/11/2020) which as previous seen in other lockdowns increased the variance. Studies [175,176] on the mobility during restricting periods show a significant reduction in movement and regionally heterogeneity. These results are consistent and correspond with the observations seen in this study.

Another key event during the studied period was the returning of student to UK universities in 2020. This produced a significant correspondence to a reduction in variance and regional homogeneity. To put into context there are approximately 164 universities and approximately 2.4 million university students along with large amount of university employees. This event involves a huge migration of people across the UK and students arriving from abroad. It is worth noting, like schools, there is some variation in the university start dates. Following a week of orientation and social activities teaching typically begins between the end of September and beginning of October. Strikingly, this event, caused reduced variance and regional homogeneity, changing propagating dynamics of cases across the England and Wales. There are many other events in the studied timeframe that caused declining variance such as ending of national restrictions imposed on the general public (03/12/2020), businesses reopening in the second national lockdown (12/04/2021), the emergence of the delta-omicron variants (late spring 2021) and the ending of all COVID-19 restrictions (24/02/2022). Noticeably, surges in cases were only seen after the ending of national restrictions and during the period delta-omicron propagation.

Strikingly, residual variance significantly decreases to a very low level towards the end of May 2021 and persists at a low level for the remaining studied period. This significant drop coincides with a change of variant causing most of the propagation in the UK. The delta variant (first identified in May 2021) spread more easily than the earlier alpha variant causing regions to homogenise during

rapid increases in cases [177,178] and during a timeframe where most restrictions were removed (July 2021). This observation is the first major variation of the disease different to the alpha variant that warranted extra precaution from the UK government and caused the dynamics of propagation to change. Furthermore, the Omicron (first identified on the 27 November 2021) is also another concerning variant occurring during the studied period which was also closely monitored by the UK government [179]. Omicron is also a more transmissible version of the disease compared to the earlier variants [180]. The higher transmission of the disease during the delta-omicron phase of the pandemic caused regions to homogenise (Fig 4.10) and regions remained close to the PL for the remaining studied period.

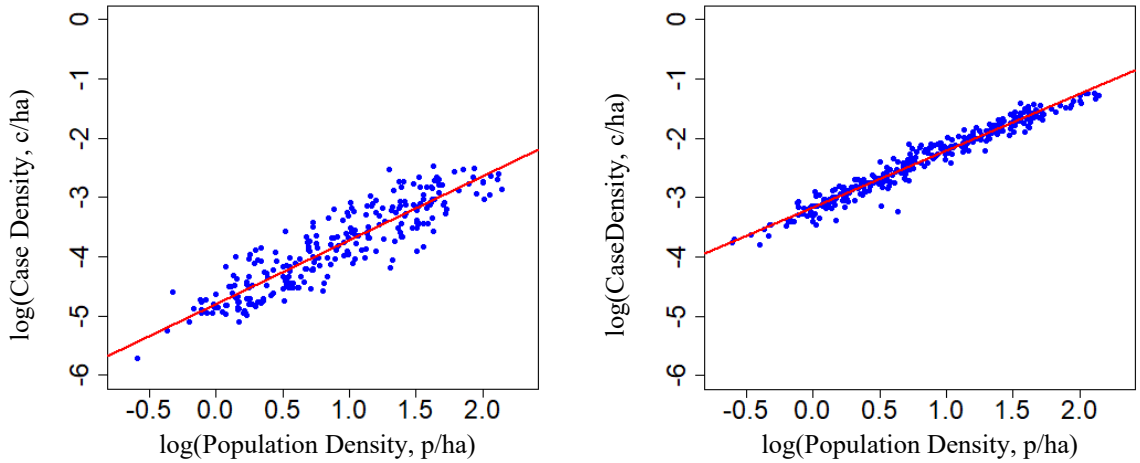


Fig 4.10. Scaling plots showing differences in variation of residuals occurring on (a) 01/05/2021 (an example pre-delta) and (b) 31/08/2021 (during delta propagation) after release of most restrictions.

Skewness provides extra information in addition to the case counts, scaling exponents and variance. Skew metrics were used to create a time series (Fig 4.9(c)) on the scaling law residuals using the GLD third moment (equation 4.8) which is capable of capturing periods of time that exhibit positive and negative skew. Overall, the distribution of residuals and how they are characterised during the pandemic drastically changed like behaviour seen in the *per capita* case distributions (Fig 4.7). In the period between March 2020 and May 2021 the isolation periods correspond to negatively skewed residuals whilst time periods dominated with spreading correspond to positively skewed residuals. This means that the periods of time where residuals follow a distribution with a strong

positive skew, the long positive tail of the skewed distribution is indicative of propagation with ‘hot spots’ and ‘super-spreading’ events. Whilst periods of time where residuals are negatively skewed is indicative of a distribution better characterised by a long tail of ‘cold spots’ and ‘super-isolated regions’. This general behaviour ended during the delta-omicron period (circa. May 2021) when negatively skewed residuals with a homogenised presentation of residuals persisted for the remaining studied period. Some of this presentation is likely due to a different government stance on testing and public health measures which changed several times during the later stages of the pandemic. In the very latter stages of the studied period (February 2022) testing was no longer compulsory and free public testing came to an end entirely causing some regions to deviate below the PL. An example (Fig 4.11) of this occurred on the 18/02/2022 where several regions break away from the PL. The ordered most deviated regions from the PL featuring the long-left tail of the GLD (Fig 4.11(c)) occurred in the regions of Pembrokeshire (0.75 p/h), Powys (0.26 p/h), Gwynedd (0.48 p/h), Ceredigion (0.42 p/h) and Carmarthenshire (0.77 p/h); all of which are Welsh low-density regions.

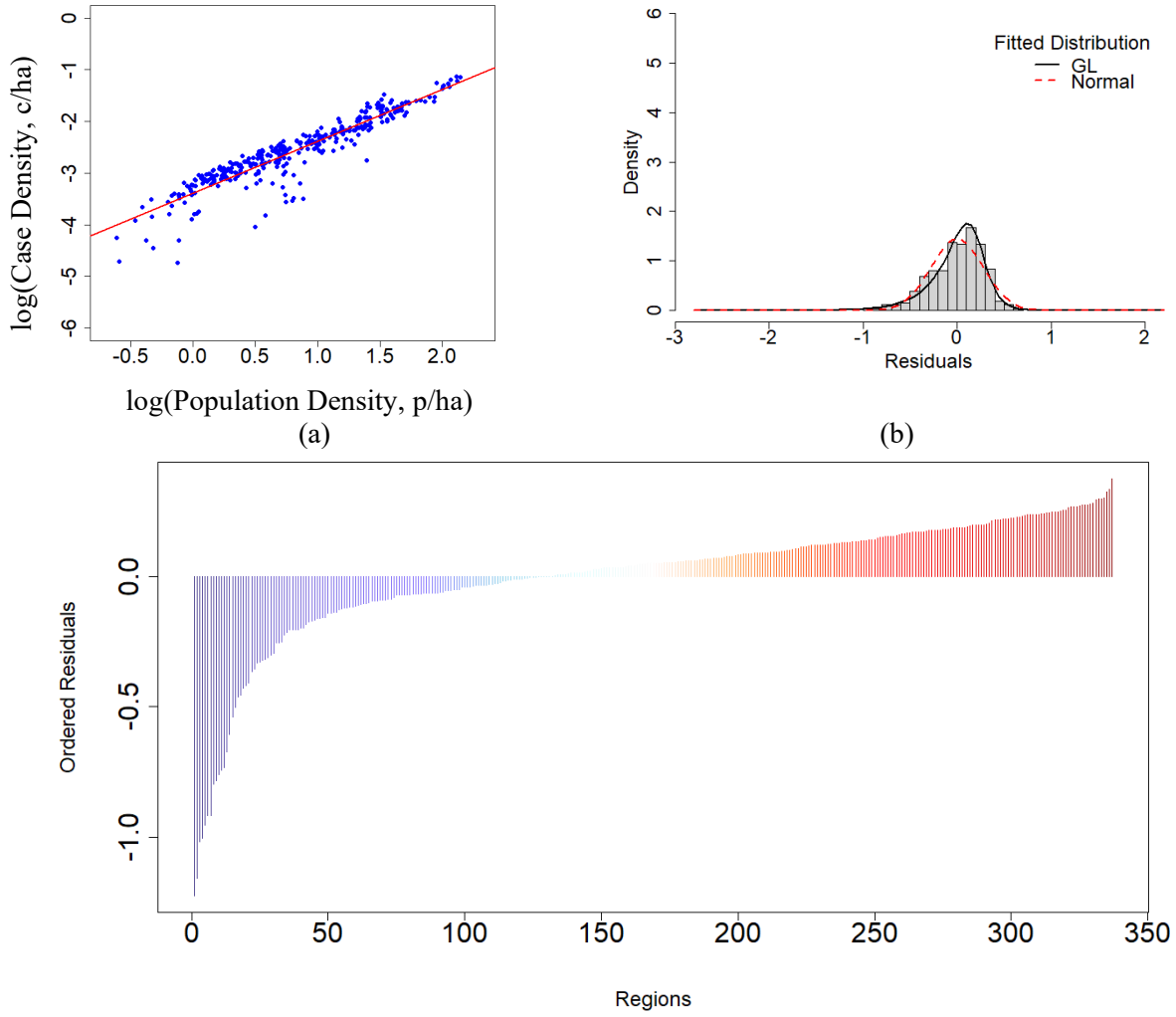
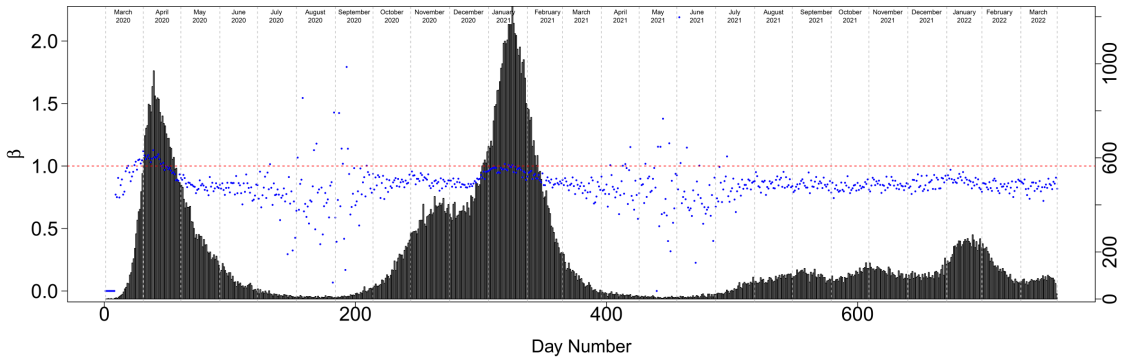


Fig 4.11. An example of residuals on 18/02/2022. (a) scaling plot, (b) histogram of residuals and (c) ordered residuals from lowest to highest value. In (c) blue indicates negative and red indicates positive residuals. The darker the colour the further away from expectation.

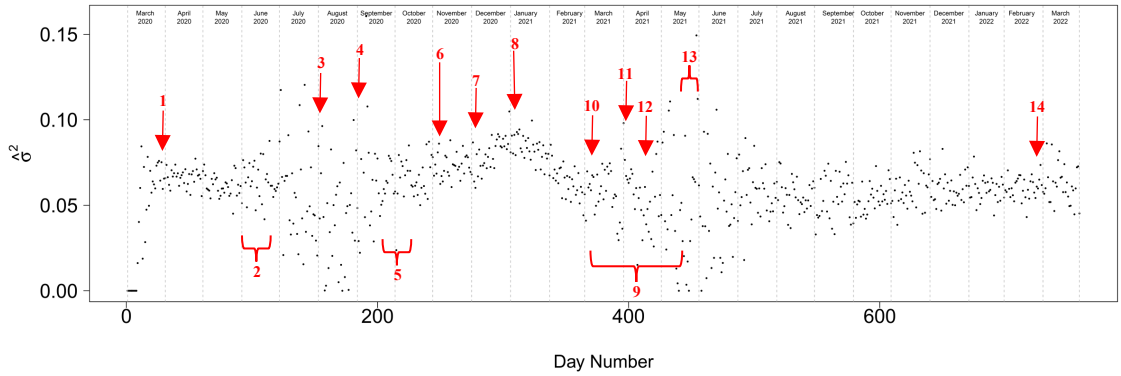
4.4.4 Daily Exponent, Variance and Skewness for Deaths

The daily scaling exponent, variance and skewness previous seen for COVID-19 cases was structured and reactive to certain events. This was the complete opposite when observing this for COVID-19 deaths (Fig 4.12). Remarkably, deaths exhibited almost constant behaviour throughout the pandemic despite government intervention, different testing regimes and a national vaccination programme. It is worth noting that there were days during the pandemic where there are low numbers of regions with either low or zero reported deaths which minimally distorts the overall presentation. Similar to cases, when the scaling exponent is greater than 1 ($\beta > 1$) deaths have a greater impact in urbanised regions, whilst when the scaling exponent is less than 1 ($\beta < 1$) it preferentially affects rural regions.

At the very beginning of the pandemic super-linear scaling was exhibited indicating that deaths preferentially affected urban regions for a short period of time. However, this scaling behaviour soon inverted (circa 10/04/2020) such that regions exhibited constant sub-linear scaling. There are approximately 25 days that are in exception to this general behaviour corresponding to fewer regions reporting cases (Fig 4.6). Thus, for the remaining part of the time series (beyond 10/04/2020) economies of scale was persistent and consistent throughout England. This indicates that the behaviour of deaths corresponds to a consistent regional homogeneity throughout England. These results are in agreement with the age demographics results in chapter 3 [38] where it was found that population dense regions have a younger demographic whilst rural regions had an older demographic and was preferentially affect by general non-transmissible death. To understand this better, age metrics are explored and integrated into this COVID-19 study.

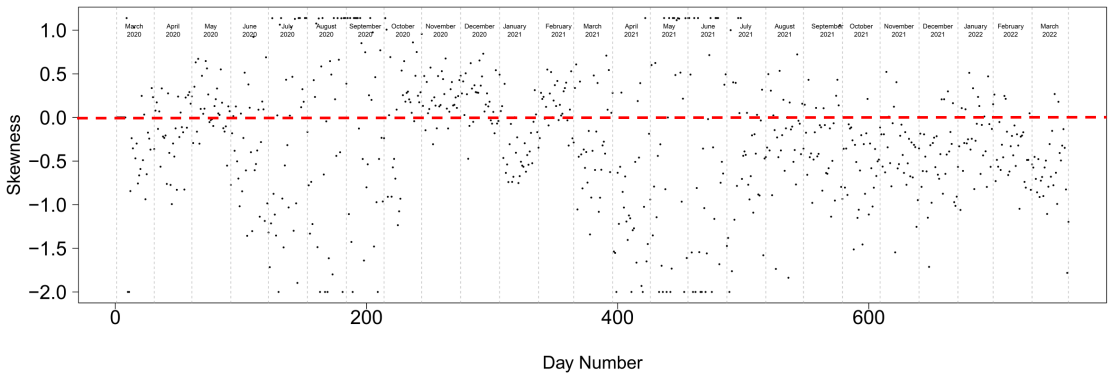


(a)



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.
13) Delta becomes the dominating Variant	14) No longer required to report a positive test nor isolate in any way with COVID-19		

(b)



(c)

Fig 4.12. Daily time series of (a) scaling exponents and (b) residual variance and (c) skewness for deaths between 01/03/2020 and 29/03/2022. In (a) blue dots represent daily exponents, the red horizontal dashed line indicates linear scaling and the bar chart is the number of raw daily deaths. In (b)

black dots represent daily variance. Red arrows and curly brackets mark key events listed in the lower panel. In (c) black dots represent skewness. The horizontal red dashed line indicates no skew. Above the line indicates positive skew whilst below the line indicates negative skew. The national restrictions in Wales preceded England beginning on 20/10/2020. Figure updated from Ref. [39] with additional data.

4.4.5 Age Demographics

Age demographics have been shown to play an important role in the scaling behaviour of general mortality (chapter 3). To understand this better in the context of COVID-19 and to see whether this trend continues, 18 age groups were included in this study ranging from 0-4 years to 85+ years old and aligned these to the English LTLA population, area and death data. For the reasons previously discussed (section 4.3.1) Welsh regions are excluded in this part of the analysis leaving the remaining 315 English LTLA regions. The density scaling models were fitted using PL (Fig 4.13(a) and Appendix B7). The preference of a single PL (equation 2.3) was fitted to the young and middle age groups, whilst preference for a segmenting PL model (equation 2.4) was fitted to 10 remaining age groups. Acceleration in urban region was exhibited in the 24-44 age groups whilst decline in urban regions was exhibited in the people aged 45 and over. The remaining age groups were better fitted using a single PL. These behaviours across all age groups are consistent with previous analysis of age metrics in chapter 3. To explore correlation and clustering between residuals, Spearman's rank correlation coefficient was selected and the hierarchical clustering algorithm. The results (Fig 4.13(b)) showed that the age categories split into two main clusters separating the young groups (0-49 years old) and older groups (aged 50+). The two main clusters exhibited almost universal anti-correlation with values reaching as low as -0.64(Aged 30-34 vs. Aged 75-79 (Fig 4.13(c)) and Aged 30-34 vs. Aged 80-84 (Fig 4.13(d))). In contrast, within the main clusters, residuals are mostly correlated with each other where Spearman's correlation coefficient values reach as high as 0.94 (Aged 70-74 vs. Aged 75-79 (Fig 4.13(e)). In context, this means that a region below the PL (negative residual) in young people have an excess of older people. Similarly, if a region is above the PL (positive residual) in older people there is fewer younger people. This is consistent with results observed in chapter 3. Conversely, relating to the age groups within the same cluster, a region with a high deviation from the

PL results in a high deviation of all the other member within that cluster. All Spearman's rank correlation coefficient values between all age categories are provided in Appendix B8.

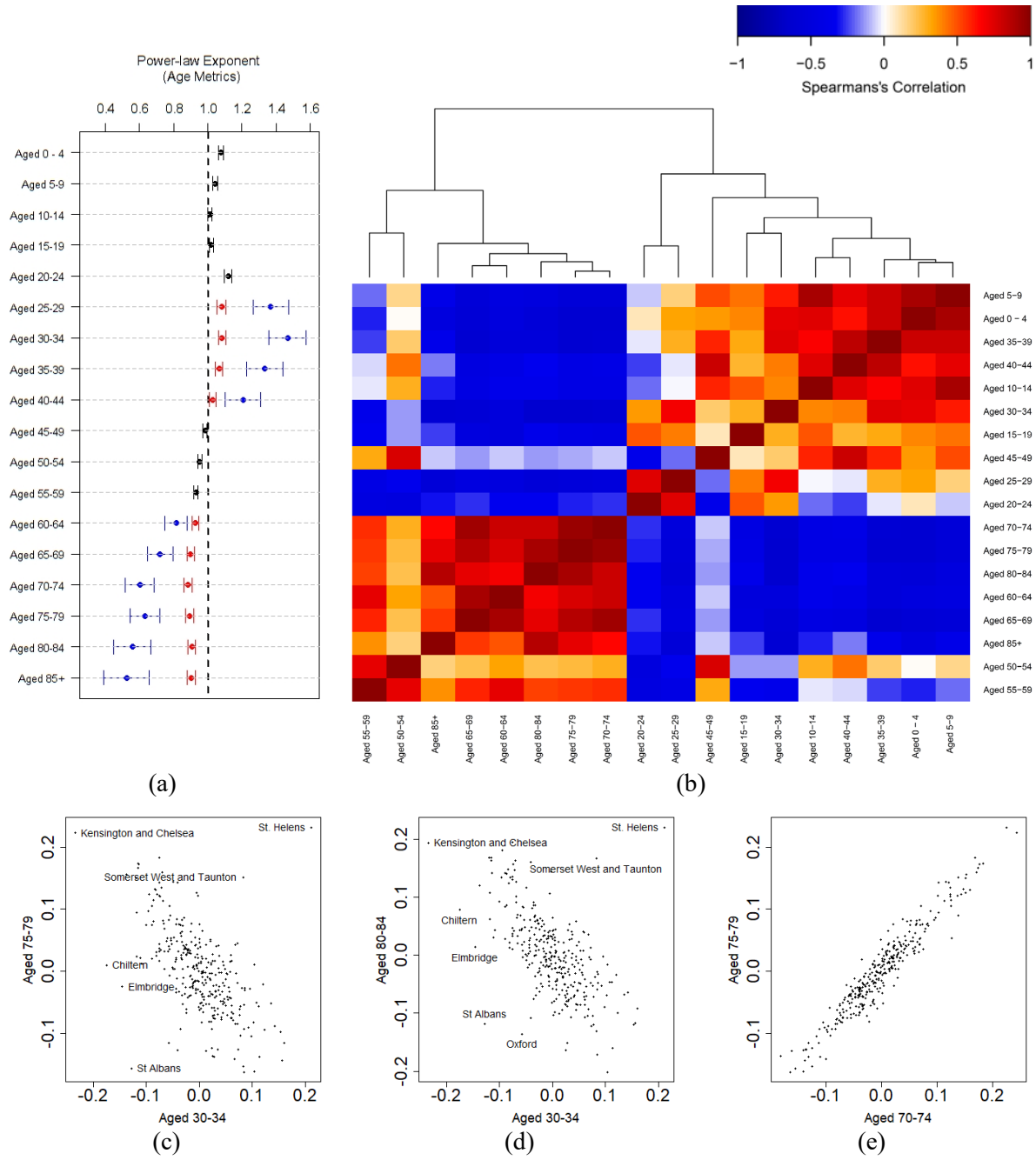


Fig 4.13. PL scaling exponents, correlation heatmap and hierarchical clustering for 18 categories of age. In (a) format and colour coding are the same as figure 3.6. In (b) format and colour coding are the same as figure 3.9 although Spearman's correlation was chosen for this heatmap opposed the Pearson 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. This figure was published in Ref. [39].

Daily scaling of COVID-19 deaths was well approximated using a single PL, however, when considering the cumulative of COVID-19 deaths across the entire studied period a segmented relationship is established indicative that high population density has a 'protective' effect (Fig 4.14(a)). The estimated fitted model parameters using equation 2.4 are: $\beta_L = 1.15, \beta_H = 0.80, d^* = 1.09$. This 'protective' effect is an artefact of age demographics since when restricting total population density to only older age groups the segmented relationship is less prominent and in some circumstances is removed in its entirety. For example, restricting total population to the age group 80-84 the segmented relationship is no longer present and a single PL model is the preferred model (Fig 4.14(b)). The estimated fitted model parameters for this age group using equation 2.3 is $\beta_d = 1.18$. Thus, the apparently 'protective' effect of COVID-19 that exists in urbanised regions disappears and outcome appear to be worse (super-linear) (Fig 4.14).

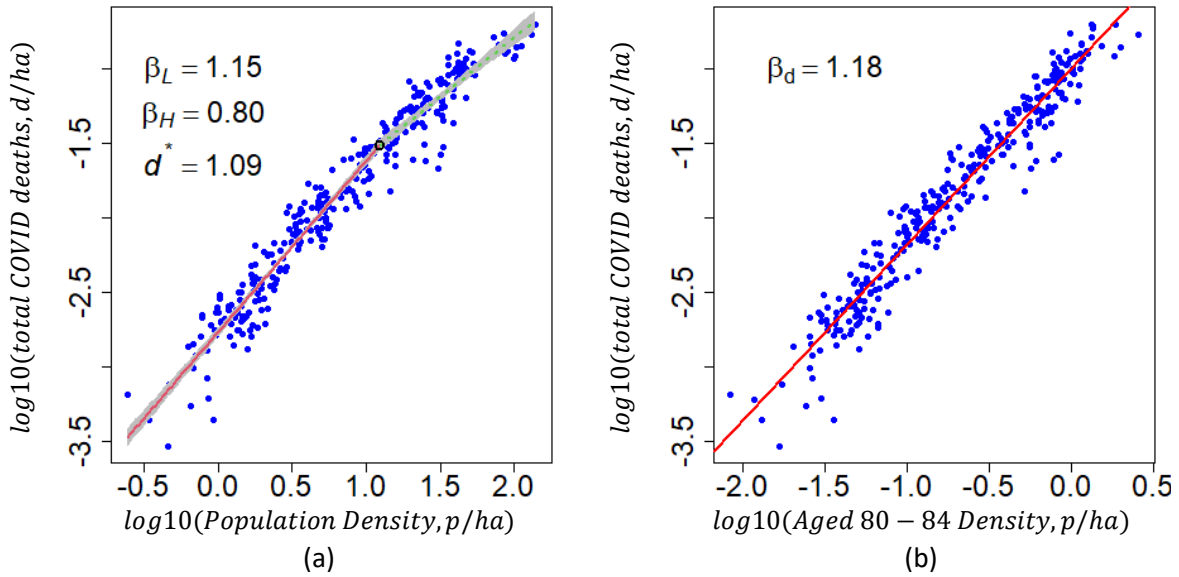


Fig 4.14. Scaling plots for cumulative COVID-19 death density vs. (a) total population density and (b) 80-84 age group. In (a) segmenting PL model. Format and colour coding are the same as figure 3.5. In (b) the segmentation is removed and instead is better fitted with a single PL. This figure was published in Ref. [39].

4.4.6 Dispersion of COVID-19 Case Residuals over Time

To better understand the distribution of residuals, the normal and GLD were explored. This is an important consideration because in a simple linear regression framework it is assumed that residuals are normally distributed, independent, homoscedastic and linear. Section 4.2.2 shows that the normal

distribution is a symmetrical distribution whilst a GLD has three parameters including a shape parameter. The GLD is capable to produce more realistic shapes such as positive and negative skew. AIC and BIC were used to decide whether the normal or GLD was the preferred model for each day during the 25-month studied period. A smaller AIC and BIC score signals a better model. The AIC and BIC differences on each day between the two models are presented in figure 4.15 indicating the model preference for cases (Fig 4.15(a)) and deaths (Fig 4.15(b)). A negative difference corresponds to a GLD (e.g., Fig 4.16(a)) as the preferred model whilst a positive value corresponds to a normal distribution (e.g., Fig 4.16(b)) as the preferred model. All daily histograms of residuals along with a fitted normal and GLD for cases and deaths are available in Appendices B9 and B10 respectively.

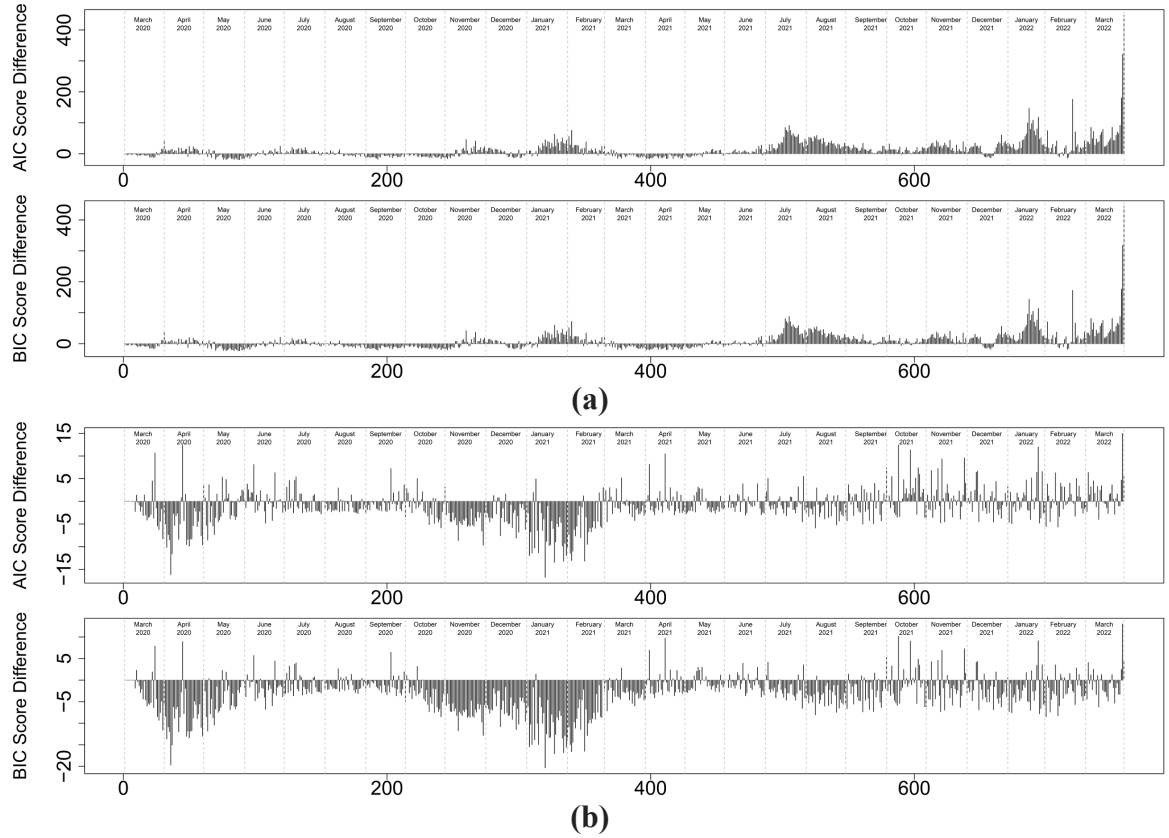
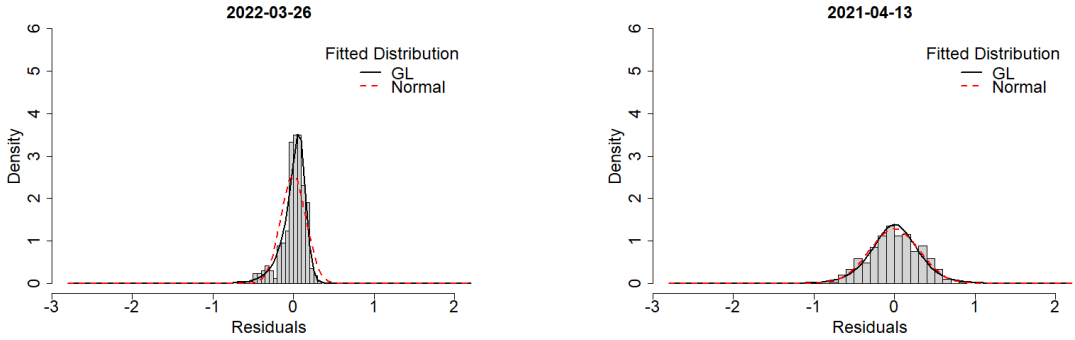


Fig 4.15. Time series of AIC and BIC differences. (a) COVID-19 cases and (b) COVID-19 death. Positive AIC/BIC indicates GL is a better fit and a negative AIC/BIC indicates normal is a better fit. Figure updated from Ref. [39] with additional data.



(a) (b)

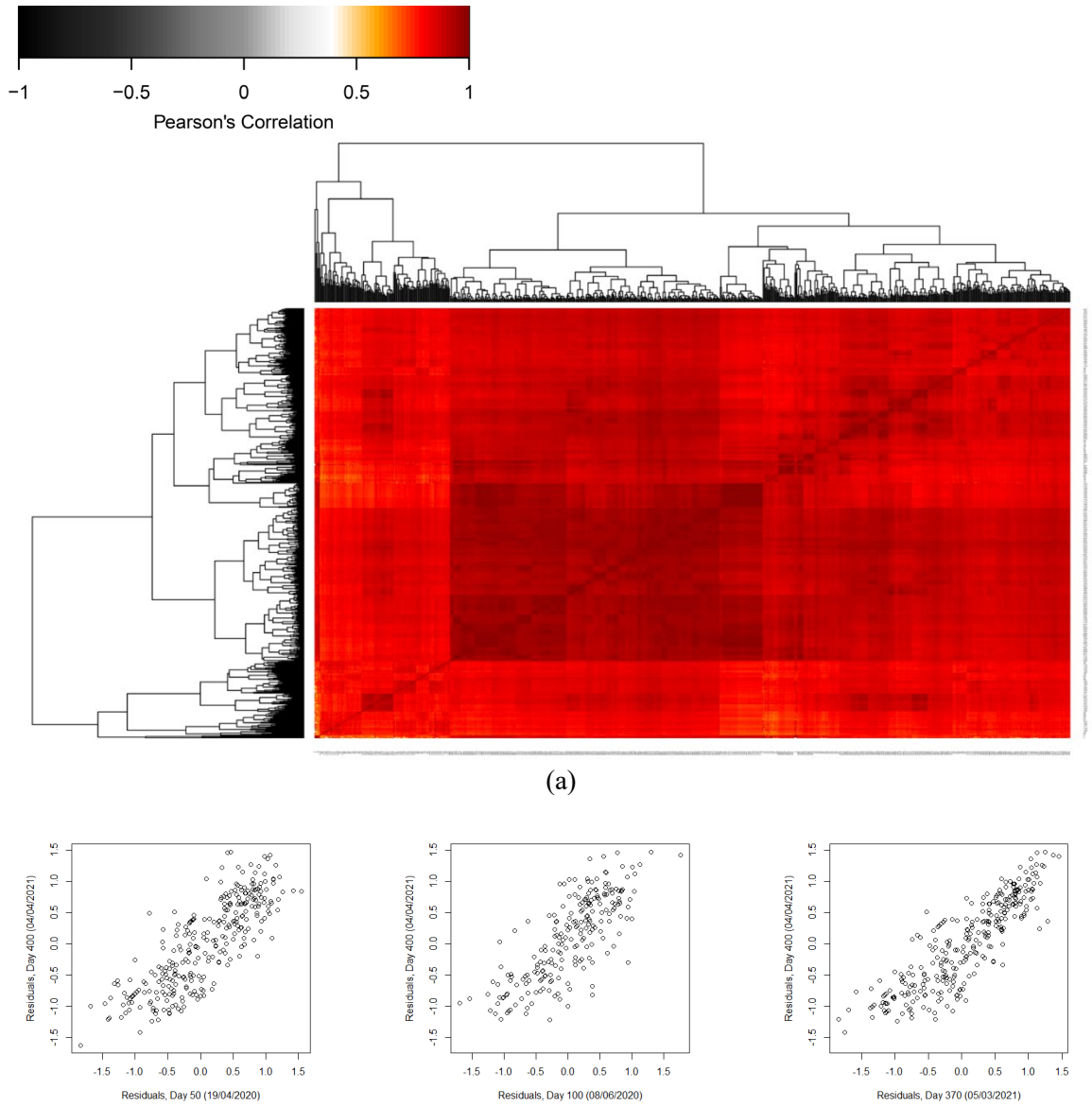
Fig 4.16. Histograms of residuals with example on the (a) 26/03/2022 and (b) 13/04/2021. The red dashed line represents the normal distribution and the solid black line represents the GLD. This figure was published in Ref. [39].

The preference between the normal and GL distributions change for both COVID-19 cases (Fig 4.15(a)) and deaths (Fig 4.15(b)). Apart from some noise, the contrast between cases and deaths is again clear. In the initial periods of lockdowns (March 2020, November 2020 and January 2021) and almost entirely post mid-May 2021 (delta-omicron phase), cases were associated with a GL distribution and negative skew. In contrast, time frames associated with fewer restrictions (August 2020, September 2020, October 2020 and April 2021) with the exception of the delta-omicron phase is associated with a normal distribution. A number of studies have documented a fat tailed and/or positive skewing in superspreading incidents [143,181–183]. In the delta-omicron timeframe, dynamics of spread is different and, although cases surged, this period is better characterised by a negatively skewed GLD featuring a long tail of regions below the PL expectation corresponding to a change in testing regimes such as fewer conditions to get a test later in the pandemic followed by no requirements at all.

4.4.7 Regional Persistence of COVID-19 Case Residuals

To understand regional persistence, the pairwise correlation of residuals for all days was assessed. The resulting heatmap, exhibiting almost universal dark red (Pearson's correlation: Fig 4.17 and Spearman's correlation: Appendix B6) indicated that a regions position and whether it is below or above the PL, remained there for almost the entire pandemic. Although there is some expectation that days relatively close to each other should have some degree of strong correlation this behaviour was

also observed in days that are very far apart. Some examples of strong correlation were observed for day 400 vs. day 50 (Fig 4.17 (b)), day 400 vs. day 100 (Fig 4.17 (c)) and day 400 vs. day 370 (Fig 4.17 (d)). The persisting behaviour remained across all 759 days despite a multitude of government interventions, different testing regimes and a national vaccination programme. In chapter 3 it was found that the inter-relationship between a host of health, well-being and age metric were relating in complex ways. This behaviour observed in chapter 3 and persisting behaviour observed in this study led to the hypothesis that the position of a region relative to the PL is pre-determined and governed by the existing socio-economic characteristics. Further, more direct study, is needed to understand and test this behaviour.



(a) (b) (c) (d)

Fig 4.17. Heatmap of Pearson's correlation coefficient between all days during the pandemic between 01/03/2020 and 29/03/2022. In (a) the orange to dark red indicates moderate to strong positive correlation. White to grey indicates no to moderate correlation and black indicates negative correlation. Example of the correlation from the heatmap are provided between (b) day 400 vs. day 50, (c) day 400 vs. day 100 and (d) day 400 vs. day 370. Figure updated from Ref. [39] with additional data.

4.5 Summary

The regional *per-capita* measures and scaling law residual distributions are generally underappreciated in the scaling literature. They are often thought of as linear, homoscedastic and normally distributed which is often not true and still needs further consideration. For example, in the context of modelling COVID-19 it has been understood that the 'super-spreading' and 'hot spot' are

characterised with positive skew, however this was insufficient to characterise the full timeframe of the COVID-19 pandemic. Cases and to, a much lesser extent deaths, exhibited both positive and negative skew. Timeframes of negative skew coincided with isolation (lockdowns) and regions that account for the long-left tail of the distribution are better regarded as 'super-isolators' and 'cold-spots'. Later stages of the pandemic also exhibit negative skew and it was almost completely universally better characterised by a GLD. This is a feature coinciding with the delta-omicron phase and a change in testing regimes causing several rural regions to break away from the PL.

The COVID-19 pandemic has been a unique opportunity to observe the daily evolution of the scaling exponent which are often thought of as being constant or very slow changing. In the case of COVID-19 cases, the scaling exponent exhibited complex structure and varied over relatively short periods of time. The scaling exponent reached a maximum of 1.27 on 25/12/2020 exhibiting super-linear scaling and reach as low as 0.70 on 03/03/2020 exhibiting sub-linear scaling. The type of scale (sub-linear and super-linear) reversed 6 times where preference of spread between rural and urban regions switched. Super-linear scaling indicates a preference of urban propagation whilst sub-linear indicates a preference of rural propagation. Contrastingly, COVID-19 mortality exhibited near constant economies of scale. In the 25-month studied period the mortality exponent is consistently below 1 and approached linear scaling in two periods occurring in March-April 2020 and in January 2021. These brief moments of linear scaling coincide with peaks in deaths. There is a higher older demographic in low population dense regions and thus, deaths preferentially affected rural environments throughout the pandemic. There are no equivalent reports in any prior scaling literature and no equivalent study of time series. Daily granularity is difficult to find. Thus, this was an opportunity to observe scaling exponents along with variance and skew in daily intervals. It is the first to report a wide variability in the scaling exponent and never been conducted over such as extended timeframe.

Variance for COVID-19 cases relative to the PL produced a considerable amount of structure and complexity throughout the pandemic. Heterogeneity (higher variance) was produced during

lockdowns across regions while reducing the number of cases. Other events during the pandemic contributed to the observed heterogeneity. For example, regional heterogeneity was seen when schools were reopened. This also caused the highest recorded residual variance during local outbreaks of case growth. Whilst homogenisation (low variance) mostly occurred during less restrictive timeframes allowing for county scale mixing. This occurred, for example, in the holiday periods (e.g., Christmas). The delta-omicron phase of the pandemic also caused the homogenisation of regions. It caused residual variance to significantly decrease and remained low during a period of significant case growth. This is an important observation. Not only does the implementation of country scale health policies and events cause changes to residual variance, but this is also affected by the variants propagating at the time. The statistical mechanics characterising COVID-19 mortality in England behaved consistent throughout the pandemic. These results were influenced by the age demographic in rural and urban regions. Rural regions featuring an older demographic are disproportionately affected by COVID-19 deaths whilst urban regions featuring a younger demographic are much less likely to be affected by COVID-19 deaths. From a policy point of view, these observations are important for future intervention and resource allocation. It also provides an underlying understanding of the potential risk of certain events and better allocation of resources to regions needing it most. In this case, it was low population density regions (rural communities). There are no prior timeseries or consideration of residual variance and skew in any prior scaling study. This shows that the statistical assumptions of standard linear regression are not met and demonstrates a need for a more flexible approach.

The dispersion of COVID-19 case and death residuals are complex and change throughout the pandemic. They exhibit positive and negative skew which are features generally underappreciated in the scaling literature. Consequently, in large parts of the studied period residuals are better characterised by a GLD which iS flexible framework and allows for more realistic shapes. This needs further consideration and demonstrates a need for a generalisation of population PL models to accommodate the complexities observed in this study. There is no equivalent statistical analysis of residuals. This study is the first to analyse the residual distributions and show that they are not always

normally distributed. Residuals are better modelled using a more flexible distribution such as the GLD allowing for skew, or otherwise, non-normal continuous data.

In the current literature, the modelling of COVID-19 propagation, have only considered positively skewed distributions. This includes the Weibull [141], Poisson [142] and Gamma [184] distributions. These distributions have been shown to be inadequate to characterise the full sweep of the pandemic due to their inability to exhibit negative skew behaviour. It has been shown that both positive and negative skew are important features of propagation and the current range of distributions in consideration are unable to capture both types of skews whilst completely disregarding the latter entirely. Negative skew was exhibited in approximately 14 months featuring a long-left tail of 'cold spots' and 'super isolating' events. The remaining timeframes are better characterised by positive skew featuring a long right tail of 'hot spots' and 'super spreading' events. This study is the first to acknowledge that both positive and negative skew are important features of propagation. Current distributions are unable to be fully appreciate and allow for negative skewed distributions. It has been shown that the GLD can uniformly capture both positive and negative behaviour: a flexible distribution able to model a range of complex shapes.

The residual behaviour relative to the PL, persisted throughout the pandemic, across all 759 days. In other words, if a residual (region) was above the PL it remained there for the entire pandemic. Similarly, if a residual (region) was below the PL it would remain. This persisting behaviour endured multiple national and local lockdowns, different testing regimes and a national vaccination programme. Such robust regional persistence has never been discovered before.

4.5.1 Limitations

The modelling of dispersion is not considered in chapter 4. It is observed in the timeseries that residual variance exhibited complex behaviour, expanding and contracting throughout the pandemic. To consider these complexities further, it is highly desirable to have a unified model that can

automatically adapt to the underlying dispersion and that can be easily implemented in practice. In chapter 5, a GLD regression model in Bayesian framework is developed, able to adapt in a simple way to different types of dispersions allowing for complex shapes such as positive and negative skew, and heteroscedasticity.

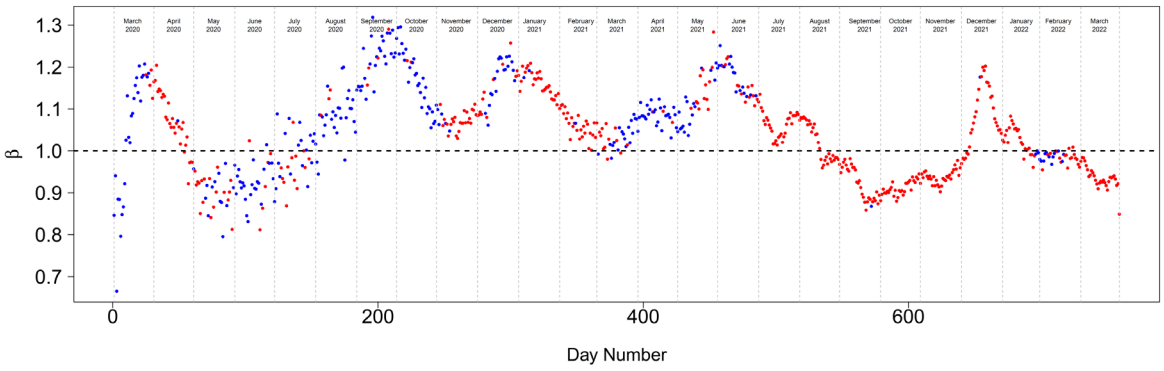
In a standard linear regression framework, it is assumed that residuals are normally distributed, homoscedastic, independent and linear. It has been shown that after fitting the COVID-19 datasets that it does not meet the requirements of standard linear regression. The GLD is a more flexible framework allowing for skewed, otherwise, non-normally distributed residuals. In this chapter, the fitting of the GLD was applied to residuals after using a standard regression model. Residuals have already been affected by model bias using normality methods and following a pre-set of assumptions that have been shown to be violated. To reduce model bias and to obtain better defined residuals, chapter 5 develops an alternative methodology, taking into account residual behaviour within the framework opposed to disregarding them and fitting them to a GLD after obtaining them from an inadequate normal model.

4.5.2 Lognormal Methods for Cases

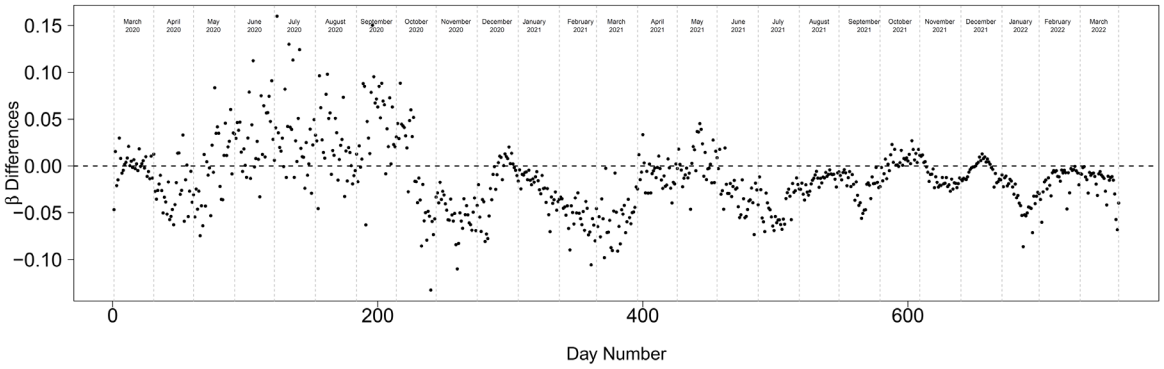
Leitão methods were applied to daily COVID-19 cases to investigate whether PL exponents are different to exponents obtained using least square estimates and to find whether the scaling is statistically significant (Fig 4.18). It was found that exponents documented using Leitão methods (Fig 4.18; panel (a)) correspond to previous reported scaling (Fig 4.9) regarding the relationships identified (i.e., sublinear, linear or super linear scaling) and the daily trajectory. The daily scaling using Leitão methods were also tested for statistical significance. If exponents are indicated in blue, this suggests that reported scaling is statistically significant (i.e., p-value > 0.05). Otherwise, non-statistically significant scaling was indicated in red. It was found that only 288 daily scaling models (out of a possible 759 days) were statistically significant. Differences in exponents were also computed to

explore the variation between methods (Fig 4.18; panel (b)). Specifically, a positive difference (i.e., β difference > 0), indicated that Leitão reported exponents are larger compared to the least square method. While a negative difference (i.e., β difference < 0), indicated that the Leitão reported exponents are smaller. It was found that larger differences between methods occurred in periods with fewer reported cases (e.g., summer 2020).

These results further highlight the need for an alternative methodology that deals with the limitations of least squared methods when assuming a normally distributed and homoscedastic system. For example, residual distributions (Fig 4.16) are found to be highly skewed. Leitão and least squared methods are built around a gaussian distribution which are unable to appreciate any features of skew, causing model bias in estimated PL exponents. It was shown that the GLD distribution can model positive and negative skewed distributions (Fig 4.16). For this reason, a GLD regression model is developed in chapter 5 to address these issues. Furthermore, the developed GLD regression model, like Leitão et al, models variance, but in a more flexible framework to allow features of heteroscedasticity.



(a)



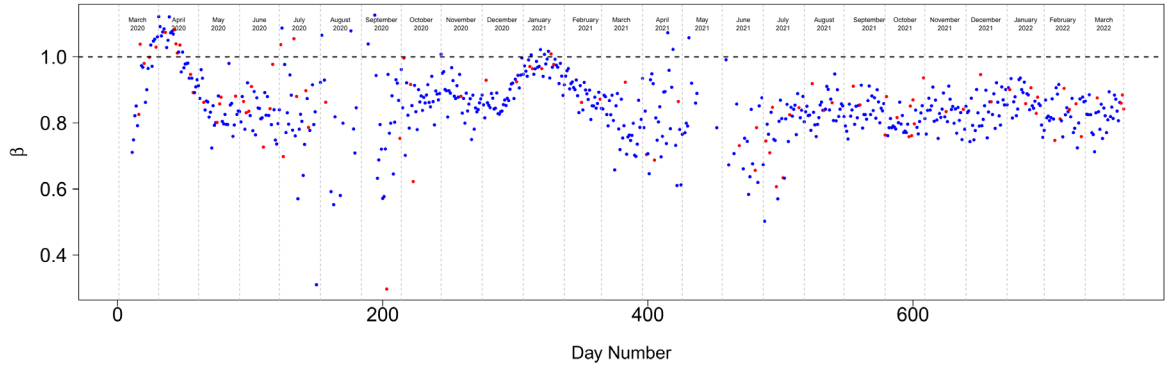
(b)

Fig 4.18. Daily time series of (a) reported exponents using Leitão methods and (b) computed differences between Leitão and least squared methods (i.e., difference = Leitão beta – least squared beta). In (a) blue dots represent statistically significant scaling (p-value > 0.05) while red dots represent non-statistically significant scaling. Also, in (a), the black dashed horizonal line represents linear scaling. In (b), black dots represent the magnitude of the difference between Leitão and least squared methods. Further distances away from 0 indicates a greater amount of variation in reported exponents. Also, in (b), the black dashed horizontal line indicates no differences between methods. Leitão methods used in this analysis are available in ref [37] and corresponding python code is available in ref [135].

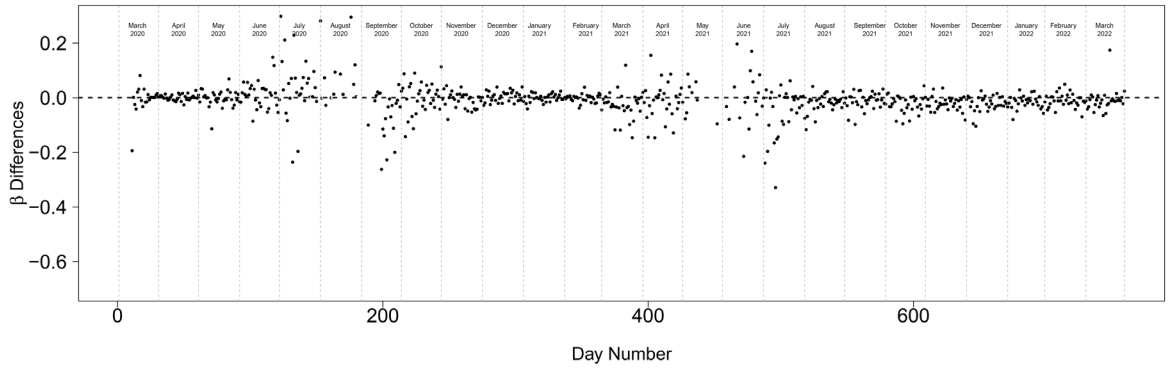
4.5.3 Lognormal Methods for Deaths

Leitão methods were also applied to daily COVID-19 deaths (Fig 4.19). Similar to the presentation in Fig 4.18, blue exponents indicate daily scaling models that are statistically significant (i.e., p-value > 0.05) while red indicate non-statistical significance (Fig 4.19; panel (a)). In contrast to cases, it was found that 590 daily death models (out of a possible 759) using Leitão methods were statistically significant. Differences were also computed to explore the variation in exponents between Leitão and least squared methods (Fig 4.19; panel (b)). In contrast to cases the differences were found to be minimal between methods throughout the pandemic. Although, similar to cases, the largest computed differences correspond to periods with fewer reported deaths. However, it is worth noting that daily

deaths drastically change throughout the pandemic (Fig 4.6; panel (b)) affecting the stability of daily reported exponents. For example, in the summer 2020 and spring 2021, scaling exponent fluctuates from sub linearity to super linearity in a couple of days. These contradictory results indicate that the model is not behaving as expected and this is certainly to the lack of data in these time periods.



(a)



(b)

Fig 4.19. Daily time series of (a) reported exponents using Leitão methods and (b) computed differences between Leitão and least squared methods (i.e., difference = Leitão beta – least squared beta). In (a) blue dots represent statistically significant scaling (p-value > 0.05) while red dots represent non-statistically significant scaling. Also, in (a), the black dashed horizonal line represents linear scaling. In (b), black dots represent the magnitude of the difference between Leitão and least squared methods. Further distances away from 0, indicates a greater amount of variation in reported exponents. Also, in (b), the black dashed horizontal line indicates no differences between methods. Leitão methods used in this analysis are available in ref [37] and corresponding python code is available in ref [135].

Chapter 5: Generalised Logistic Regression Model

5.1 Introduction

The relationship between two variables modelling the response Y in terms of some predictor X can be seen extensively in many different applications in such fields as engineering [185], medicine [38,186], energy [187] and crime [26,27,38]. In a standard linear regression framework [33,188], classical regression assumes residuals are linear, independent, homoscedastic and normally distributed. However, in many cases these assumptions are not valid or appreciated leading to model bias. In chapter 4, these techniques, in a PL setting using data from the COVID-19 pandemic, have been shown to have extensive flaws such as the violation of many of the assumptions of standard linear regression. Residuals are better characterized by a GLD and residual variance exhibited complex behavior, expanding and contracting throughout the pandemic. Fortunately, there are techniques available extending the standard linear regression allowing for a more flexible approach [189] commonly referred to as a generalized linear model (GLM). These techniques usually consider a distribution from the exponential family although other distributions can be considered where context compels another choice. In the most common distributions within the exponential family there are very few that exhibit both positive and negative skew. There are none that exhibit both positive and negative skew mapped to continuous data in $-\infty < x < \infty$. Thus, there is a need to consider other continuous distributions in a GLM-style framework outside the distributions that are mentioned in table 5.1 to model real and both positive and negative skewed data.

Distribution	**pdf $f(x)$**	**Range**
Normal	$\frac{1}{\sigma_{norm}\sqrt{2\pi}} exp\left\{\frac{-(x-\mu)^2}{2\sigma_{norm}^2}\right\}$	$-\infty < x < \infty$
Poisson	$\frac{\mu^x \exp(-\mu)}{x!}$	$x = 0,1,2,\ldots$
Binomial	$\binom{n}{x}\binom{\mu}{n}^x \left(1-\frac{\mu}{n}\right)^{n-x}$	$x = 0,1,\ldots,n$
Gamma	$\frac{1}{\Gamma(v)}\binom{v}{\mu}^v x^{v-1}\exp\left(-\frac{vx}{\mu}\right)$	$x > 0$
Inverse Gaussian	$\sqrt{\frac{\gamma}{2\pi x^3}} exp\left\{\frac{\gamma(x-\mu)^2}{2\mu^2 x}\right\}$	$x > 0$
Exponential	$\lambda\exp(-\lambda x)$	$x \in [0,\infty)$
Pareto	$\frac{\alpha x_m^\alpha}{x^{\alpha+1}}$	$x \in [x_m,\infty)$
Weibull	$\begin{cases}\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}\exp\left\{-\left(\frac{x}{\lambda}\right)^k\right\} & x \geq 0 \\ 0 & x < 0\end{cases}$	$x \in [0,+\infty)$
Beta	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$ $where\ B(\alpha,\beta)$ $= \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ $and\ \Gamma\ is\ the\ gamma\ function$	$x \in [0,1]\ or\ x \in (0,1]$

Table 5.1. A selection of common distributions that belong to the family of exponential distributions. These are available in Refs. [190–192]**. Distribution definitions and notation used in this table are common in the statistics literature and do not correspond to the list of Latin and Greek symbols at the beginning of this thesis.**

In the current literature there are a few proposed regression models, other than the normal approach. The closest to the proposed regression model in section 5.2.3 is a regression model where the response Y assumes a form of the GLD, different to the form considered in this chapter. Specifically, the distribution under consideration is an exponential-logistic {generalized Weibull} (E-L {GW}) distribution with a specific purpose of modelling life-time data using a survival function [193]. Although the distribution is mapped to $-\infty < x < \infty$, the application in this study only

considers positive valued data. This is particularly useful with application to lifetime data, but there is still a need for an approach that allows for a wider application.

There are many variations of the GLD [194–196]. It is an extensively studied distribution in various disciplines that can be found in (Johnson *et al.* [153]) and the references therein. The form considered in this chapter is the type I GLD [152–155,197]. It consists of three parameters including a location, scale and shape allowing it to be an extremely versatile and flexible distribution. The main advantage of the type I GLD is that it can accommodate a wider range of shapes including positive and negative skew, otherwise, non-normal continuous data. It is a mathematically a simple distribution; in that it is a closed form distribution under a single mathematical expression. The GLD also has been shown to have nice statistical properties using method of moments (MOM) corresponding to the expectation, variance and skew [154] and parameters have been shown to be well approximated using maximum likelihood estimates (MLE) [152,197].

In this chapter a Bayesian duel GLM has been developed where the response, Y, assumes a type I GLD is proposed. In the Bayesian GLD regression model, it will allow for non-linear and non-normal data modelling both the expectation and dispersion through a range of linear regression functions, including but not limited to, null (Appendix C1), simple (Appendix C2), polynomial (Appendix C3) and multiple linear regression structures (Appendix C4)). The resulting joint posterior distribution includes a complicated normalizing constant and was not proven to be analytically tractable thus, MCMC methods are applied to fit each of the Bayesian estimate model parameters. The intractable form, in a regression setting has never been seen the statistics literature, thus, the Bayesian framework developed in this chapter will propose suitable methods to deal with such forms. The additional dispersion parameters are indicative of heteroscedasticity along with the GLD model parameters indicative of positive and negative skew. To monitor convergence, Gelman-Rubin convergence diagnostics were applied along with deviance information criterion (DIC) scores to see how this model compares with other regression models using normality methods. Overall, this developed Bayesian GLD regression model is a flexible framework capable of modelling skewed and

heteroscedastic data sets. A normality approach to skewed and heteroscedastic data produces fundamental and systematic model bias relative to the improved methodology proposed in this chapter.

In the final section of this chapter, suggestions and concluding remarks are made about how the Bayesian GLD regression model can be adapted to make further improvements and outline the limitations of the proposed model which offers an opportunity for future continual development. In general, the dual GLD expectation and dispersion regression model is a less restrictive framework, accessible to a wide range of systems and accounts for a wide range of different shapes and heteroscedasticity.

5.1.1 Introducing the Generalized Logistic Distribution

Several forms of the logistic distribution have been proposed in the statistics literature [153,197–202]. To recall, the type I logistic distribution has a probability distribution function (pdf) and cumulative distribution function (cdf) given by:

$$f(x;\alpha) = \frac{\alpha e^{-x}}{(1+e^{-x})^{\alpha+1}} \tag{5.1}$$

and

$$F(x;\alpha) = \frac{1}{(1+e^{-x})^{\alpha}}, \tag{5.2}$$

respectively where $\alpha > 0$ and $-\infty < x < +\infty$. The distribution is negatively skewed for $0 < \alpha < 1$ and positively skewed for $\alpha > 1$. In the cases where $\alpha = 1$ this reduces the type I logistic distribution to the standard logistic distribution [153].

An extension of the type I logistic distribution has been proposed [152,153,197,199], considering a three-parameter form, by introducing location θ and scale σ parameters. This is

commonly known as the type I generalised logistic distribution (GLD). The probability density function (pdf) and the cumulative density function (cdf) of this form is as follows:

$$f(x;\theta,\sigma,\alpha) = \frac{\alpha}{\sigma}\frac{e^{\frac{-(x-\theta)}{\sigma}}}{\left\{1+e^{\frac{-(x-\theta)}{\sigma}}\right\}^{\alpha+1}} \tag{5.3}$$

and

$$F(x;\theta,\sigma,\alpha) = \frac{1}{(1+e^{-(x-\theta)/\sigma})^{\alpha}} \tag{5.4}$$

where θ, σ and α are the location, scale and shape parameters respectively. $\theta \in \mathbb{R}$, $\alpha > 0$, $\sigma > 0$ and $-\infty < x < +\infty$. If $\alpha < 1$ the distribution is left-skewed and if $\alpha > 1$ the distribution is right-skewed. If $\theta = 0$, $\sigma = 1$ and $\alpha = 1$, the equation in 5.3 reduces to the standard logistic distribution in 5.1 [153]. It is possible to derive the parameters, θ , σ and α using the MOMs which have been shown to be asymptotically unbiased and consistent [197] by equating the first k sample moments to their theoretical moments and solving these k simultaneous equations for the parameter estimates. Let $x_1, \dots, x_n$ be a random sample from $F(\,\cdot\,;\theta,\sigma,\alpha)$ and let $\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$. The $R'th$ moment for the GLD is given by:

$$\mu'_R = E[X^R] = \int_{-\infty}^{+\infty} x^R GLD(x;\theta,\sigma,\alpha)dx \tag{5.5}$$

where $GLD(x;\theta,\sigma,\alpha)$ is the type I GLD. Thus, the first three moments of the GLD correspond to the expectation, variance and skew are well defined as:

$$E(\bar{x}) = E(X) = \theta + \sigma(\psi(\alpha) - \psi(1)) \tag{5.6}$$

$$E\left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}\right) = Var(X) = \sigma^2(\pi^2/6 + \psi'(\alpha)) \tag{5.7}$$

Also, the sample skewness coefficient is given by:

$$b_1 = \frac{(n-1)^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left\{(n-1)^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right\}^{3/2}}$$

and satisfies:

$$b_1 = Skew(X) = \frac{\psi''(\alpha) - \psi''(1)}{(\psi'(\alpha) + \psi'(1))^{\frac{3}{2}}} \quad 5.8$$

where the functions $\psi(\cdot)$, $\psi'(\cdot)$ and $\psi''(\cdot)$ are the digamma function and its first and second derivatives, respectively.

5.1.1.1 GLD Maximum Likelihood Estimation

To obtain the maximum likelihood estimation (MLE) for the GLD, it begins by defining the likelihood function which is given by:

$$\begin{aligned} L(\theta, \sigma, \alpha | x_1, \dots, x_n) &= GLD(x_1, \dots, x_n; \theta, \sigma, \alpha) \\ &= GLD(x_1; \theta, \sigma, \alpha) \times GLD(x_2; \theta, \sigma, \alpha) \times GLD(x_n; \theta, \sigma, \alpha) \\ &= \prod_{i=1}^{n} GLD(x_i; \theta, \sigma, \alpha) \end{aligned}$$

where $i = 1, 2, \dots, n$ are the number of independent observations, θ is the location, σ is the scale and α is the shape parameter. Thus, by taking a logarithm of $GLD(x_1, \dots, x_n; \theta, \sigma, \alpha)$ and using the GLD equation given in 5.3 the GLD log likelihood is as follows:

$$\log GLD(x_1, \dots, x_n;\ \theta, \sigma, \alpha) = log\left(\frac{\alpha}{\sigma}\right) - \sum_{i=1}^{n}\frac{x_i - \theta}{\sigma} - (\alpha + 1)\sum_{i=1}^{n} log\left(1 + e^{\frac{-(x_i - \theta)}{\sigma}}\right) \quad 5.9$$

The MLE exist if $\theta \in \mathbb{R},\ \hat{\sigma} > 0$ and $\hat{\alpha} > 0$ which simultaneously maximize the log likelihood function.

To obtain parameter estimates, let the log likelihood estimate of (θ, σ, α) be denoted by $(\hat{\theta}, \hat{\sigma}, \hat{\alpha})$ and find $(\hat{\theta}, \hat{\sigma}, \hat{\alpha})$ that maximizes L. For example, to estimate the shape parameter $\hat{\alpha}$, differentiate (w.r.t $\hat{\alpha}$), and set this to equal zero such that:

$$\frac{\partial \log GLD(x;\ \theta, \sigma, \alpha))}{\partial \hat{\alpha}} = \frac{n}{\hat{\alpha}} - \sum_{i=1}^{n} \log\left(1 + e^{\frac{-(x_i - \hat{\theta})}{\hat{\sigma}}}\right) = 0$$

and then finally solving for $\hat{\alpha}$:

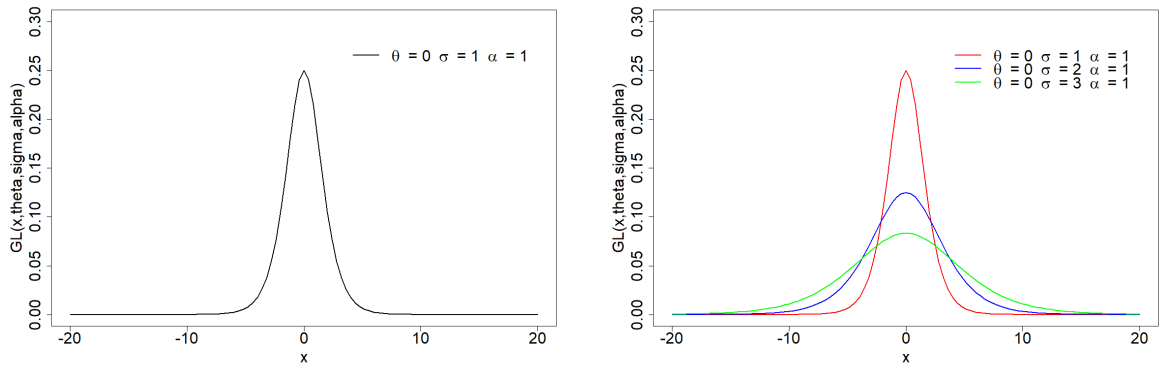
$$\hat{\alpha} = \frac{n}{\sum_{i=1}^{n} \log\left(1 + e^{\frac{-(x_i - \hat{\theta})}{\hat{\sigma}}}\right)} \qquad 5.10$$

from which it can be seen that $\hat{\alpha} \to \infty$ if $\theta \to -\infty$ and that $\hat{\alpha} \to 0$ if $\hat{\sigma} \to 0$ [152].

5.1.1.2 GLD Graphical Presentation

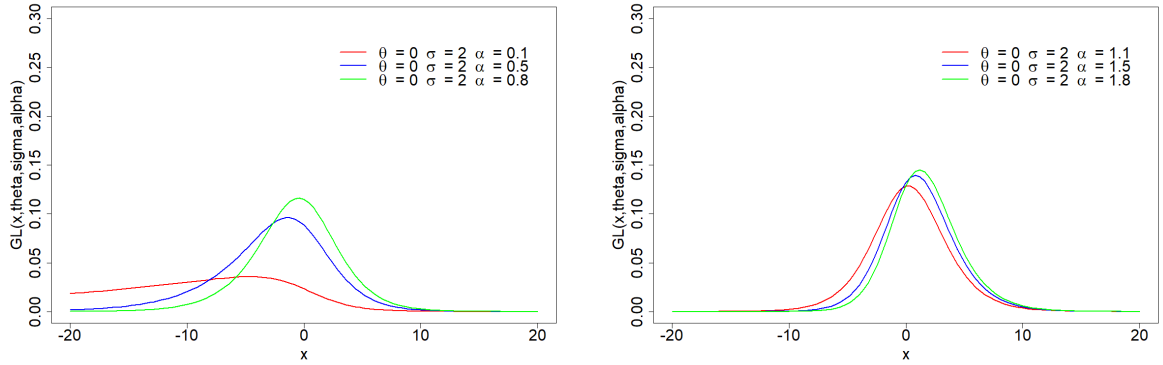
The behaviour of a random variable X following the GLD with different values of θ, σ and α is illustrated in this section. The three-parameter type I GLD can be reduced to the symmetrical standard logistic distribution with the following values $\theta = 0$ and $\sigma = \alpha = 1$ (Fig 5.1(a)). The effect of σ with fixed $\theta = 0$ and $\alpha = 1$ still retains the symmetrical shape of the logistic distribution but increasing the σ widens the span over X (Fig 5.1(b)). The value of α determines the skewness of the distribution (Fig 5.2). When $\alpha < 1$ the distribution is negatively skewed (Fig 5.2(a)) whilst the distribution is positively skewed (Fig 5.2(b)) when $\alpha > 1$. It can also be seen that when $\alpha < 1$ and increasing α values meeting this condition that the GLD tends to have a heavier and longer left tail (Figs 5.2(a) and 5.3(a)). Conversely, it can also be seen that increasing the value of α when $\alpha > 1$ that the right tail of

the GLD becomes heavier. The combination of changing the GLD θ, σ and α parameters (Fig 5.3) demonstrate the flexibility of the GLD able to characterise a range of different complex shapes.



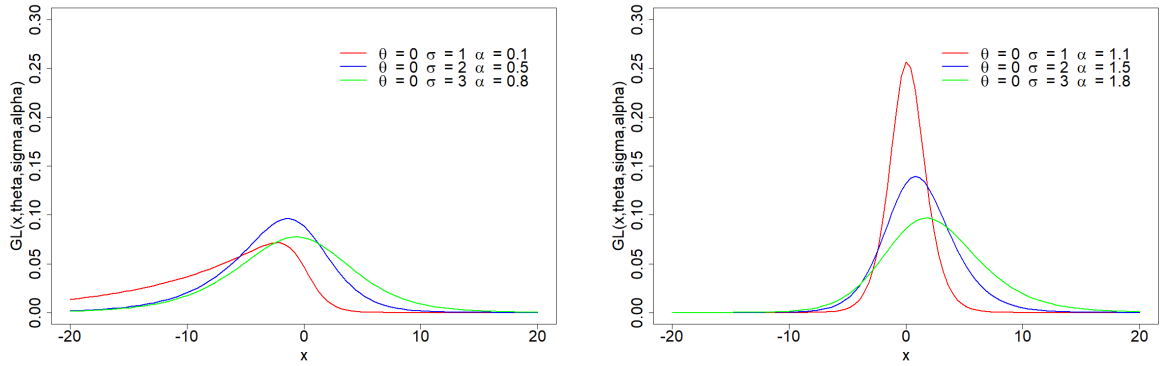
(a) (b)

Fig 5.1. The standard logistic distribution (a) with the following parameterisation: $\theta = 0$, $\sigma = \alpha = 1$ and the generalised logistic distribution (b) with different values of σ with fixed location and shape parameters such that $\theta = 0$ and $\alpha = 1$.

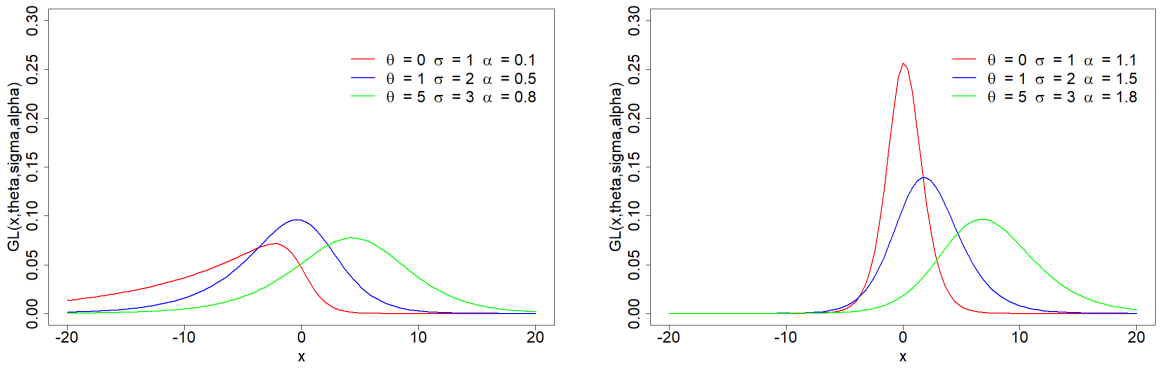


(a) (b)

Fig 5.2. The GLD for different values of α with $\theta = 0$ and $\sigma = 2$.



(a) (b)



(c) (d)

Fig 5.3. The GLD for different values of θ, σ and α. In panels (a)-(b) the GLD location parameter is fixed equal to 0. In panels (c)-(d) the GLD location parameter varies.

In the study of COVID-19 cases and deaths (chapter 4), preference in the extended daily timeseries between the GLD and the normal distributions to model residuals switch several times. Thus, to understand whether the GLD was able to collapse to the normal distribution and to explore the behaviour under different conditions various simulations were performed. It is important that the proposed modelling framework captures universally both symmetrical and non-symmetrical data effectively. In this study 4 simulations were performed. In each, 10,000 randomly normally distributed data were obtained each with a mean of 0 and different standard deviation ($\sigma = 1, 3, 5, 7$). The curves of the normal and GLD were fitted to each of the simulated datasets to observe the variations between the two fitted distributions. The resulting histograms and curves show that the fitted GLD curve compared to the normal curve has slightly heavier tails for simulated data with a low standard deviation (Fig 5.4(a)). Whereas the tails of the GLD and normal distributions become more similar in the case where the standard deviation is higher (Fig 5.4(b-d)). In addition, in each of the simulations (Fig 5.4 (a-d)) the peak of the distribution is slightly elevated in the GLD. This is prominent in the case with low standard deviation which is likely caused by the slightly heavier tails. Furthermore, as expected considering that the data are normally distributed, the GLD shape parameter in each of the simulations is consistent and almost equal to 1 indicating no skew. The GLD location parameter is also consistent and close to the mean of the normally distributed data ($\mu = 0$) whereas the scale parameter increases with increasing standard deviation. Overall, the GLD can model the

'bell shape' curve of the normal distribution with slight variations in the peak and tails depending on conditions of the simulated data.

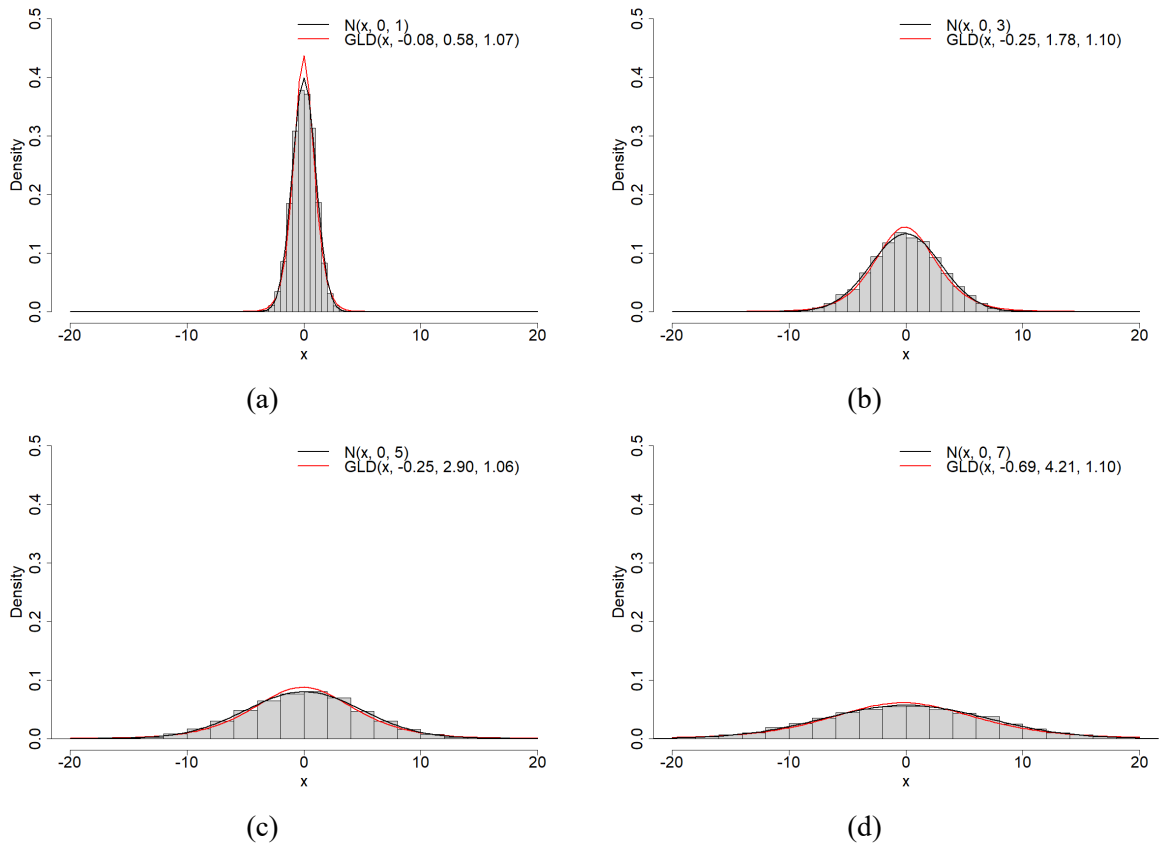


Fig 5.4. Histograms of 10,000 randomly normally distributed data with fitted normal (black line) and GLD (red line) curves. The simulated data all have a mean of 0 with standard deviation (a) $\sigma = 1$, (b) $\sigma = 3$, (c) $\sigma = 5$ and (d) $\sigma = 7$.

5.2 Introducing the Generalized Logistic Distribution Regression Model

This section proposes the GLD regression model, beginning with linear models that have previously been shown to be extensively unsatisfactory such that the statistical assumptions are not always followed (e.g., chapter 4; Figs 4.9 and 4.15). An extension to the linear model is then introduced: GLM. These methods are more flexible and adaptable which have been an underpinning part of the proposed GLD regression model that will be introduced in section 5.2.3.

5.2.1 Linear Models

Considering the modelling of a response Y in terms of some predictors $X_1, \dots, X_n$. This can be modelled in the form of a linear regression model [203] given by:

$$E(Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} = \sum_{j=0}^{p} x_{ij}\beta_j = x_i^T \boldsymbol{\beta} = [X\boldsymbol{\beta}]_i \quad 5.11$$

for $i = 1, \dots n$ where:

$$X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ 1 & x_{2,1} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{2,1} & \cdots & x_{n,p} \end{pmatrix}$$

and $\beta = \left(\beta_0, \dots, \beta_p\right)^T$ is a vector of fixed but unknown parameters describing the dependence of Y_i on x_i. The four ways of defining the linear model in 5.11 are equivalent [204], but the most economical is the matrix form:

$$E(Y) = X\boldsymbol{\beta} \quad 5.12$$

The $n \times p$ matrix X consists of a known (observed) constraint and is commonly referred to as the design matrix. The ith row of X is x_i^T, the explanatory data corresponding to the ith observation of the response. The jth column of X contains the n observations of the jth explanatory variable [204].

Strictly the only requirement for a model to be linear is that the relationship between the response variables Y and any explanatory variables can be written in the form 5.12 [204]. There are various common forms of linear models including the null (Appendix C1), simple linear (Appendix C2), polynomial (Appendix C3) and multiple (Appendix C4) regression models. Despite the structure of the linear model, it is useful for statistical analysis if the three further assumptions [204] are valid:

1. $Y_1, \ldots, Y_n$ are independent random variables,

2. $Y_1, \ldots, Y_n$ are normally distributed and,

3. $Var(Y_1) = Var(Y_2) = \cdots = Var(Y_n)$. Otherwise $Y_1, \ldots, Y_n$ are homoscedastic. This common variance is denoted by σ^2_{norm}.

With these assumptions the linear model completely specifies the distribution of Y, in that $Y_1, \ldots, Y_n$ are independent and $Y_i \sim N(x_i^T \boldsymbol{\beta}, \sigma^2_{norm})$ for $i = 1, \ldots n$. Otherwise, $Y = x_i^T \boldsymbol{\beta} + \varepsilon_i$, where $\varepsilon_i, \ldots \varepsilon_n$ are IID and $N(0, \sigma^2_{norm})$ random variables [204]. The regression coefficients describe the pattern by which the response depends on the explanatory variable. Briefly, in the least squares estimation, the chosen $\widehat{\boldsymbol{\beta}}$ the estimate of $\boldsymbol{\beta}$ will make the estimated means $\hat{E}(Y) = X\widehat{\boldsymbol{\beta}}$ as close as possible to the observed values $\boldsymbol{y}$ [204].

5.2.2 Generalised Linear Model

The GLM [189,191,205] extends the linear model defined in equation 5.12 to allow a more flexible family of probability distributions. The linear regression model is restrictive in that the errors in the model are assumed to be normally distributed amongst other assumptions including independence and homoscedasticity. An extension to the standard linear model is the GLM. The GLM framework is less restrictive and can model more realistic world problems in different application areas. It allows for a wide range of distribution including the normal distribution but many others too. Thus, a degree of non-linearity in the model structure [189] is possible. It has been shown [191] that $y_1, \ldots, y_n$ are observations of the response variables $Y_i, \ldots, Y_n$ assumed to be independently generated by the same exponential distribution with means $\mu_i \equiv E(Y_i)$ linked to the explanatory variables $X_i, \ldots, X_n$ through:

$$g(\mu_i) = \eta_i \equiv \beta_0 + \sum_{r=1}^{p} \beta_r x_{ir} \equiv x_i^T \boldsymbol{\beta} \quad 5.13$$

where g is a smooth monotonic and differentiable function commonly referred to as the 'link function', x_i is a $p \times 1$ vector of explanatory variables such that:

$$x_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix} \text{ so } x_i^T = [x_{i1} \quad \cdots \quad x_{ip}]$$

and $\boldsymbol{\beta}$ is the $p \times 1$ vector of unknown parameters such that:

$$\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

5.2.2.1 The Exponential Family

The nice statistical properties of the normal distribution are shared with a wider range of other distributions called the exponential family of distributions [191]. The exponential family of distributions, amongst many others, includes the Binominal, Gaussian and Poisson distributions (Table 5.1). A distribution can be shown that it belong to the exponential family if the distribution can be put into canonical form [189,205]. The distribution in the GLM framework is an important consideration and depends on the type of data. For a continuous skewed response, for example, the gamma or inverse gaussian would be suitable candidate distributions [205]. However, in some circumstances distributions in the exponential family may have μ_i that is restricted to certain values. To illustrate the Poisson distribution is restricted to $\mu_i \in \mathbb{R}_+$ and the Bernoulli distribution is restricted to $\mu_i \in (0,1)$. If this is not considered correctly there may be a problem with the model such that there may exist a possible x_i and $\boldsymbol{\beta}$ that does not satisfy g such that $\eta_i \neq g(\mu_i)$. Therefore when choosing a link function it is important that the link function maps the set of allowed values for μ_i onto $\mathbb{R}$ [204].

After exploring a range of the common distributions (table 5.1) belonging to the exponential family there are few that are mapped to negative and positive continuous real data. The only distribution that does within table 5.1 is the normal distribution. Whilst the gamma and inverse Gaussian, for example, are intended for continuous skewed response, but neither are mapped to negative numbers [205]. That said, there is clearly a lack of representation in the exponential family that represents a continuous skewed response on $-\infty < x < \infty$. Fortunately, it is not required that the distribution strictly belongs to the exponential family and sometimes context compels another choice. Thus, it is possible to fit distributions that are not in the exponential family using the GLM-style approach, but there are some additional complications [189,205].

5.2.3 Proposed GLD Regression Model

Section 5.1.1 presented the flexibility and capabilities of the GLD which can model skewed continuous data mapped to $-\infty < x < +\infty$. In this section, the GLD is developed in a GLM-style framework and proposes a duel GLM to model both the mean and the dispersion of the data with linear regression structure defined in equation 5.12.

Let $Y_1', \ldots, Y_n'$ be a sample of n continuous independent random variables, where each Y_i' follows a GLD with location θ_i, scale, σ_i and shape α_i, that is, $Y_i' \sim GLD(x; \theta_i, \sigma_i, \alpha_i)$ for $i = 1, \ldots, n$. Thus, the duel GLM using a GLD to model the mean and dispersion is given by:

$$\begin{aligned} E(Y_i') &= g_1'(Y_i') = x_i^T \boldsymbol{\beta} = \eta_i \\ log(Var(Y_i')) &= g_2'(Y_i') = x_i^T \boldsymbol{\beta}' = \eta_i' \end{aligned} \qquad 5.14$$

where x_i are observations, $g_k'(Y_i')$, for $k = 1, 2$ are strictly monotone, differentiable functions called the link functions; that is, they are flat and at least twice differentiable. The functions $g_1'(Y_i')$ and $g_2'(Y_i')$ relate mean and dispersion regression functions to the linear predictor η_i and η'_i, respectively, such that $\theta_i = (g_1')^{-1}(\eta_i)$, with $(g_1')^{-1}(\cdot)$ being the inverse function of $g_1'(\cdot)$ and $\sigma_i = (g_2')^{-1}(\eta_i')$,

with $(g_2')^{-1}(.)$ being the inverse function of $(g_2')(\cdot)$. The set $\boldsymbol{\beta}$ of mean regression coefficients $\beta = \{\beta_0, \quad \beta_1, \quad \cdots, \quad \beta_p\}$ and the set $\boldsymbol{\beta'}$ of dispersion regression coefficients $\beta' = \{\beta'_0, \quad \beta'_1, \quad \cdots, \quad \beta'_p\}$ are vectors of fixed but unknown model parameters describing the pattern by which the response depends on the explanatory variable. In this framework, the identity link function is considered for $g_1'(\cdot)$ and the logarithmic link function is considered for $g_2'(\cdot)$.

This variation of the model in the above was chosen very carefully whilst other variations of the model were considered. In equation 5.14 the mean regression coefficients are retained in $x_i^T \boldsymbol{\beta}$ that are better fitted using the GLD regression model as opposed to other normality methods. This approach also gains additional dispersion regression coefficient in $x_i^T \boldsymbol{\beta'}$ providing information about heteroscedasticity along with the GLD model parameters in θ, σ and α which can create many different complex shapes (section 5.1.1.2) including positive and negative skew. In this form all the model parameters generally converge well, and it allows $x_i^T \boldsymbol{\beta'}$ to be used in the GLD second moment corresponding to variance. Overall, this proposed model is a flexible approach, is applicable to a wider range of applications and represents a significant contribution within GLM domain able to deal with continuous, non-normal data mapped to $-\infty < x < +\infty$.

5.3 Bayesian Framework

Here it is shown how the model proposed in section 5.2.3 will be conducted in a Bayesian framework. First, the prior and posterior distributions of the GLD regression model are introduced followed by the MCMC methods applied to the joint posterior distribution along with Gelman-Rubin convergence techniques to monitor the convergence of the chains. DIC, a Bayesian version of the AIC score [40] is used to observe goodness-of-fit and monitor whether the GLD regression model is a better model than a normal regression model.

5.3.1 Bayesian Inference: Prior Distribution

The GLD regression model in equation 5.14 will be conducted in a Bayesian framework. In a Bayesian framework, prior distributions need to be specified. When no information is provided, assign non-informative priors to the GLD regression model parameters assuming that that the mean and variance parameters $\beta = \{\beta_0, \quad \beta_1, \quad \cdots, \quad \beta_p\}$ and $\beta' = \{\beta'_0, \quad \beta'_1, \quad \cdots, \quad \beta'_p\}$ respectively are normally distributed with mean zero and a large variance to express prior ignorance such that:

$$\beta_0 \sim N(\mu_{\beta_0}, \sigma^2_{norm\,\beta_0}) \qquad \beta'_0 \sim N(\mu_{\beta'_0}, \sigma^2_{norm\,\beta'_0})$$

$$\beta_1 \sim N(\mu_{\beta_1}, \sigma^2_{norm\,\beta_1}) \qquad \beta'_1 \sim N(\mu_{\beta'_1}, \sigma^2_{norm\,\beta'_1})$$

$$\vdots \qquad \vdots$$

$$\beta_p \sim N(\mu_{\beta_p}, \sigma^2_{norm\,\beta_p}) \qquad \beta'_p \sim N(\mu_{\beta'_p}, \sigma^2_{norm\,\beta'_p})$$

For α a gamma distribution is assumed with shape and scale parameters equal to one. These priors are given by:

$$\alpha \sim \Gamma(a_1, b_1)$$

All priors are independent. For the normal distribution the mean parameters are μ_{β_p} and $\mu_{\beta'_p}$ and the variance parameters are $\sigma^2_{norm\,\beta_p}$ and $\sigma^2_{norm\,\beta'_p}$ $for\ p = 0, 1, \ldots$ corresponding to the total number of model parameters. For the gamma distribution it has shape and scale parameters a_1, b_1 respectively.

5.3.2 Bayesian Inference: Posterior Distribution

Statistical inference is performed on the posterior distribution of the parameter vector given by:

$$\psi_{par} = (y'_i; \boldsymbol{\beta}, \boldsymbol{\beta}', \alpha)$$

where $\beta = \{\beta_0, \quad \beta_1, \quad \cdots, \quad \beta_p\}$ is the set of mean regression parameters linked to the linear predictor η_i and $\beta' = \{\beta'_0, \quad \beta'_1, \quad \cdots, \quad \beta'_p\}$ is the set of variance regression parameters linked to the linear predictor η'_i. Given the priors and likelihood function, the joint posterior density function can be expressed as:

$$f_{post}(\psi_{par}|y') \propto f_{like}(y'|\psi_{par}) f_{prior}(\psi_{par}) \tag{5.15}$$

where the likelihood is given by:

$$f_{like}(y'|\psi_{par}) = \prod_{i=1}^{n} GLD(y'_i;\ \boldsymbol{\beta}, \boldsymbol{\beta}', \alpha)$$

and $f_{prior}(\psi_{par})$ is completed with independent prior distributions, given by:

$$f_{prior}(\psi_{par}) = f(\beta_0) \cdot f(\beta_1) \cdot \cdots \cdot f(\beta_p) \cdot f(\beta'_0) \cdot f(\beta'_1) \cdot \cdots \cdot f(\beta'_p) \cdot f(\alpha)$$

The joint posterior distribution was not found to be analytically tractable, thus, to generate samples, MCMC methods were used. Specifically, the random walk Metropolis-Hasting (MH) algorithm. The R code for the priors and join posterior described above are available in Appendix H1. It should be noted that the code was implemented using the simple linear form along with data obtained from the COVID-19 pandemic. Thus, this code will need adapting for other data sets and regression functions other than the simple linear form.

5.3.3 Metropolis Hastings Algorithm

In general, MCMC methods are a broad set of algorithms with application in mathematics and science [206–208], although more recently these methods have had impact in statistics and simulation. They are particularly useful in a Bayesian framework where these methods can be applied when sampling from Bayesian posterior distributions usually referred to as the target distribution [40]. The basic idea

of the MCMC algorithm is that it sequentially samples from a candidate distribution (usually from a well-defined and known distribution) and then in each iteration corrects the previous draw to better approximate the target distribution. Thus, each draw θ^t depends on previous draw θ^{t-1} and eventually converges to the target distribution. The sampling is done iteratively in a way that in each iteration the draws from a candidate distribution become closer to the target distribution [40,209].

In section 5.3.2 the posterior distribution is defined as $f_{post}(\psi_{par}|y')$ where ψ_{par} is the vector of parameters in the GLD regression model and y' denotes the vector of observations. The posterior distribution $f_{post}(\psi_{par}|y')$ may not be analytically tractable making it difficult to draw sample and thus applying MCMC techniques can help to solve this problem. Suppose there exists a distribution $q(.|.)$ such that [40,210,211]:

1. state space of ψ_{par} is the same as in the target distribution;

2. sample draws from $q(.|.)$ converge to the target distribution;

3. random draws from the candidate distribution are easily available.

If the criteria above are met, then, the distribution $q(.|.)$ is a good choice as the candidate distribution [40]. The random walk M-H algorithm proceeds as follows:

1. Draw a starting point ψ_{par}^0 from the candidate distribution based on some approximate estimate;

2. Repeat for $t = 1, \dots, m'$ where m' is the number of MCMC chains and

    a. Draw candidate sample from the candidate distribution such that $\psi_{par}^{cand} \sim q\left(.\,|\psi_{par}^{t-1}\right)$

b. Obtain the acceptance ratio, λ, which is given by:

$$\lambda = \frac{f_{post}(\psi_{par}^{cand}|y')q(\psi_{par}^{t-1}|\psi_{par}^{cand})}{f_{post}(\psi_{par}^{t-1}|y')q(\psi_{par}^{cand}|\psi_{par}^{t-1})} \quad 5.16$$

c. Set

$$\psi_{par}^{t} = \begin{cases} \psi_{par}^{cand} & with\ probability\ \min(\lambda, 1) \\ \psi_{par}^{t-1} & otherwise \end{cases} \quad 5.17$$

5.3.4 Implementation of the Metropolis Hasting Algorithm

In the M-H algorithm, for regression coefficient parameters in $x_i^T\boldsymbol{\beta}$, and the dispersion regression coefficients in $x_i^T\boldsymbol{\beta}'$,the proposal distribution is chosen to be normal such that:

$$\boldsymbol{\beta}_p \sim N(\mu_{\beta_p}^{(t-1)}, \sigma_{norm\,\beta_p}^2)$$

$$\boldsymbol{\beta}'_p \sim N\left(\mu_{\beta'_p}^{(t-1)}, \sigma_{norm\,\beta'_p}^2\right)$$

where $p = 0, 1, \ldots$ is the number of parameters in each of the mean and dispersion vectors for some value of $\sigma_{norm\,\beta_p}^2$ and $\sigma_{norm\,\beta'_p}^2$. For the parameters α, the proposal distribution is chosen to be a gamma distribution such that:

$$\alpha \sim \Gamma(a_1^{(t-1)}, b_1^{(t-1)})$$

where $a_1^{(t-1)} = \frac{(\sigma^{t-1})^2}{(\sigma_\sigma^2)}$ and $b_1^{(t-1)} = \frac{(\sigma^{t-1})^2}{(\sigma_\sigma^2)}$.

In the t^{th} iteration, for $t = 1, \ldots, m'$, where m' is the number of MCMC iterations, the proposal distributions are centred at the value from the previous iteration for some suitable value for

the variance parameter [212]. An important consideration is the choice of the dispersion parameters for each of the proposal distributions. In summary, a consequence is that a large variance will produce large variation in iteration t from the previous iteration $t-1$ and subsequently will lead to a small acceptance rate λ. However, a small variance will produce small variation in iteration t from the previous iteration $t-1$ and will lead to high acceptance rate λ [209]. It has been proposed [212,213] that the optimal choice for the variance in the circumstance that the proposal is normal to be $\sigma_{norm}^2 = c^2\Sigma$ where $c \approx 2.4/\sqrt{d}$ where d is the dimension of the parameters vector ψ_{par} and Σ is the variance-covariance matrix based on the curvature of the posterior at the mode.

5.3.5 Gelman Convergence Diagnostics

It is important that the chains in the MCMC converge to the target distribution. This can be monitored using Gelman and Rubin's convergence diagnostics [210,211]. The convergence in MCMC chains is diagnosed when the chains have bypassed the initial values set at the beginning of the algorithm and the output of all chains are indistinguishable. The number of chains can be increased until convergence occurs. In this analysis an R statistic is computed which indicates whether convergence has occurred. If the R statistic is less than 1.1, then, the chains are stationary whilst an R statistic greater than 1.1 indicates the chains are non-stationary.

5.3.6 Deviance information criterion (DIC)

To assess the performance of the Bayesian GLD regression model and to compare it to the existing normal linear regression model, DIC is used. The DIC is a Bayesian version of the AIC [40]. To recall, the AIC is given by:

$$AIC = -2logf(y'|\hat{\psi}_{par}^{mle}) + 2k_{pen} \qquad 5.18$$

where $f(y'|\hat{\psi}_{par}^{mle})$ is the MLE (rewarding goodness of fit) and k_{pen} is penalty that is an increasing function of the number of estimated parameters. To obtain the DIC two changes are made to equation

5.18 such that the MLE $\hat{\psi}_{par}$ is replaced with the posterior mean $\hat{\psi}_{par} = E(\psi_{par}|y')$ and k_{pen} is replaced with a data-based bias correction. Therefore, the DIC is defined as:

$$DIC = -2logf(y'|\hat{\psi}_{par}^{Bayes}) - f_{DIC} \quad 5.19$$

where $f(y'|\hat{\psi}_{par}^{Bayes})$ is the GLD likelihood function and f_{DIC} is the effective number of parameters defined as:

$$f_{DIC} = 2\big((logf(y'|\hat{\psi}_{par}^{Bayes}) - E_{post}(logf(y'|\psi_{par}))\big) \quad 5.20$$

where the expectation in the second term is an average of ψ_{par}over its posterior distribution which can also be computed using the simulations $\psi^{m\prime}$, $m' = 1, \dots M$ where M' is the number of MCMC iterations such that:

$$computed\ f_{DIC} = 2\left(logf_{like}(y'|\psi_{par}) - \frac{1}{M'}\sum_{m\prime=1}^{M\prime} logf_{like}(y'|\psi_{par}^{m})\right) \quad 5.21$$

where $f_{like}(y'|\psi_{par})$ is the GLD likelihood function and ψ_{par} is the parameter vector space. Like the AIC score, a lower DIC score corresponds to a better fit.

5.4 Bayesian GLD Regression Model Remarks

In this chapter, a Bayesian GLD regression model has been developed to model expectation and dispersion of continuous data in a regression setting. It is a flexible and variable proposed methodology capable to apply to a range of disciplines. It provides mean regression coefficient and gains additional dispersion regression coefficient relating to heteroscedasticity along with the GLD location, scale and shape parameters accounting for complex distributions such as positive and negative skewing. In previous normality methods applied in chapters 3 and 4, assuming a normally

homoscedastic system produces large model bias, relative to the improved Bayesian GLD regression model.

The GLD regression model developed in chapter 5 can model linear regression using a range of functions, amongst others, null (Appendix C1), simple linear (Appendix C2), polynomial (Appendix C3) and multiple (Appendix C4) linear regression. The only requirement is that the linear function can be written in the form of equation 5.12. The generalised form allows a wide range of applications without homoscedastic, linear and normality assumptions. Instead, in the developed dual GLD regression model (mean and dispersion) it is assumed that y_i' is a GLD such that $y_i' \sim GLD(x; \theta, \sigma, \alpha)$ to accommodate a wider range of shapes (Figs 5.1-5.3). This Bayesian GLD regression model will be demonstrated using PLs and COVID-19 data from the pandemic. This will be conducted in chapter 6, although it is worth emphasising that the application is much wider.

5.5 Summary and Conclusions

In a normal regression model, it is assumed that residuals are normal, homoscedastic and linear. Chapter 4 clearly showed that residuals sometimes fail to meet these assumptions and demonstrated a need for a more flexible approach. The GLM is an extension of the standard linear model that allows the response, usually from the exponential family, to be a distribution other than the normal distribution. However, other distributions outside the exponential family can be considered, where context compels a different choice. In chapter 4, the GLD was explored, which better characterized the full sweep of the COVID-19 pandemic allowing for positive and negative skew, otherwise, non-normally distributed residuals. However, a limitation acknowledged within chapter 4 was that the GLD residual model (section 4.2.2) was applied after using normality methods.

This chapter continued to provide an overview of the GLD: a simple, flexible, variable and widely applicable distribution. It is mathematically a simple distribution in that the form is a single expression. It is a well-studied distribution with MOM that can easily derive moments such as the

expectation, variance and skew. Considering the flexible nature of the GLD, in this chapter a dual Bayesian GLM-style GLD regression model was explored.

In the proposed framework, expectation and dispersion is modeled through some linear regression function (e.g., null, simple linear, polynomial, multiple regression). Bayesian expectation and dispersion estimations of the parameters of interest are obtained using MCMC methods and it is no longer required to have linear, normality and homoscedastic assumptions. Thus, the developed Bayesian GLD regression model is a more flexible approach capable of looking at skewed or otherwise non-normally distributed continuous data. In this approach a complicated non-analytically tractable posterior distribution is obtained, thus, Bayesian model parameter estimates are fitted using MCMC methods. The DIC, a Bayesian version of Akaike information criterion, are computationally obtained using the MCMC chains. These methods are used to monitor goodness-of-fit and as a comparison tool to observe model preference.

In the developed Bayesian GLD regression model, expectation regression coefficients are retained. However, Bayesian expectation estimates obtained through the mean regression coefficients are better fitted using the GLD regression model developed in this chapter. This is because the assumptions of standard linear regression are not met with skewed and heteroscedastic data, thus in this case, applied normality methods produce fundamental and systematic model bias. This occurs in many studies in a range of disciplines. Therefore, the methodology developed in this chapter will benefit many studies using linear regression that do not meet the requirements of standard linear regression. Nevertheless, in scenarios that do meet standard linear regression requirements the Bayesian GLD regression model, in general, is still a more flexible framework, able to model both the expectation and dispersion.

The Bayesian dispersion estimates obtained through the regression coefficients are indicative of heteroscedasticity along with the GLD location, scale and shape parameters which can characterise a range of complex shapes including positive and negative skew. If the shape parameter, $\alpha < 1$ the

distribution is left-skewed and if it was $\alpha > 1$ the distribution is right-skewed. The simulation in this study shows that the GLD can also model symmetrical data like the normal distribution, although the tails of the GLD distribution are slightly heavier. Overall, the additional information, obtained through the extra parameters will help further understand the data sets under examination.

The framework is flexible and applicable to a wide range of disciplines using regression. The developed framework models both the expectation and dispersion of continuous data. The modelling of dispersion is less common in the GLM-style framework and has never been attempted using the GLD. The intractable form of the posterior distribution has never been dealt with in a regression setting, thus, the overall framework developed in this chapter using MCMC techniques deals with the complex mathematical forms never seen before. The Gelman convergence diagnostics and DIC methods also deal with the intractable framework proposed in this chapter.

In chapter 6, the developed methodology within this chapter will be tested and illustrated using PLs and data from the COVID-19 pandemic, running the developed model over an extended timeframe. This will allow convergence to be monitored over a complex and extended timeseries, and to compare the results with previous normality methods (chapter 4).

5.5.1 Possible Extensions and Wider Application

The GLD regression model is widely applicable since the only criteria is that the regression structure can be expressed as equation 5.12. To exemplify possible extensions, consider the dementia data from chapter 3. The data were modelled using a segmenting regression model in a PL setting (Fig 5.5(a)). Briefly, to account for segmenting behavior, expectation is fitted below and above a critical point. Future work on the GLD regression model must consider segmenting behavior to extend the systems that would benefit using the proposed framework. This extension would increase the parameter vector space to account for fitting below and above the critical point modelling both the expectations and dispersion separately. In the context of rural-urban scaling the reasoning behind a consistent and

persistent critical point for a host of indicators needs further work to be able to explain why rural-urban regions are fundamentally different. Alternatively, using the same dementia data, a polynomial regression function (Fig 5.5(b)) could fit the data. For example, the cubic fitting in figure 5.5(b) would include additional expectation and dispersion regression coefficients in a PL setting with a quadratic or cubic fitting. Although scientifically a cubic is perhaps not required since the GLD model is capable of this fitting as well as more complex forms of heteroscedasticity. Although, for now, the purpose of figure 5.5(b) is to demonstrate the general application of the model with cubic structure. Overall, the generalization in the proposed GLD regression model allows for a range of systems with linear regression but would be enhanced with adaptation to the framework allowing for segmentation and other relationships.

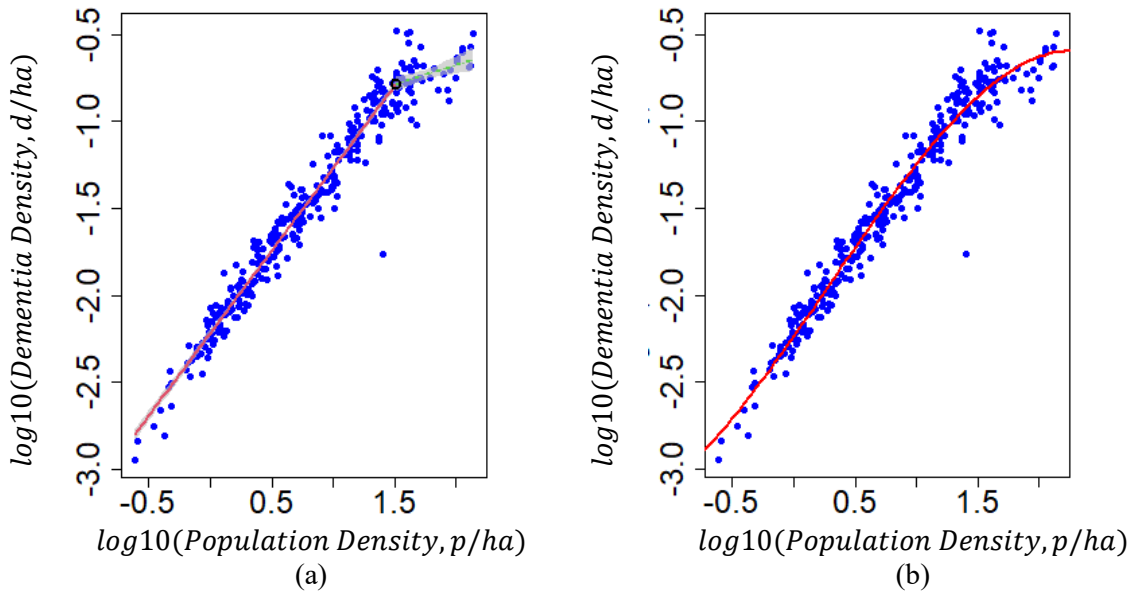


Fig 5.5. Log transformed dementia data with a segmented fit (a) and a cubic polynomial fit (b). The blue dots represent LTLA (residuals). The red and green lines in (a) are linear regression lines below and above the critical point respectively. The critical point is represented in the black circle. The red line in (b) represents a cubic fit of the form $y' = \beta_0 + \beta_1(x) + \beta_2(x)^2 + \beta_3(x)^3$.

Other work should include a residual analysis, to continue to explore better defined residuals obtained from the Bayesian GLD regression model and compare them to previous normal regression models to investigate how they vary. This will also be useful in a goodness-to-fit analysis to continue to test the GLD regression model and to further establish its place in the statistics realm.

5.5.2 Limitations

Overall, the GLD regression model is widely applicable to many different disciplines using linear regression. It is a significant improvement to normal regression models such that the model no longer requires assumptions of homoscedasticity and normality which have been fundamentally biased in a restricting framework. The many benefits and improvements using the developed GLD regression model, have been shown, but limitations must be acknowledged that need addressing, ideally placed as development opportunities and future work.

In chapter 5, a GLM was developed, where the response Y is assumed to be a GLD in a Bayesian framework. The GLD is mapped to continuous data (i.e., $-\infty < x < \infty$) and thus it inherently models expectation and dispersion in the case when X is continuous. The model only considers continuous data and exhibited the key features using a real continuous data set (chapter 6). Thus, the proposed model would be considerably better exhibited for the case when X is discrete. This is an opportunity for future work. This is a recommendation along with suggestions made in section 5.5.1 to account for segmented and cubic behaviour.

Chapter 6: Application of the GLD Regression Model

6.1 Introduction

To illustrate the key features of the Bayesian GLD regression model developed in chapter 5 and to demonstrate the wide application, the model is run and tested using COVID-19 data. This data set is the same data previously analysed in chapter 4, where the statistical assumptions of standard linear regression were not met. It was shown that residual variance behaved in a complex way, expanding and contracting throughout the pandemic. Currently, the modelling of COVID-19 has been done using positively skewed distributions such as the Weibull [141], Poisson [142] and Gamma [214] distribution. It was shown that these distributions were unable to fully characterise the full sweep of the COVID-19 pandemic. Instead, it was found that in large timeframes of the pandemic the residuals were not normally distributed and negatively skewed. Residuals were better characterised by a GLD. They exhibited positive and negative skew corresponding to certain events and government intervention. Positive skew was indicative of 'hotspots' and 'superspreading' events whilst negative skew was indicative of 'coldspots' and 'superisolating' events: both important features of propagation with the latter underappreciated in the modelling of COVID-19.

In this chapter, the Bayesian GLD regression model, developed in chapter 5, is applied to PL models and data from the COVID-19 pandemic to explore the additional information and how this

compares with previous methods. In a normal PL regression setting, the assumptions of homoscedasticity and normality are false, thus, using the GLD regression model, mean regression model parameters are adjusted to account for periods of heteroscedastic and non-normally distributed residuals. The additional dispersion parameters indicative of heteroscedasticity changes from increasing with magnitude to decreasing with magnitude depending on the day of the pandemic corresponding to positive and negative skew respectively.

In previous chapters, it has been shown that conventional methods inherited within scaling using least squared methods are fundamentally bias due to the restrictive framework and following a set of limiting statistical assumptions. For example, inference on estimated parameters to test the reliability, are assessed based on the assumption that the variance in the error term ε is σ_{norm}^2 such that $\varepsilon \sim N(0, \sigma_{norm}^2)$. Meanwhile, an initial solution to this problem has been proposed, highlighting the differences in reported exponents when assuming a fixed or free fluctuation [37]. In this work, authors show differences in estimated exponents depending on the model in consideration. Using Leitão methods, this has also been investigated in chapters 3 and 4, showing the limitations of restricting the variance. However, to date, the statistical modelling is built around the Gaussian distribution. Analysis of daily COVID-19 data have shown that this is insufficient. Consequently, chapters 5-6, continues to investigate a generalisation of fluctuations and proposes that modelling is built around the GLD. The GLD has been shown to be flexible allowing of positive and negative skew. Specifically, it is no longer a requirement that the error term in the model is $\varepsilon \sim N(0, \sigma_{norm}^2)$ and instead is now assumed $\varepsilon \sim GLD(\theta, \sigma, \alpha)$. This is built into a GLM-style framework using Bayesian methods and, in this form, includes a shape parameter α allowing for complex shapes (positive and negative skew). As a result of the modelling developed in chapter 5 and discussions relating to variance throughout, there is reservations in the conclusions drawn in chapters 3 and 4. Thus, it will be worthwhile revisiting these datasets in future work, retrospect to the progress made in the modelling process.

6.2 Dataset and R Packages

To explore the methodology developed in chapter 5, COVID-19 data was applied which has previously been explored using standard PL models in chapter 4 where it was found that the statistical assumptions of standard linear regression were not met. Section 4.3.1 describes the data and methods in more detail. To summarize, the data set includes COVID-19 case densities (i.e., COVID-19 cases/Regional Area) and population densities (i.e., Population/Regional Area) in England and Wales using LTLAs (337) to define area boundaries. This dataset presents an ideal opportunity to run the proposed model over an extended time series under different conditions such as a varying number of observations to illustrate the key features of the model. It is also an opportunity to compare the GLD regression model results with previous results using standard normal linear regression techniques. There are in total 337 lower tier local authorities and 759 days of data beginning on the 01/03/2020 and ending on the 29/03/2022.

The data were analysed using the statistical software R version (3.6.2) [157] with the glogis (1.0-1) [170], rio (0.5.27) [215], xlsx (0.5.7) [162], matrixStats (0.61.0) [216], lubridate (1.7.9.2) [167], fitdistrplus (1.1-3) [168], fGarch (3042.83.2) [217] and coda (0.19-4) [218] packages.

6.3 Application of the GLD Regression Model

To recall, to obtain parameters estimates under a PL framework the data are log transformed and forms a simple linear regression model. To recall the standard population density PL model is given by:

$$Y_D = Y_0 {P_D}^{\beta} 10^{\varepsilon} \qquad 6.1$$

and the logarithmic version of 6.1 is:

$$\log(Y_D) = \log(Y_0) + \beta \log(P_D) + \varepsilon \qquad 6.2$$

where $Y_D = Y/A$ and $P_D = P/A$ is the population and indicator density where A in the area of a given region and ε are residuals that are IID with common $N(0, \sigma^2)$. The structure in equation 6.2 forms a simple linear regression. Thus, rewriting this in standard regression form gives:

$$Y_i' = \beta_0 + \beta_1 x_i' + \varepsilon_i \qquad 6.3$$
$$E(Y_i') = \beta_0 + \beta_1 x_i'$$

where $Y_i' = \log(y_i)$, $\beta_0 = \log(y_0)$, $\beta_1 = \beta$, $x_i' = \log(N_i)$ and $\varepsilon_i' = \log(\varepsilon_i)$. In this form, the exponent in β_1 indicates the type of scaling relationship and whether it is sublinear ($\beta_1 < 1$), linear ($\beta_1 = 1$) or superliner ($\beta_1 > 1$). However, it has been shown that statistical assumptions, using normality methods, are false, and consequently this caused estimated model parameters to be fundamentally bias.

In chapter 5, the GLD regression model was developed where it is assumed that the response Y_i' is a GLD such that $Y_i' \sim GLD(x; \theta, \sigma, \alpha)$ to accommodate a wider range of distributions including positive and negative skew. Thus, using the model in chapter 5 and the data from the COVID-19 pandemic equation 5.14 can be rewritten in simple liner regression form using the GLD regression model as follows:

$$E(Y_i') = g_1'(Y_i') = x_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_i = \eta_i$$
$$log\big(Var(Y_i')\big) = g_2'(Y_i') = x_i^T \boldsymbol{\beta}' = \beta_2' + \beta_3' x_i = \eta_i' \qquad 6.4$$

where the estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ for β_0 and β_1 respectively, are the mean regression model parameters, and estimates of $\hat{\beta}_2'$ and $\hat{\beta}_3'$ for β_2' and β_3' respectively, are the dispersion regression model parameters along with x_i *log(population densities)* and y_i' *log(COVID-19 Case)* densities. In this model mean regression coefficients (β_0 and β_1), previously obtained in chapter 4, known as the pre-exponential factor and exponent are retained respectively. The exponent indicates the type of scaling

behavior and whether it is sublinear, linear or superliner. Additional dispersion regression parameters (β_2' and β_2') indicative of heteroscedasticity along with the GLD model parameters: location θ, scale σ and shape α were also obtained.

6.4 Implementation

In the M-H algorithm the first phase of the algorithm is often referred to as the warm-up (or burn-in) phase, thus, it is often that these draws are disregarded. The number of iterations is set to $m' = 20{,}000$ and the first 10,000 draws are disregarded as burn-in, and convergence is monitored in the MCMC chains using Gelman and Rubin's convergence statistics. Briefly, the statistic obtained is known as the R statistic indicating whether the chains in the MCMC are stationary. As, proposed in section 5.3.4, the optimal standard deviation for the normal distribution is $\sigma_{norm}^2 = c^2\Sigma$ where $\approx 2.4/\sqrt{d}$ where d is the dimension of the parameters vector ψ_{par} and Σ is the variance-covariance matrix based on the curvature of the posterior at the mode. Considering that the parameter vector space is $\psi_{par} = (y_i'; \beta_0, \beta_0, \beta_0', \beta_0', \alpha)$, thus, $d = 5$, $c = 1.07$ and $\Sigma = 0.01$ it implies that the optimal standard deviation for the normal proposal is 0.12.

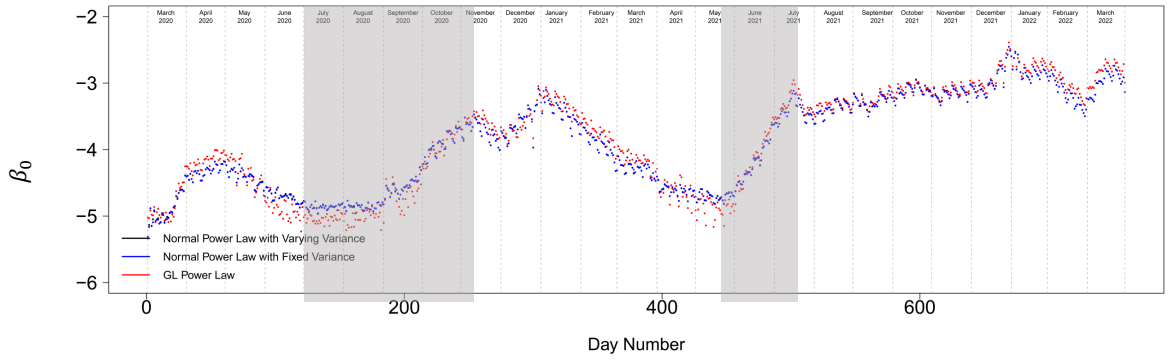
Overall convergence was recognised for the mean regression coefficients β_0 and β_1 along with all GLD model parameters θ, σ, α. Although, dispersion regression coefficients β_2' and β_3' show some signs of convergence, they did so to a much lesser extent. The complete daily Gelman and Rubin's diagnostics are available in Appendices E1 and E2 along with daily MCMC trace plots in Appendix E3 and distributions of the estimates of the posterior distributions for each parameter in appendix E4. The R code to implement the walking M-H algorithm as described above is available in Appendix H2.

6.5 Key Features

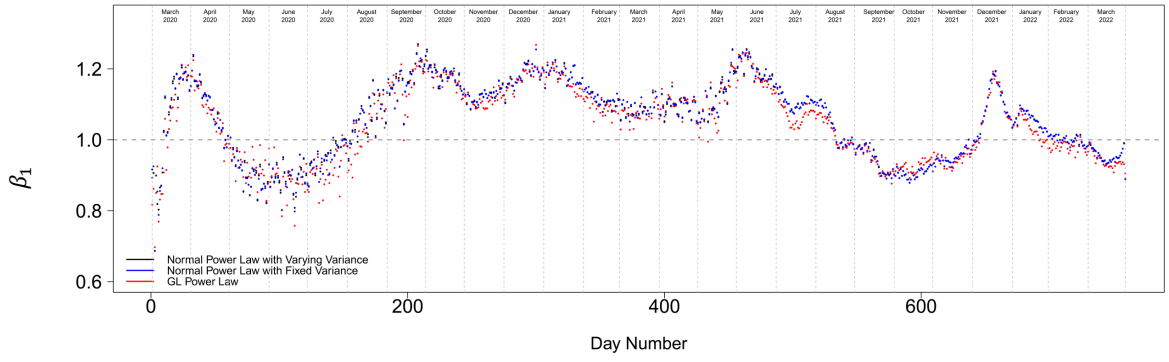
In this specific COVID-19 application, the dual Bayesian GLD regression model (equation 6.4) has seven parameters (β_0, β_1, β_2', β_3', θ, σ and α). Specifically, it has mean regression parameters β_0 and β_1, dispersion regression parameters β_2' and β_3' along with the GLD location θ, scale σ and shape α parameters. The daily model parameters (β_0 (Fig 6.1), β_1 (Fig 6.1), β_2' (Fig 6.2), β_3' (Fig 6.2), θ (Fig 6.3), σ (Fig 6.3) and α (Fig 6.3) are presented as an extended timeseries to assess the trajectory and structure. The GLD regression model parameters (represented as red), when possible, are compared with the normal simple linear regression model with fixed variance (represented as blue) and the normal simple linear regression model with varying variance (represented as black) (see methods in Appendix D1). The mean posterior and standard deviation along with credible intervals are available in Appendix F1.

The general tracking of the mean regression coefficients β_0 and β_1 presented in figures 6.1(a) and 6.1(b) are remarkably similar in both the normal (with fixed and varying variance) and the GLD regression models. The exception in β_0 indicated in the grey shaded area (figure 6.1(a)) coincide with a lower number of regions reporting cases. The fitting of β_0 and β_1 using both normal methods with varying and fixed variance appear identical. Thus, to expose the differences between the normal and GLD regression model the differences were computed in β_0 (Fig 6.1(c)) and β_1 (Fig 6.1 (d)) between the standard normal (chapter 4) and the Bayesian GLD regression models. This presentation better highlighted the variations in β_0 and β_1 between the two models at different times during the pandemic. If differences are greater than 0 then the GLM has decreased the mean intercept (β_0) and mean slope (β_1) whilst values less than 0 indicates an increase. The magnitude of the difference in the parameter β_0 between the model significantly alters in the summer 2020 and spring 2021 coinciding with fewer regions reporting cases whilst alteration in the β_1 parameter is much less. Furthermore, these differences in model parameters reflect the reduced bias in estimated parameters when compared to standard linear regression. The flexibility in the error terms ε allows for features of skew by following the GLD. Furthermore, the generalisation in the assumed variance, allows for variation

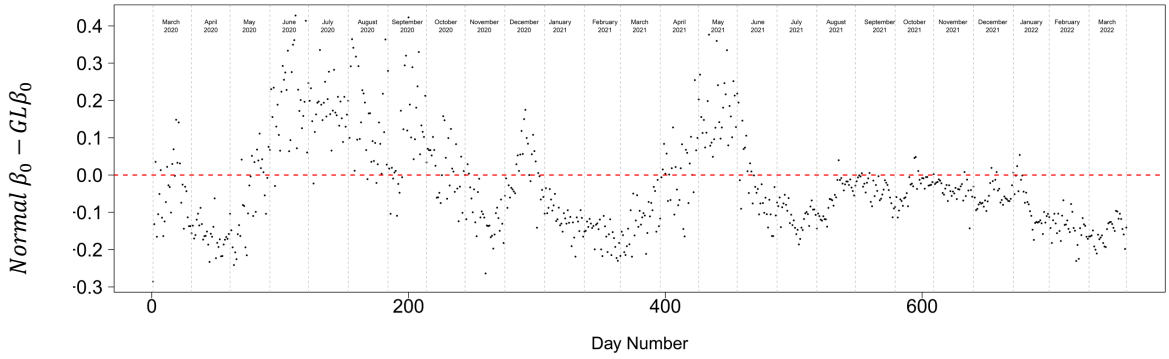
from day-to-day, in comparison, to standard methods where it is strictly fixed throughout the pandemic. The corresponding linear regression plots using standard linear regression and GLD regression parameters are available in Appendix G1.



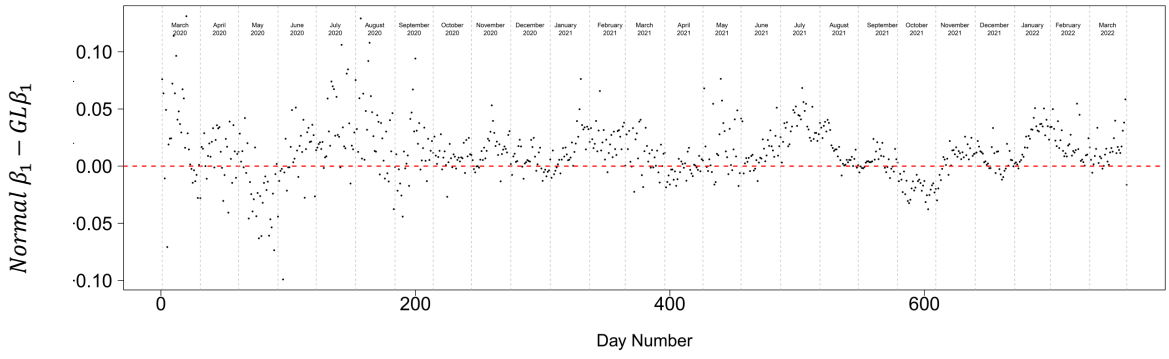
(a)



(b)



(c)



(d)

Fig 6.1. Daily time series of model parameters (panel (a) and (b)) and corresponding differences (panel (c) and (d)). Red represents GLD regression model; blue represents normal regression model with fixed variance and black represents normal regression model with varying variance. The grey shaded areas in (a) indicate time periods with fewer regions reporting cases. The horizonal grey dashed line in (c) and (d) represents no change in the estimated parameters. The studied timeframe is between 01/03/2020 to the 29/03/2022.

A key advantage given by the GLD regression model is the additional parameters accounting for heteroscedasticity in the data (Fig 6.2). The structure in the dispersion parameter β_2' between the normal and GLD regression models are almost identical, but significantly reduced in the latter. Whilst comparisons in the dispersion parameter β_3' between the two models are very different. The fitting of the β_3' in the normal regression model is almost 0 for the entire timeseries but considering that the normal methods assume a homoscedastic system this result was anticipated. Whilst the reported β_3' in the GLD regression model indicates periods of heteroscedasticity in the data. This result is a significant improvement to previous models using normality methods where homoscedastic assumptions do not hold all the time. The reported β_3' indicating heteroscedasticity is prominent for approximately 17 months of the timeseries reaching as high as 2.156 (05/03/2020). Generally, a positive β_3' indicates that residuals with a high x_i deviate from the expectation $E(Y')$ whilst homogenising residuals with a low x_i (Fig 6.2(c)). A negative β_3' has the opposite effect where low $x_i's$ deviate from $E(Y')$ whilst residuals with high $x_i's$ homogenise (Fig 6.2(d)).

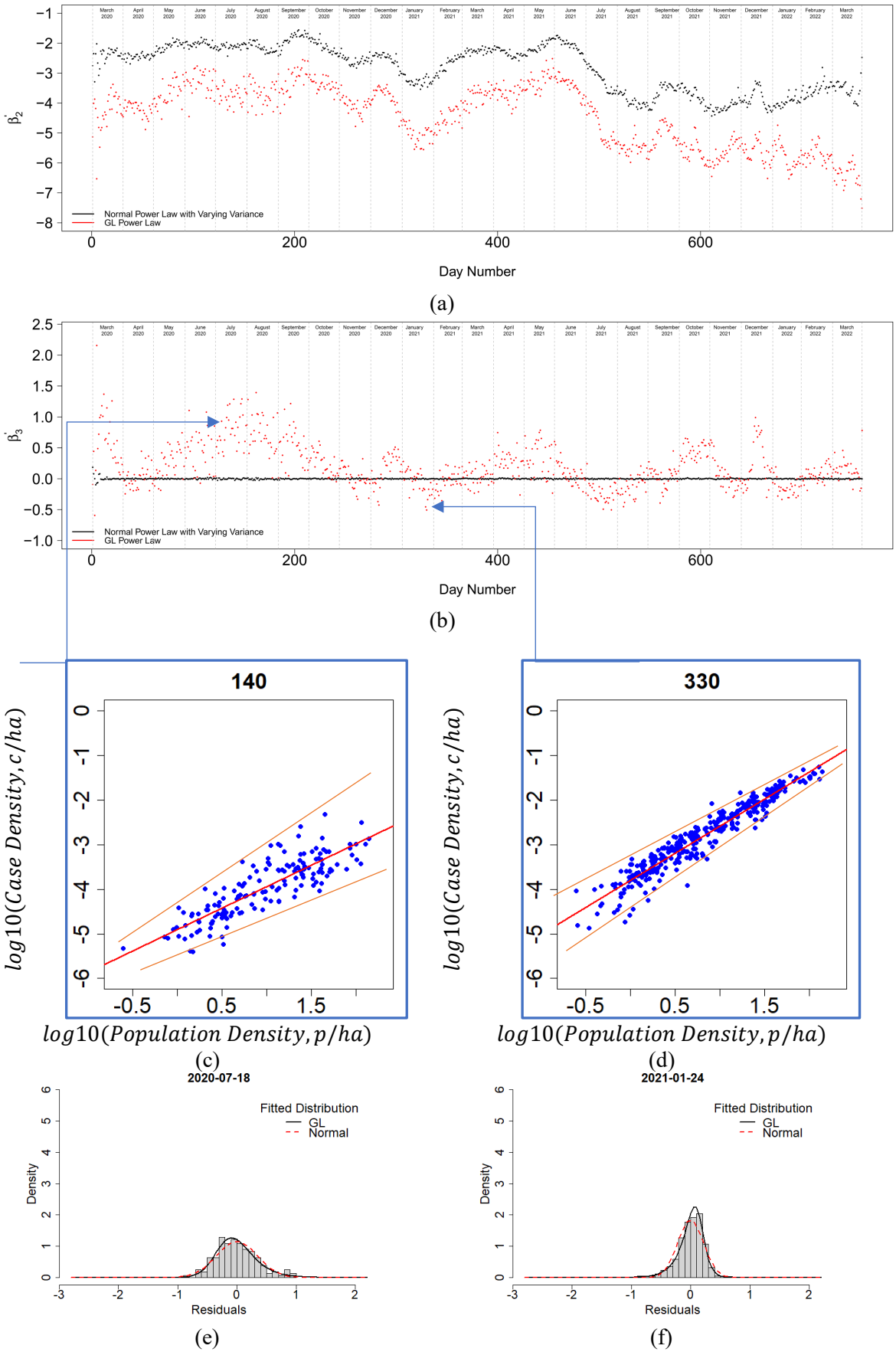


Fig 6.2. Daily timeseries of the (panels (a) and (b)) dispersion parameters and examples of heteroscedasticity (panels (c) and (d)) corresponding to skewed residuals (panels () and (f)). The

daily reported (a) β_2' and (b) β_3' dispersion model parameters. Red represents the GLD regression model and black represents normal regression model with varying variance. Examples include (c) positive β_3' indicating heteroscedasticity affecting high value of x, mostly corresponding to (e) positively skewed residuals and (d) negative β_3' indicating heteroscedasticity affecting low value of x, corresponding to (f) negatively skewed residuals. The studied timeframe is between 01/03/2020 to the 29/03/2022.

In the context of COVID-19, the GLD model provides additional information about the pandemic and the propagation of cases throughout England and Wales. In time frames with positive heteroscedasticity (i.e., $\beta_3' > 0$), rural regions homogenised whilst urban regions deviate from the expectation $E(Y')$. This happens in March 2020, summer 2020, December 2020, spring 2021, October 2021 and December 2021. Before the propagation of delta and omicron variants, a more transmissible version of the disease, these observations correspond to positively skewed residuals; a long right tail of 'hotspots' and 'superspreading' events mostly featuring urban regions (For example on day 140 in Fig 6.2(c, e)). Conversely in time frames with negative heteroscedasticity (i.e., $\beta_3' < 0$), urban regions tended to homogenise whilst rural regions deviate from the expectation $E(Y')$. This behaviour is observed in the periods: April 2020, November 2020, February 2021, July-August 2021, November 2021 and January 2022. These timeframes correspond to negatively skewed residuals; a long-left tail of 'cold spots' and 'super isolating' events mostly featuring rural regions (For example on day 330 in Fig 6.2(d, f)).

Another key advantage of the GLD regression model is the GLD location parameter θ (Fig 6.3(a)), scale σ (Fig 6.3(b)) and the additional shape α (Fig 6.3(c)) parameters which can accommodate a wide range of distributions in the data including positive and negative skew (Figs 5.1-5.3). This is a significant improvement considering that residuals are not always normally distributed in a standard linear regression setting and can exhibit far more complexity (chapter 4; Figs 4.9, 4.12 and 4.15). To draw upon the differences between normal and GLD regression, model parameters, where possible, were obtained for the mean μ (location) and variance σ^2 (scale) (model in Appendix D1) from the normal regression model with varying variance.

The discontinuity from day to day in θ (Fig 6.3(a)) are during periods of time when there are fewer regions reporting cases. This happens at the beginning of the pandemic, in the summer 2020 and spring 2021 (Fig 4.6(a)). This is an example when scaling is not in equilibrium and is in a transition state. Overall, the location (Fig 6.3(a)) and scale (Fig 6.3(b)) parameters in both models are similar although slightly elevated in the case for scale. To account for skew in the model, the GLD regression model also has a shape (α) parameter (Fig 6.3(c)). This is not comparable with the normal regression model since it is assumed that residuals are normally distributed, thus, a shape parameter is not needed and therefore does not exist in the normal regression model. To recall, the shape (α) parameter in the GLD regression model indicates the type of distribution and whether there exhibits any positive or negative skew in the data. If $\alpha < 1$ then this indicates negative skew, whilst if $\alpha > 1$ indicates positive skew. This varied considerably from very positive (Fig 6.3(d)) to very negative (Fig 6.3(f)) and in the two cases the distributions are characterised with either a long left or right tail. This result is significant. It illustrates another key feature of the GLD regression model such that it can accommodate a wide range of shapes, no longer needing normality assumptions and consequently the model is extensively accessible to a wide range of systems.

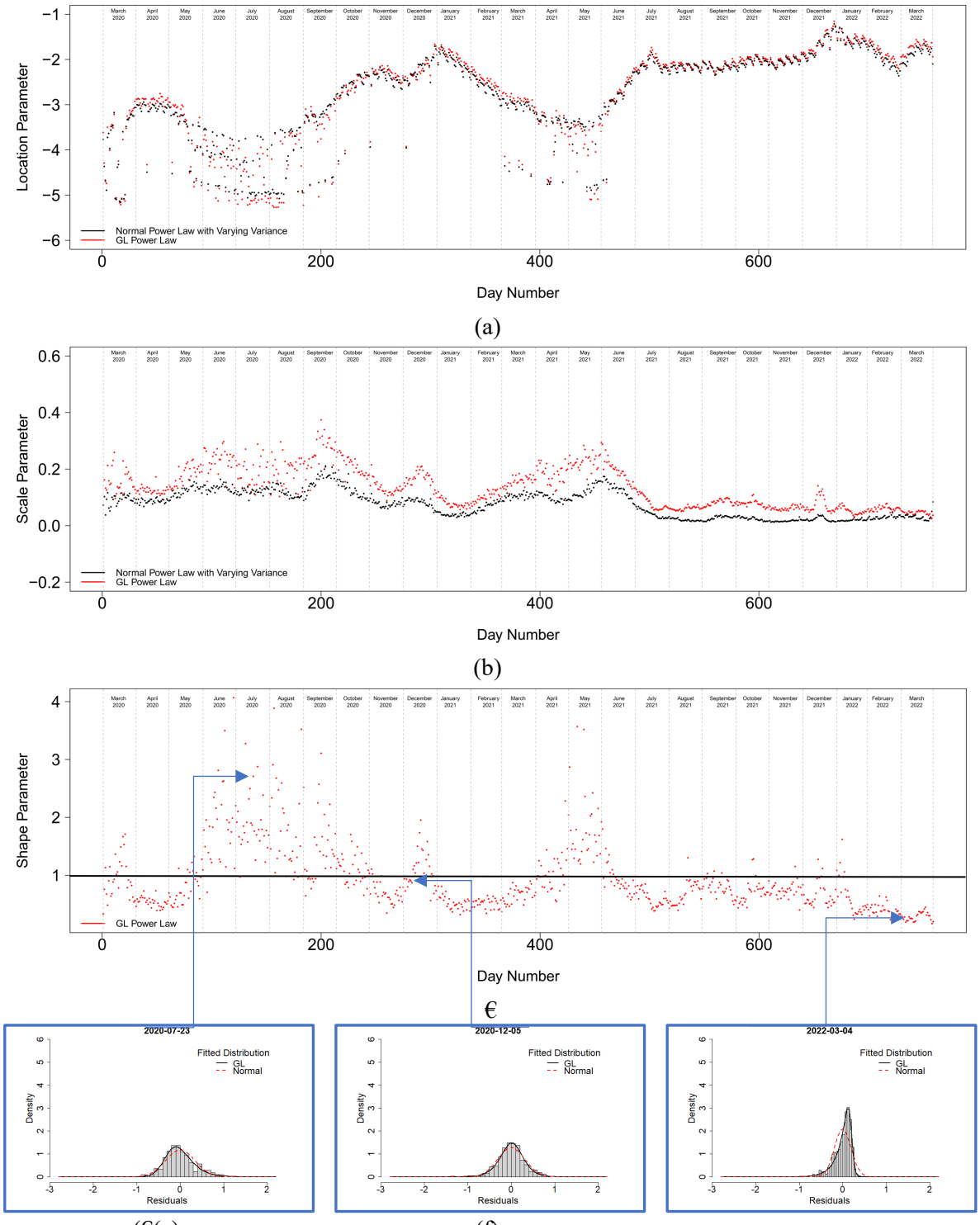


Fig 6.3. Daily time series of (a) location (b) scale and (c) shape parameters and examples of residual distributions with (d) positive €w, (e) no skew and (f) negative skew. Red points in panels (a)-(c) represent the GLD regression model and black points represent the normal regression model with varying variance. The red curve in the panels (d)-(f) represents the normal distribution and the black curve represents the GL distribution. The horizontal black line in (c) indicates no skew. The studied timeframe is between 01/03/2020 to the 29/03/2022.

6.6. Model Checking

The DIC score was used to measure the goodness of fit (section 5.3.6). When selecting a model, low DIC scores represent a better fit. The differences between the DIC scores obtained from fitting the two models (normal with varying variance and GLD) were obtained for each day in the 25-month period (Fig 6.4). A positive value corresponds to a GLD regression model as the preferred model whilst a negative value corresponds to a normal regression model as the preferred model. The daily histograms for cases can be found in Appendix B9.

The near universal preference of the GLD regression model towards the end of the time series corresponds to a homogenised and negatively skewed presentation of residuals. This happens during the UK propagation of delta and omicron variants. These variants are known to be more transmissible versions of the disease compared to previous versions and caused a 'surge' in the number of cases. Whilst preference to a normal regression model occurs when n, the number of regions reporting cases, is not at full capacity (indicated in the grey shaded areas in Fig 6.4). Overwhelmingly, the GLD regression model is the preferred model when n tended towards full capacity, such that all regions were reporting at least 1 case of COVID-19.

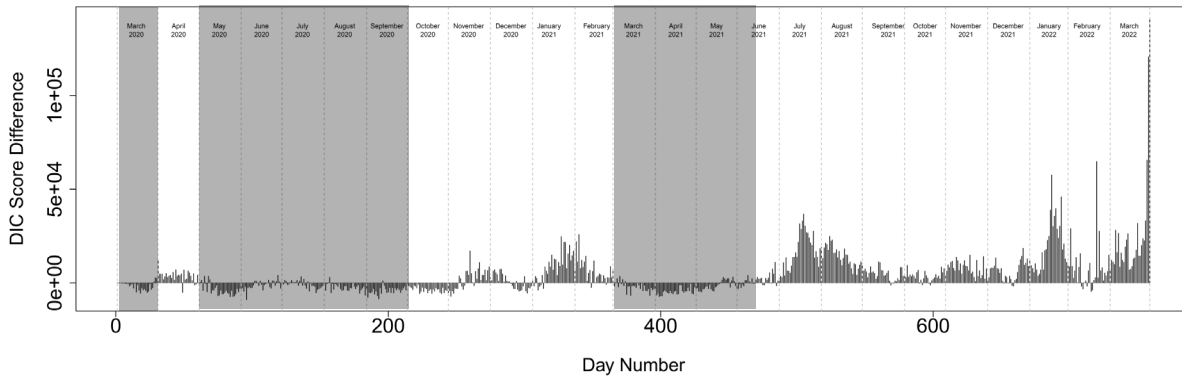


Fig 6.4. DIC score over time. Positive DIC indicates GL regression is a better fit and a negative DIC indicates normal regression model is a better fit. The grey shaded periods correspond to fewer regions with reported cases.

6.7 Summary

To demonstrate the capabilities and the key features, the developed Bayesian GLD regression model is applied using COVID-19 case density data previously employed to PL models using simple linear regression techniques in chapter 4. This dataset is ideal to run and test the model under different conditions (e.g., varying n), to illustrate the meaning of Bayesian parameter estimates and draw upon comparisons with earlier normality methods.

Expectation and dispersion parameter regression coefficients were obtained along with GLD location, scale and shape parameters. The expectation regression coefficients are retained, although improved by fitting them using the Bayesian GLD regression model developed in chapter 5. Previous normality methods are fundamentally and systematically biased since the underlying assumptions of normal linear regression are violated throughout the pandemic. The fitted parameter differences are highlighted showing the corrections made by the improved Bayesian GLD regression model. An additional asset given by the Bayesian GLD regression model is the dispersion regression parameter estimates accounting for heteroscedasticity in the data, a feature of the model that varies significantly in magnitude in the extended timeseries. The comparison of the dispersion exponent between the normal regression model with varying variance and the GLD regression model was striking. The normal regression model exhibited almost constant homoscedasticity: a data characteristic that has been shown to be untrue. The Bayesian GLD regression model, however, exhibited complex heteroscedasticity throughout the pandemic. Generally, heteroscedasticity affects high values of x when $\beta_3' > 0$ whilst heteroscedasticity affects low values of x when $\beta_3' < 0$. This additional information provided further information about the distribution of residuals (regions) around the PL. Timeframes when $\beta_3' > 0$ tended to homogenize low density regions whilst causing heterogeneity in highly dense regions and vice versa when $\beta_3' < 0$.

In addition, the Bayesian GLD regression model developed in chapter 5 also provided location, scale and shape parameters accounting for a wide range of complex shapes. This was an important advantage over previous normality methods. In standard linear regression it is assumed that residuals are normally distributed although, in chapter 4, it was shown to be untrue for large periods of the pandemic. Instead, residuals exhibited both positive and negative skew indicative of propagation. Positive skew featured a long tail of 'hot spots' and 'super-spreading' events whilst a negative skew featured a long tail of 'cold spots' and 'super-isolators'. Specifically, negative skew was indicated when $\alpha > 1$, no skew was indicated when $\alpha = 1$ and positive skew was indicated when $\alpha > 1$. This behavior was observed in chapter 4 (section 4.4.6) after applying normality methods. The advantage shown in this chapter is that the skewed behavior is developed within the Bayesian GLD regression framework itself, thus, correcting for any previous model bias produced by applying normality methods.

The Bayesian GLD regression model parameters converged well over the 759 daily data, with some day-to-day discontinuity on days when n was small corresponding to the number of regions reporting cases. Overwhelmingly, using COVID-19 data when n is at either full or near full capacity (i.e., the number of regions reporting at least one case) it was found that the developed Bayesian GLD regression model is the preferred model. The fitted Bayesian GLD regression model, when possible, compared to previous normality methods gives systematically different parameter estimates. The assumptions of a normal linear regression model are often violated, thus, producing model bias when applying normal regression methods. All parameters have been shown to systematically improve when applying the Bayesian GLD regression model.

The Bayesian GLD regression model is a significant improvement to previous normality methods. This is the first to apply PL models and data from the COVID-19 pandemic. The heteroscedasticity and skewed considerations are significant improvements to previous normality methods. It provides additional information and corrects the model bias produced when assuming a

homoscedastic and normal system. There are no heteroscedastic GLD regression models in the statistics literature capable of modelling negatively skew distribution. Thus, using the developed framework, it is the first study to apply it to the COVID-19 pandemic scaling.

6.7.1 Further Limitations

After running and testing the GLD regression model developed in chapter 5 some additional limitations are recognised. It was shown that model parameters β_0, β_1, θ, σ and α converge very well whilst dispersion regression model parameters β_2' and β_3' converge to a much lesser extent. Considering the extensively large dataset, the model parameters overall converge well, but possible improvements on the convergence rate for β_2' and β_3' could improve the overall fitting of the GLD regression model. Currently, non-informative priors are applied, and future applications of the model developed in chapter 5, may consider informative prior. In a timeseries dataset, like the COVID-19 pandemic, this might correspond to incorporating previous day information into the prior, but priors for other datasets, not in the form of timeseries needs further consideration.

In the application of COVID-19, case and population densities (continuous case) data was used in the GLD regression model developed in chapter 5. Conventional scaling studies, use discrete values such as raw indicators and population data. Thus, future work must consider modelling expectation and dispersion developed in chapter 5 for the case when X is discrete and consider the implications of uncertainty and model parameters.

Chapter 7: Overall Conclusions

7.1 Overall Summary

The scaling literature has focused on urbanised systems whilst neglecting studying other types of environments such as low population dense regions. The study of cities is motivated by the economics [219] and locational [220,221] theories highlighting the importance of urban spaces. These relate from spatial influences to economic factors and ideas relating to cost and transport [18]. However, it is thought that the scaling behaviour of urban regions are self-similar underpinning the urban scaling hypothesis. Recent studies of crime and property using rural-urban parliamentary constituencies to define regional boundaries challenge this hypothesis showing that many of the studied metrics exhibited a segmented relationship with a change point positioned at a similar population density. The methodology applied in these studies are available in chapter 2. The inherited methodology in conventional scaling use normal linear regression and assumes a homoscedastic and normally distributed residuals. These have been shown to be fundamentally biased and it is often that these assumptions are violated.

Chapter 3 extends this study by including non-transmissible mortality and age metrics. The population density framework was applied to this set of metrics. These were well approximated by PLs (chapter 2) and were used to investigate the scaling relationships of 67 metrics. The residuals obtained from the preferred single or segmented PL models were used to better understand regional variation and inter-correlation in both: between indicators and between regions. It is shown that

almost all mortality declined in urban regions which is achieved by the ability that urban regions are able to attract younger people (25-39) and repel older people (45+). Most critical population densities obtained from the diverse range of metrics (crime, property, mortality and age) forms a bimodal distribution with a median of 27 p/h. After fitting the preferred model, residual analysis using correlation and similarity heatmaps along with hierarchical clustering, revealed three clusters: crime and young people; mortality and older people; property and middle-aged people. The most striking result is the correlation between crime and property clusters. The inter-correlation between these clusters revealed that a subset of deaths, linked to life choices, are anti-correlated with property transaction value. This protective effect of the property transaction value above the scaling expectation results in below expectation in some preventable deaths (e.g., suicide) and vice versa. Transposing the indicator residuals provided us with the regional residuals which led us to the analysis of regional variation. Analysis of the regional residuals separated England and Wales into two main clusters mostly separating northern and southern regions exhibited almost universal anti-correlation. If crime, mortality, property transaction values are above the scaling PL then they are below in the other. Further exploration of regional residuals, using self-organising maps provided a more definite picture of the variation presented in England and Wales and these were projected onto a geomap. In this analysis England and Wales clustered into four main groups. The green cluster (Fig 3.14) mostly located on the coast is characterised by high mortality and an older population. The blue cluster, mostly northern regions are mainly characterised by a deflated property value transaction. The red cluster have an inflated property value transaction, low mortality and younger population. Lastly, the yellow cluster that only consists of the City of London and St. Edmundsbury are characterised with high crime. Overall, this study shows that *per-capita* models are fundamentally biased and that policy makers need to understand that these models are skewed toward highly dense regions. There is a need for better developed statistical models when considering health care resources and allocation to improve the general health care system.

The study of property, crime, mortality and age was cumulative data over a relatively wide timeframe. COVID-19 (chapter 4) was a unique opportunity to study the scaling of a transmissible

disease at daily time points. This is the first study to analyse daily scaling at this granularity. The scaling exponents are thought to be slow changing or constant, but the analysis of COVID-19 cases showed that this was not true. They were extremely structured and complex. This was also observed in other statistical measures such as variance and skew. These statistical behaviours were all sensitive to the dynamics of propagation, government intervention, changing public behaviour and the different variants of the disease propagating at the time. Scaling exponents reversed six times where preference of spread in rural-urban regions switched. Regional heterogeneity timeframes coincided with restrictions while released restrictions tended to homogenise and reduce the distance of individual regions to the scaling law. An exceptional timeframe occurred during the delta-omicron era where the trajectory of variance significantly reduced and persisted low for the remaining studied period causing regions to homogenise. Skew also changed many times from being strongly negative to being strongly positive. The periods of positive skew characterise a long tail of 'super-spreading' regions and 'hot spots' events coinciding with timeframes with fewer restrictions. Whilst a negative skew characterises a long tail of 'super-isolating' regions and 'cold-spots' events which is a feature of propagation under-appreciated in the modelling of COVID-19. This is an extremely important consideration of propagation since this feature dominated for 12 months out the possible 25-month studied timeframe. Conversely, scaling exponents for COVID-19 mortality were almost constant for the entire studied period. It showed sub-linear scaling indicating that rural regions were most effected by mortality throughout the pandemic. This behaviour was consistent with the previous mortality study (chapter 3) where it was documented that rural regions have a higher general mortality and more elderly population. From a policy point of view these observations and patterns are important. They provide insight into the expected effect following health policies and changing public behaviour allowing for better preparation and monitoring for future interventions.

During the study of COVID-19, variance and skew were extremely variable and residuals were often not normally distributed and better characterised by a GLD. These observations led to the development of the GLM where the response Y, is assumed to be a GLD.

The corresponding developed Bayesian GLD regression model, available in chapter 5, retains mean regression coefficients and gains additional dispersion regression coefficients along with GLD location, scale and shape parameters. The complexity of the variance and residual distributions observed in chapter 4 are better fitted with the additional heteroscedasticity and distribution variability considered by the GLD regression model. The normal regression model assumes a normal, linear and homoscedastic system, although it has been shown that these assumptions are often violated. The additional parameters along with considerations of skewed distributions and heteroscedasticity in the GLD regression model have shown to systematically improve the fitting of model parameters and consequently reduce model bias.

In chapter 6, the key features of the model were tested and demonstrated using PLs and data from the COVID-19 pandemic. The data was applied using the developed Bayesian GLD regression model proposed in chapter 5. Within this framework a complicating, non-analytically tractable posterior distribution was obtained and therefore MCMC methods were applied. Specifically, the M-H algorithm was applied over 20,000 iterations and disregarded the first 10,000 iterations as 'burn-in'. Expectation and dispersion regression coefficient are obtained along with GLD location, scale and shape parameters. After applying Gelman-Rubin convergence diagnostics it was found that model parameters generally converge well over the course of 759 days. The additional dispersion regression coefficients provide information about the heteroscedasticity in the data. A positive β_3' (i.e., $\beta_3' > 0$) indicates heteroscedasticity affecting residuals with higher values of x whilst negative β_3' (i.e., $\beta_3' < 0$) indicates heteroscedasticity affecting residuals with lower values of x. The magnitude of β_3' indicates more severe levels of heteroscedasticity. In the context of COVID-19, using density data and PL models, periods where $\beta_3' > 0$ tended to homogenise rural regions whilst urban regions deviated from the PL mostly corresponding to positively skewed residuals. Whereas periods $\beta_2' < 0$ had the opposite effect such that urban regions homogenised whilst rural regions deviated from the PL corresponding to negatively skewed residuals. In addition, the shape parameter obtained from the GLD provided information about the types of skewness (positive or negative). Positive skew occurred

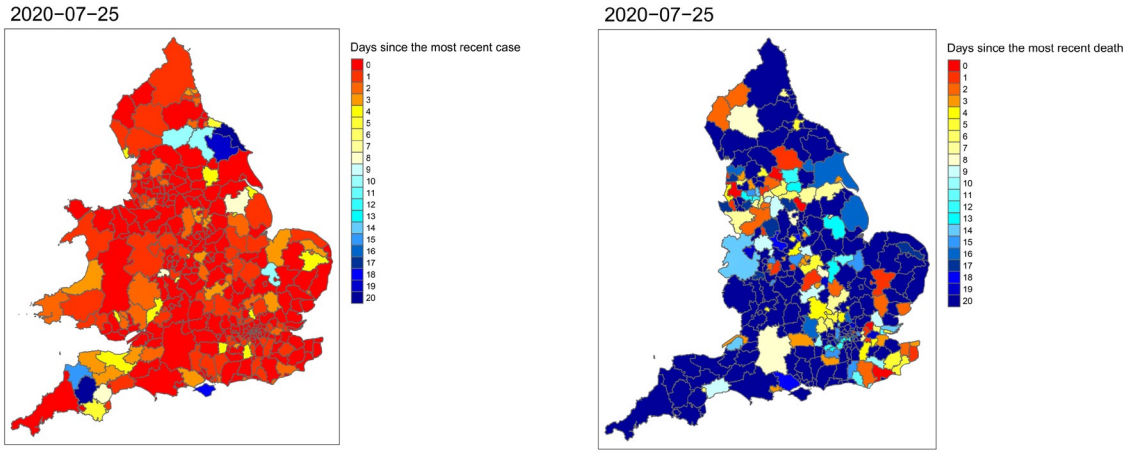
when $\alpha > 1$, negative skew occurred when $\alpha < 1$ and a symmetrical distribution was established when $\alpha = 1$. Conversely, periods of negative skew characterise a long tail of 'super-isolators' and 'cold spots'.

7.2 Other Contributions

The methods and discussions in this study has also benefited other work. At the earlier stages of the COVID-19 pandemic, initial discussions, started with the notion that population density is a proxy for interaction and meeting size. For example, regions with a highly dense population would have more opportunities for circulation of the infectious disease since there are higher proportions of people and a greater number of infectious interactions between them. Though during this research, it is found that the hypothesis of population density as a proxy for interaction is not entirely true and that propagation is far more complex. Here it is noted that the discussions encouraged by the topics in this thesis led to the work available in [222]. However, the analysis was primarily conducted by others, thus an in-depth presentation of this work is not available within this thesis. This study was first published during the earlier stages of the pandemic and represents the first 170 days. This would now benefit from being updated to investigate whether the initial findings still hold or if they are fundamentally different as the pandemic evolved. Since the publication of this thesis, there have been different circulating variants, numerous government interventions and changes to public behaviour.

Other contributions include making the daily COVID-19 data sets (cases and deaths) and corresponding code easily accessible on an open public website (https://github.com/jacksutton13/regional_COVID_geoplots). The idea was to allow the public to quickly obtain this daily data and generate maps of England and Wales (Fig 7.1). The data on the site and corresponding code views the data from a different perspective seen in chapters 4 and 5. Instead discrete COVID-19 cases and deaths counts are retained and maps of England and Wales are computed observing the number of days since the most recent case (Fig. 6.1(a)) or death (Fig. 6.1(b)). If a region is blue, then there has been at least 9 days since the last case or death caused by COVID-19. The darker shade of blue indicates a greater number of days since the last case or death. The

motivation of this site was to provide people quick access to COVID-19 data. The idea of this presentation was to help the public view the data in a comprehensible format and view the most up-to-date data to make 'safer' and 'better' data informed decisions.



(a) (b)

Fig 7.1. Maps of England and Wales exhibiting the number of days since the last COVID-19 (a) case and (b) death. The red-blue colour spectrum indicates the number of days since the last case or death. Red indicates a case or death close to the specified day whilst blue indicates a case or death far away from the specified date.

7.3 Final Remarks

Currently, this study considers a wide range of indicators including crime, property, mortality, and age in chapter 3 along with COVID-19 in chapter 4. It would be interesting to extend the current datasets to include more health-related metrics and to explore other disciplines (e.g., climate change, GDP etc). In the initial period of study, normal PL models were applied whilst more recently it was found that these methods lead to model bias and sometimes violate the statistical assumptions. Thus, any future work would benefit from the developed and improved GLD regression model proposed in chapter 5. Having said that, the proposed framework in chapter 5 and the algorithms (e.g., M-H algorithm) within it are complicated, thus the work in chapter 5 would benefit with an established and universal R package. This would make the proposed model in chapter 5 more easily assessable for any future work in scaling and more general work using linear regression.

As aforementioned, it was found in chapter 3 that there is a consistent change point at a similar critical population density which almost forms a unimodal distribution. The reasons for a critical population density are unclear and the processes that cause rural-urban regions to be fundamentally different needs further consideration.

A limitation of the methodologies imposed in this study is the dealing with missing data. In all studies within this thesis, regions with no reported data are considered missing due to the logarithmic transformation of the data not allowing for 0 values. This is inherited in the urban scaling field and there are currently few alternatives that would deal with missing data better. Incrementing the data has been suggested [156] but this is not a satisfying replacement to existing methods. This is because regions reporting 0 values are valid, but analysis ignoring this data are missing this important information. This will also impact the reporting of scaling exponents using least squared methods especially in case where there are zero-inflated values. A recent contribution [156] examines some possible solutions dealing with 0 values in urban scaling. This includes the weighted ordinary least squared method, where regions are positioned into bins with similar attributes in a way that 0 values do not exist. Another method proposed in this study is the Hurdle and Poisson model where 0 values are viewed as a binary outcome (presence/absence) rather than on a continuous scale and dealt with separately. There is also possible scope, that methods in ref [156] if found to be successful, to incorporate them into the proposed Bayesian GLD model. Furthermore, there are also issues related to what determines the boundaries of a city [223] as well as on going attempts to understand how the productivity of cities might relate to networks and related mathematics [224,225].

References

1. Auerbach F. Das Gesetz der Bev¨olkerungskonzentration. Petermanns Geogr Mitt. 1913;59.
2. Lotka AJ. Elements of Physical Biology. Baltimore: Waverly Press; 1925.
3. Rybski D. Auerbach's legacy. Environ Plan A. 2013;45.
4. Zipf GK. National unity and disunity: The nation as a bio-social organism. Bloomington: Principia Press; 1941.
5. Nordbeck S. Urban Allometric Growth. Geogr Ann Ser B, Hum Geogr. 1971;53: 54–67. doi:10.1080/04353684.1971.11879355
6. Beckmann MJ. City Hierarchies and the Distribution of City Size. Econ Dev Cult Change. 1958;6: 243–248. doi:10.1086/449769
7. Smith HF. An empirical law describing heterogeneity in the yields of agricultural crops. J Agric Sci. 1938;28: 1–23. doi:10.1017/S0021859600050516
8. Taylor L. Aggregation, variance and the mean. Nature. 1961;189: 732–735.
9. Perry JN. Taylor's Power Law for Dependence of Variance on Mean in Animal Populations. Appl Stat. 1981;30: 254. doi:10.2307/2346349
10. Rybski D, Buldyrev S V., Havlin S, Liljeros F, Makse HA. Scaling laws of human interaction activity. Proc Natl Acad Sci. 2009;106: 12640–12645. doi:10.1073/pnas.0902667106
11. Gaffeo E, Di Guilmi C, Gallegati M, Russo A. On the mean/variance relationship of the firm size distribution: Evidence and some theory. Ecol Complex. 2012;11: 109–117. doi:10.1016/j.ecocom.2012.05.001
12. Eisler Z, Bartos I, Kertész J. Fluctuation scaling in complex systems: Taylor's law and beyond. Adv Phys. 2008;57: 89–142. doi:10.1080/00018730801893043

13. Petri G, Expert P, Jensen HJ, Polak JW. Entangled communities and spatial synchronization lead to criticality in urban traffic. Sci Rep. 2013;3: 1–8. doi:10.1038/srep01798

14. de Menezes MA, Barabási A-L. Fluctuations in Network Dynamics. Phys Rev Lett. 2004;92: 028701. doi:10.1103/PhysRevLett.92.028701

15. Duranton G, Puga D. Micro-foundations of urban agglomeration economies. In book: Handbook of Regional and Urban Economics. In: Henderson J V., Thisse JF, editors. Elsevier. 2004. pp. 2063–2117. doi:10.1016/S1574-0080(04)80005-1

16. Glaeser E. Agglomeration Economies. University of Chicago Press; 2010.

17. Arcaute E, Ramasco JJ. Recent advances in urban system science: Models and data. PLoS One. 2022;17: 1–16. doi:10.1371/journal.pone.0272863

18. Rybski D, Arcaute E, Batty M. Urban scaling laws. Environ Plan B Urban Anal City Sci. 2019;46: 1605–1610. doi:10.1177/2399808319886125

19. Marshall A. Principles of Economics. Macmillan and Co; 1890. Available: http://archive.org/%0Adetails/principlesecono00marsgoog

20. Pumain D. Pour Une Théorie Évolutive Des Villes. JSTOR. 1997;26: 119–34. Available: http://www.jstor.org/stable/44381391

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

22. Bettencourt LMA, Lobo J, Strumsky D, West GB. Urban Scaling and Its Deviations: Revealing the Structure of Wealth, Innovation and Crime across Cities. Añel JA, editor. PLoS One. 2010;5: e13541. doi:10.1371/journal.pone.0013541

23. Van Raan AFJ, Van Der Meulen G, Goedhart W. Urban scaling of cities in the Netherlands. PLoS One. 2016;11: 1–18. doi:10.1371/journal.pone.0146775

24. Alves LGA, Ribeiro H V, Mendes RS. Scaling laws in the dynamics of crime growth rate. Physica A. 2013;392: 2672–2679. doi:10.1016/j.physa.2013.02.002

25. Gomez-Lievano A, Youn H, Bettencourt LMA. The Statistics of Urban Scaling and Their Connection to Zipf's Law. Lambiotte R, editor. PLoS One. 2012;7: e40393.

doi:10.1371/journal.pone.0040393

26. Hanley QS, Lewis D, Ribeiro H V. Rural to Urban Population Density Scaling of Crime and Property Transactions in English and Welsh Parliamentary Constituencies. Perc M, editor. PLoS One. 2016;11: e0149546. doi:10.1371/journal.pone.0149546

27. 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. Gomez-Lievano A, editor. PLoS One. 2018;13: e0192931. doi:10.1371/journal.pone.0192931

28. Nordbeck S. Urban Allometric Growth. Geogr Ann Ser B. 1971;53: 54–67.

29. van Raan AFJ, van der Meulen G, Goedhart W. Urban Scaling of Cities in the Netherlands. Rozenblat C, editor. PLoS One. 2016;11: e0146775. doi:10.1371/journal.pone.0146775

30. United Nations. World Urbanisation Prospects: The 2014 Revision. New York: Department of Economic and Social Affairs, Population Division; 2015.

31. Bettencourt LMA, Lobo J, Youn H. The hypothesis of urban scaling: formalization, implications and challenges. 2013. doi:10.1.1.369.8990

32. Ribeiro H V, Rybski D, Kropp JP. Effects of changing population or density on urban carbon dioxide emissions. Nat Commun. 2019;10: 1–9. doi:10.1038/s41467-019-11184-y

33. Faraway JJ. Linear Models with R. Chatfield C, Tanner M, Zidek J, editors. London: Chapman and Hall/CRC; 2005.

34. Belsley DA, Kuh E, Welsch RE. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: Wiley; 1980.

35. Anderson RL, Bancroft, Theodore A. Statistical Theory in Research. Hill M, editor. New York: Wiley; 1952.

36. Chatfield C. Model Uncertainty, Data Mining and Statistical Inference. J R Stat Soc Ser A Stat Soc. 1995;158: 419. doi:10.2307/2983440

37. Leitão JC, Miotto JM, Gerlach M, Altmann EG. Is this scaling nonlinear? R Soc Open Sci. 2016;3: 150649. doi:10.1098/rsos.150649

38. 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.

doi:10.1038/s41598-020-74015-x

39. Sutton J, Shahtahmassebi G, Ribeiro H V, Hanley QS. Population density and spreading of COVID-19 in England and Wales. PLoS One. 2022;17: 1–19. doi:10.1371/journal.pone.0261725

40. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian Data Analysis. Third. London: Chapman and Hall/CRC; 2014.

41. Ribeiro H V, Sunahara AS, Sutton J, Perc M, Hanley QS. City size and the spreading of COVID-19 in Brazil. PLoS One. 2020;15: 1–12. doi:10.1371/journal.pone.0239699

42. Alves LGA, Mendes RS, Lenzi EK, Ribeiro H V. Scale-Adjusted Metrics for Predicting the Evolution of Urban Indicators and Quantifying the Performance of Cities. Rozenblat C, editor. PLoS One. 2015;10: e0134862. doi:10.1371/journal.pone.0134862

43. Pumain D. Scaling Laws and Urban Systems. SFI Working Paper 2004-02-002. 2004. Available: https://www.santafe.edu/research/results/working-papers/scaling-laws-and-urban-systems

44. Irigoien X, Huisman J, Harris RP. Global biodiversity patterns of marine phytoplankton and zooplankton. Nature. 2004;429: 863–867. doi:10.1038/nature02593

45. UN/DESA. The World's Cities in 2018. World Urban Prospect 2018 Revis. 2018; 34.

46. Demographia. Demographia World Urban Areas. Demographia. 2015;202106: 132.

47. Biiby P, Brindley P. Urban and Rural Area Definitions for Policy Purposes in England and Wales: Methodology. Office of National Statistics; 2013.

48. Anonymous. 2011 Rural-Urban Classification for Output Areas in England. Defra Rural Stat. 2015.

49. NHS England Analytical Services (Finance). Technical Guide to Allocation Formulae and Pace of Change. 2016.

50. England N. Fair Shares: A guide to NHS Allocations. 2018.

51. Technical NEAS (Finance). Guide to Allocation Formulae and Pace of Change. 2016.

52. Green AE. The North-South Divide in Great Britain: An Examination of the Evidence. Trans Inst Br Geogr. 1988;13: 179. doi:10.2307/622505

53. Hacking JM, Muller S, Buchan IE. Trends in mortality from 1965 to 2008 across the English north-south divide: comparative observational study. BMJ. 2011;342: d508–d508. doi:10.1136/bmj.d508

54. Keeble D, Bryson J. Small-Firm Creation and Growth, Regional Development and the North—South Divide in Britain. Environ Plan A Econ Sp. 1996;28: 909–934. doi:10.1068/a280909

55. Salvatore M, Pozzi F, Ataman E, Huddleston B, Bloise M. Mapping global urban and rural population distributions. 2005.

56. Swiecki-Sikora AL, Henry KA, Kepka D. HPV Vaccination Coverage Among US Teens Across the Rural-Urban Continuum. J Rural Heal. 2019;35: 506–517. doi:10.1111/jrh.12353

57. Li K, Chen Y, Wang M, Gong A. Spatial-temporal variations of surface urban heat island intensity induced by different definitions of rural extents in China. Sci Total Environ. 2019;669: 229–247. doi:10.1016/j.scitotenv.2019.03.100

58. Burnham KP, Efron B. The Jackknife, the Bootstrap and Other Resampling Plans. Biometrics. 1983;39: 816. doi:10.2307/2531123

59. Efron B. Nonparametric Standard Errors and Confidence Intervals. Can J Stat. 1981;9: 139–158.

60. Efron B. Nonparametric Estimates of Standard Errors: The Jackknife, the Boostrap and Other Methods. Biometrika. 1981;68: 589–599.

61. Efron B. Bootstrap Methods: Another Look at the Jackknife. Annals of Statistics. 1992. pp. 569–593. doi:10.1007/978-1-4612-4380-9_41

62. L. RM. Statistical Computing with R. CRC Press; 2007.

63. Bradley E, J. TR. An introduction to bootstrap. CRC Press; 1994.

64. Buckland ST, Davison AC, Hinkley D V. Bootstrap Methods and Their Application. Biometrics. 1998;54: 795. doi:10.2307/3109789

65. Davison AC, Hinkley D V. Bootstrap Methods and Their Applications. Cambridge: Cambridge University Press; 1997.

66. Stevens AA, Tappon SC, Garg A, Fair DA. Functional Brain Network Modularity Captures

Inter- and Intra-Individual Variation in Working Memory Capacity. Ward LM, editor. PLoS One. 2012;7: e30468. doi:10.1371/journal.pone.0030468

67. Barabási AL, Oltvai ZN. Network biology: Understanding the cell's functional organization. Nat Rev Genet. 2004;5: 101–113. doi:10.1038/nrg1272

68. Kaiser M. A tutorial in connectome analysis: Topological and spatial features of brain networks. Neuroimage. 2011;57: 892–907. doi:10.1016/j.neuroimage.2011.05.025

69. Sporns O. Neworks of the Brain. MIT Press; 2010.

70. Crofts JJ, Higham DJ. A weighted communicability measure applied to complex brain networks. J R Soc Interface. 2009;6: 411–414. doi:10.1098/rsif.2008.0484

71. Newman M. Networks. Oxford: Oxford University Press; 2010. doi:10.1093/acprof:oso/9780199206650.001.0001

72. Cherifi C, Cherifil H, Karsai M, Musolesi M. Complex Networks & Their Applications VI. Lyon; 2017.

73. Costa LDF, Rodrigues FA, Travieso G, Boas PRV. Characterization of complex networks: A survey of measurements. Adv Phys. 2007;56: 167–242. doi:10.1080/00018730601170527

74. Newman MEJ. Modularity and community structure in networks. Proc Natl Acad Sci U S A. 2006;103: 8577–82. doi:10.1073/pnas.0601602103

75. Barber MJ. Modularity and community detection in bipartite networks. Phys Rev E. 2007;76: 066102. doi:10.1103/PhysRevE.76.066102

76. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theory Exp. 2008;2008: P10008. doi:10.1088/1742-5468/2008/10/P10008

77. Mehrle D, Strosser A, Harkin A. Walk-modularity and community structure in networks. Netw Sci. 2015;3: 348–360. doi:10.1017/nws.2015.20

78. Girvan M, Newman MEJ. Community structure in social and biological networks. Proc Natl Acad Sci U S A. 2002;99: 7821–6. doi:10.1073/pnas.122653799

79. Kohonen T. Self-Organizing Maps. 3rd Editio. Kohonen T, Huang T, Schroeder M, editors. Berlin: Springer; 2001.

80. Kohonen T. The self-organizing map. Proc IEEE. 1990;78: 1464–1480. doi:10.1109/5.58325

81. Vesanto Juha, Himberg Johan, Alhoniemi Esa PJ. Self-Organizing Map in Matlab: Som Toolbox. 2000.

82. Oliver E, Vallés-Ṕerez I, Baños RM, Cebolla A, Botella C, Soria-Olivas E. Visual Data Mining With Self-organizing Maps for "Self-monitoring" Data Analysis. Sociol Methods Res. 2018;47: 492–506. doi:10.1177/0049124116661576

83. Cohen D. Precalculus: A Problems-Oriented Approach. 6th ed. Cengage Learning; 2004.

84. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc Ser B (Statistical Methodol. 2001;63: 411–423. doi:10.1111/1467-9868.00293

85. Benesty J, Chen J, Huang Y, Cohen I. Pearson Correlation Coefficient. 2009. pp. 1–4. doi:10.1007/978-3-642-00296-0_5

86. R. The R Project for Statistical Computing. [cited 23 Feb 2021]. Available: https://www.r-project.org/

87. Muggeo VMR. segmented: an R Package to Fit Regression Models with Broken-Line Relationships. R News. 2008;8: 20–25. Available: https://cran.r-project.org/doc/Rnews/

88. Muggeo VMR. Estimating regression models with unknown break-points. Stat Med. 2003;22: 3055–3071.

89. Muggeo VMR. Testing with a nuisance parameter present only under the alternative: a score-based approach with application to segmented modelling. Stat Comput Simul. 2016;86: 3059–3067.

90. Muggeo VMR. Interval estimation for the breakpoint in segmented regression: a smoothed score-based approach. Aust N Z J Stat. 2017;59: 311–322. Available: https://cran.r-project.org/doc/Rnews/

91. Meyer D, Buchta C. proxy: Distance and Similarity Measures. 2019. Available: https://cran.r-project.org/package=proxy

92. Canty A, Ripley BD. boot: Bootstrap R (S-Plus) Functions. 2019.

93. Wehrens R, Kruisselbrink J. Flexible Self-Organizing Maps in kohonen 3.0. J Stat Softw.

2018;87: 1–18. doi:10.18637/jss.v087.i07

94. Wehrens R, Buydens LMC. Self- and Super-organizing Maps in R : The kohonen Package. J Stat Softw. 2007;21: 1–19. doi:10.18637/jss.v021.i05

95. Kassambara A, Mundt F. factoextra: Extract and Visualize the Results of Multivariate Data Analyses. 2019. Available: https://cran.r-project.org/package=factoextra

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

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

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

99. Fox J, Weisberg S. An {R} Companion to Applied Regression. Third. Thousand Oaks {CA}: Sage; 2019. Available: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

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

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

102. Milligan GW, Cooper MC. An examination of procedures for determining the number of clusters in a data set. Psychometrika. 1985;50: 159–179. doi:10.1007/BF02294245

103. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. J Open Source Softw. 2019;4: 1686. doi:10.21105/joss.01686

104. Wilke CO. cowplot: Streamlined Plot Theme and Plot Annotations for “ggplot2.” 2019.

105. Revelle W. psych: Procedures for Psychological, Psychometric, and Personality Research. Evanston, Illinois; 2019. Available: https://cran.r-project.org/package=psych

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

107. Hijmans RJ. raster: Geographic Data Analysis and Modeling. 2020. Available: https://cran.r-

project.org/package=raster

108. Wickman H, François R, Henry L, Muller K. dlpyr: A Grammar of Data Manipulation. CRAN Repository. 2021. pp. 1–2. Available: https://cran.r-project.org/package=dplyr

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

110. Tennekes M. tmap : Thematic Maps in R. J Stat Softw. 2018;84: 1–39. doi:10.18637/jss.v084.i06

111. Cheng J, Karambelkar B, Xie Y. leaflet: Create Interactive Web Maps with the JavaScript "Leaflet" Library. 2019. Available: https://cran.r-project.org/package=leaflet

112. Appelhans T, Detsch F, Reudenbach C, Woellauer S. mapview: Interactive Viewing of Spatial Data in R. 2019. Available: https://cran.r-project.org/package=mapview

113. Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J. shiny: Web Application Framework for R. 2020. Available: https://cran.r-project.org/package=shiny

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

115. Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. Third international AAAI conference on weblogs and social media. 2009.

116. Caminha C, Furtado V, Pequeno THC, Ponte C, Melo HPM, Oliveira EA, et al. Human mobility in large cities as a proxy for crime. Amaral LAN, editor. PLoS One. 2017;12: e0171609. doi:10.1371/journal.pone.0171609

117. Grasmick H, Tittle C, Bursik R, Arneklev B. Testing the Core Empirical Implications of Gottfredson and Hirschi's General Theory of Crime. J Res Crime Delinq. 1993;30: 5–29. doi:10.1177/0022427893030001002

118. Pratt TC, Cullen FT. The Empirical Status of Gottfredson and Hirschi's General Theory of Crime: A Meta-Analysis. Criminology. 2000;38: 931–964. doi:10.1111/j.1745-9125.2000.tb00911.x

119. Wikström P-OH. Why crime happens: A situational action theory. Analytical Sociology.

Chichester, United Kingdom: John Wiley & Sons, Ltd; 2014. pp. 71–94. doi:10.1002/9781118762707.ch03

120. Wikström P-OH. Crime as alternative: Towards a cross-level situational action theory of crime causation. In: McCord J, editor. Beyond Empiricism: Institutions and Intentions in the Study of Crime. Transaction Publishers; 2004. pp. 1–38.

121. Kinney JB, Mann E, Winterdyk JA. Crime Prevention. Winterdyk JA, editor. Crime Prevention: International Perspectives, Issues, and Trends. 1 Edition. | Boca Raton, FL : CRC Press, 2017.: CRC Press; 2017. doi:10.1201/9781315314211

122. Bettencourt LMA. The Origins of Scaling in Cities. Science (80-). 2013;340: 1438–1441. doi:10.1126/science.1235823

123. Advisory Commitee Allocation Resource. Public health grant: exposition book for proposed formula for 2016-17 target allocations: technical guide: an engagement on behalf of the Advisory Committee on Resource Allocation (ACRA). 2015.

124. England N. Fair Shares: A guide to NHS Allocations. 2018.

125. Lee D, Cho YS, Goh K-I, Lee D-S, Kahng B. Recent advances of percolation theory in complex networks. J Korean Phys Soc. 2018;73: 152–164. doi:10.3938/jkps.73.152

126. Arcaute E, Molinero C, Hatna E, Murcio R, Vargas-Ruiz C, Masucci AP, et al. Cities and regions in Britain through hierarchical percolation. R Soc Open Sci. 2016;3. doi:10.1098/rsos.150691

127. Wynder E, Covey L, Mabuchi K, Mushininski M. Environmental Factors in Cancer of the Larynx, A second Look. Cancer. 1976;38: 1591–1601.

128. South AP, den Breems NY, Richa T, Nwagu U, Zhan T, Poojan S, et al. Mutation signature analysis identifies increased mutation caused by tobacco smoke associated DNA adducts in larynx squamous cell carcinoma compared with oral cavity and oropharynx. Sci Rep. 2019;9: 1–9. doi:10.1038/s41598-019-55352-y

129. Amiri S, Behnezhad S. Cancer Diagnosis and Suicide Mortality: A Systematic Review and Meta-Analysis. Arch Suicide Res. 2019;0: 1–19. doi:10.1080/13811118.2019.1596182

130. Barnard-Kelly KD, Naranjo D, Majidi S, Akturk HK, Breton M, Courtet P, et al. Suicide and

Self-inflicted Injury in Diabetes: A Balancing Act. J Diabetes Sci Technol. 2019. doi:10.1177/1932296819891136

131. Alattas M, Ross CS, Henehan ER, Naimi TS. Alcohol policies and alcohol-attributable cancer mortality in U.S. States. Chem Biol Interact. 2020;315: 108885. doi:10.1016/j.cbi.2019.108885

132. Alves LGA, Ribeiro H V, Mendes RS. Scaling laws in the dynamics of crime growth rate. Phys A Stat Mech its Appl. 2013;392: 2672–2679. doi:10.1016/j.physa.2013.02.002

133. Alves LGA, Ribeiro H V, Lenzi EK, Mendes RS. Distance to the Scaling Law: A Useful Approach for Unveiling Relationships between Crime and Urban Metrics. Perc M, editor. PLoS One. 2013;8: e69580. doi:10.1371/journal.pone.0069580

134. Arcaute E, Hatna E. Scaling Laws: Insights and Limitations. 2020. pp. 45–66. doi:10.1007/978-3-030-36656-8_4

135. Leitão JC, Miotto JM, Gerlach M, Altmann EG. Python code to analyse scaling laws. 2016 [cited 16 Mar 2023]. doi:10.5281/ZENODO.49367

136. Morawska L, Cao J. Airborne transmission of SARS-CoV-2: The world should face the reality. Environ Int. 2020;139: 105730. doi:10.1016/j.envint.2020.105730

137. Anderson EL, Turnham P, Griffin JR, Clarke CC. Consideration of the Aerosol Transmission for COVID-19 and Public Health. Risk Anal. 2020;40: 902–907. doi:10.1111/risa.13500

138. Prather KA, Wang CC, Schooley RT. Reducing transmission of SARS-CoV-2. Science (80-). 2020;368: 1422–1424. doi:10.1126/science.abc6197

139. 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. doi:10.1080/02786826.2020.1749229

140. 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. doi:10.1098/rsif.2013.0789

141. 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. doi:10.1007/s42770-020-00331-z

142. 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. doi:10.1098/rsos.201382

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

144. 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. doi:10.1016/j.idm.2020.03.002

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

146. 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. doi:10.1111/jors.12510

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

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

149. Kendall M. A new measure of rank correlation. Biometrika. 1938;30: 81–93. doi:10.1093/biomet/30.1-2.81

150. Thirumalai C, Kumar D. K, Sidhardha K, Kumar Reddy DV. An assessment framework of SIAM/ARAI fuel efficiency using semi-supervised and similarity methods. 2017 International Conference on Trends in Electronics and Informatics (ICEI). IEEE; 2017. pp. 957–961. doi:10.1109/ICOEI.2017.8300849

151. Mohammad A, Kibria BM, Shakil M. Normal distribution. In Normal and Student st Distributions and Their Applications. Paris: Atlantis Press; 2014.

152. Shao Q. Maximum Likelihood Estimation for Generalised Logistic Distribution. Commun Stat - Theory Methods. 2002;31: 1687–1700. doi:10.1081/STA-120014908

153. Johnson LN, Kotz S, Balakrishnan N. Continuous Univariate Distributions Volume 2. Second. New York: Wiley and Sons; 1995.

154. Dubey SD. A new derivation of the logistic distribution. Nav Res Logist Q. 1969;16: 37–40.

155. Windberger T, Zeileis A. Structural Breaks in Inflation Dynamics within the European Monetary Union. East Europ Econ. 2014;52: 66–88. doi:10.2753/EEE0012-8775520304

156. Finance O, Cottineau C. Are the absent always wrong? Dealing with zero values in urban scaling. Environ Plan B Urban Anal City Sci. 2019;46: 1663–1677. doi:10.1177/2399808318785634

157. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2003. Available: https://www.r-project.org/

158. Pebesma E. Simple features for R: Standardized support for spatial vector data. R J. 2018;10: 439–446. doi:10.32614/rj-2018-009

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

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

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

162. Dragulescu AA. xlsx: Read, write, format Excel 2007 and Excel 97/2000/XP/2003 files. R package version 0.5.7. https://CRANR-project.org/package=xlsx. 2014.

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

164. Wickham H. The Split-Apply-Combine Strategy for Data Analysis. J Stat Softw. 2011;40: 1–29. Available: http://www.jstatsoft.org/v40/i01/

165. Bivand R, Keitt T, Rowlingson B. rgdal: Bindings for the “Geospatial” Data Abstraction Library. 2021. Available: https://cran.r-project.org/package=rgdal

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

167. Grolemund G, Wickham H. Dates and Times Made Easy with Lubridate. J Stat Softw. 2011;40: 1–25. Available: https://www.jstatsoft.org/v40/i03/

168. Delignette-Muller ML, Dutang C. An {R} Package for Fitting Distributions. J Stat Softw.

2015;64: 1–34. Available: https://www.jstatsoft.org/v64/i04/

169. Wuertz D, Setz T, Chalabi Y, Boudt C, Chausse P, Miklovac M. Rmetrics - Autoregressive Conditional Heteroskedastic Modelling. 2020. Available: https://cran.r-project.org/package=fGarch

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

171. Komsta L, Novomestky F. moments: Moments, Cumulants, Skewness, Kurtosis and Related Tests. R Packag Version 013. 2012; 1–15.

172. Meyer D, Buchta C. proxy: Distance and Similarity Measures. 2010. Available: http://cran.r-project.org/package=proxy

173. Revelle W. Package "psych" - Procedures for Psychological, Psychometric and Personality Research. R Packag. 2015; 1–358. Available: http://personality-project.org/r/psych-manual.pdf

174. Lemon, J; Bolker, B; Oom, S; Klein E. plotrix - Various Plotting Functions. 2016. Available: https://cran.r-project.org/package=plotrix

175. 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. doi:10.1073/pnas.2012326117

176. 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: 013032. doi:10.1103/PhysRevResearch.3.013032

177. Duong D. Alpha, Beta, Delta, Gamma: What's important to know about SARS-CoV-2 variants of concern? CMAJ. 2021;193: E1059–E1060. doi:10.1503/cmaj.1095949

178. Shiehzadegan S, Alaghemand N, Fox M, Venketaraman V. Analysis of the Delta Variant B.1.617.2 COVID-19. Clin Pract. 2021;11: 778–784. doi:10.3390/clinpract11040093

179. Burki TK. Omicron variant and booster COVID-19 vaccines. Lancet Respir Med. 2022;10: e17. doi:10.1016/S2213-2600(21)00559-2

180. Duong B V., Larpruenrudee P, Fang T, Hossain SI, Saha SC, Gu Y, et al. Is the SARS CoV-2 Omicron Variant Deadlier and More Transmissible Than Delta Variant? Int J Environ Res

Public Health. 2022;19. doi:10.3390/ijerph19084586

181. 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. doi:10.1016/j.epidem.2020.100430

182. 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. doi:10.1073/pnas.2011802117

183. Fukui M, Furukawa C. Power Laws in Superspreading Events. medRxiv. 2020; 1–41.

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

185. Yang Y, Sun L, Guo C. Aero-material consumption prediction based on linear regression model. Procedia Comput Sci. 2018;131: 825–831. doi:10.1016/j.procs.2018.04.271

186. Dai Q, Liu X-Q, Wang T-M, Vukicevic D. Linear regression model of DNA sequences and its application. J Comput Chem. 2007;28: 1434–1445. doi:10.1002/jcc.20556

187. Fumo N, Rafe Biswas MA. Regression analysis for prediction of residential energy consumption. Renew Sustain Energy Rev. 2015;47: 332–343. doi:10.1016/j.rser.2015.03.035

188. Freedman DA. Statistical Models: Theory and Practice. Cambridge University Press; 2009.

189. Nelder AJA, Wedderburn RWM, Journal S, Statistical R, Series S. Generalized Linear Models. 1972;135: 370–384.

190. Barndorff-Nielsen O. Information and exponential families: in statistical theory. John Wiley & Sons; 2014.

191. Dobson AJ, Barnett AG. An introduction to Generalized Linear Models. Third. Carlin BP, Faraway JJ, Tanner M, Zidek J, editors. London: CRC Press; 2008.

192. Silva L da. Modelagem geométrica do par roda-trilho com descrição por meio de NURBS. Universidade de São Paulo. 2021. doi:10.11606/D.3.2021.tde-25102021-145823

193. Aljarrah MA, Famoye F, Lee C. Generalized logistic distribution and its regression model. J

Stat Distrib Appl. 2020;7: 7. doi:10.1186/s40488-020-00107-8

194. Asgharzadeh A, Esmaeili L, Nadarajah S, Shih S. A generalised skew logistic distibution. REVSTAT. 2013;11: 317–338.

195. Nadarajah S. The skew logistic distribution. AStA Adv Stat Anal. 2009;93: 187–203.

196. Jamalizadeh A, Behboodian J, Balakrishnan N. A two-parameter generalised skew-normal distribution. Stat Probab Lett. 78: 1722–1726.

197. Zelterman D. Parameter estimation in the generalized logistic distribution. Comput Stat Data Anal. 1987;5: 177–184. doi:10.1016/0167-9473(87)90013-2

198. Zelterman D. Computational statistics and data analysis. Comput Stat Data Anal. 1989;7: 67–69.

199. Dubey SD. A new deviation of the logistic distribution. Nav Res Logist Q. 16: 37–40.

200. Stukel TA. Generalized Logistic Models. J Am Stat Assoc. 1988;83: 426–431.

201. Asgharzadeh A, Esmaeili L, Nadarajah S. Balakrishnan skew logistic distribution. Commun Stat - Theory Methods. 2016;45: 444–464. doi:10.1080/03610926.2013.823205

202. Asgharzadeh A, Esmaeili L, Nadarajah S, Shih S. A generalized skew logistric distribution. REVSTAT. 2013;11: 317–338.

203. Faraway JJ. Linear Models with R. Chatfiled C, Tanner M, Zidek J, editors. London: Chapman and Hall/CRC; 2005.

204. Ogden H, Overstall A. Lecture: Statistical Modelling. In: University of Warwick. 2019.

205. Faraway JJ. Extending the Linear model with r: Generalized linear, mixed effects and nonparametric regression models. Second. London: Chapman and Hall/CRC; 2016.

206. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH. Equation of State Calculations by Fast Computing Machines. J Chem Phys. 1953;21: 1088–1092.

207. Hastings AWK. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika. 1970;57: 97–109.

208. Peskun P. Optimum Monte-Carlo Sampling Using Markov Chains. Oxford Jounals. 1973;60: 607–612.

209. Garmerman D, Lopes HF. Markov Chain Monte Carlo. Second. London: Chapman and

Hall/CRC; 2006.

210. Gelman A, Rubin DB. Inference from Iterative Simulation Using Multiple Sequences. Stat Sci. 1992;7: 457–511.

211. Brooks SP, Gelman A. General Methods for Monitoring Convergence of Iterative Simulations. J Comput Graph Stat. 1998;7: 434–455. doi:10.1080/10618600.1998.10474787

212. Shahtahmassebi G, Moyeed R. An application of the generalized Poisson difference distribution to the Bayesian modelling of football scores. Stat Neerl. 2016;70: 260–273. doi:10.1111/stan.12087

213. Tanner, A M. Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions. Third. New York: Springer; 1998.

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

215. Chan C, Chan GCH, Leeper TJ, Becker J. rio: A Swiss-army knife for data file I/O. 2021.

216. Bengtsson H. matrixStats: Functions that Apply to Rows and Columns of Matrices (and to Vectors). 2021. Available: https://cran.r-project.org/package=matrixStats

217. Wuertz D, Chalabi Y, Miklovic M, Boudt C, Chausse P. FGarch: Rmetrics -autoregressive conditional heteroskedastic modelling. Zu Rich Diethelm Wuertz R Metrics Found. 2009.

218. Plummer M, Best N, Cowles K, Vines K. CODA: Convergence Diagnosis and Output Analysis for MCMC. R News. 2006;6: 7–11. Available: https://journal.r-project.org/archive/

219. Sarkar S, Arcaute E, Hatna E, Alizadeh T, Searle G, Batty M. Evidence for localization and urbanization economies in urban scaling. R Soc Open Sci. 2020;7: 191638. doi:10.1098/rsos.191638

220. Hahn FH, Losch A, Woglom WH, Stolper WF. The Economics of Location. Econ J. 1956;66: 504. doi:10.2307/2227869

221. Berry BJI, Baskin CW, Christaller W. Central Places in Southern Germany. Econ Geogr. 1967;43: 275. doi:10.2307/143299

222. 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.

doi:10.1371/journal.pone.0239699

223. Arcaute E, Hatna E, Ferguson P, Youn H, Johansson A, Batty M. Constructing cities, deconstructing scaling laws1. Arcaute E, Hatna E, Ferguson P, Youn H, Johansson A, Batty M. Constructing cities, deconstructing scaling laws. J R Soc Interface. 2015;12(102). J R Soc Interface. 2015;12. doi:10.1098/rsif.2014.0745

224. Alves LGA, Rybski D, Ribeiro H V. Commuting network effect on urban wealth scaling. Sci Rep. 2021;11: 1–10. doi:10.1038/s41598-021-02327-7

225. Prieto Curiel R, Cabrera-Arnau C, Bishop SR. Scaling Beyond Cities. Front Phys. 2022;10: 1–10. doi:10.3389/fphy.2022.858307