Mathematical modelling of social phenomena in urban areas

Carmen Cabrera-Arnau

A thesis presented for the degree of
Doctor of Philosophy

Department of Mathematics
University College London

August, 2021

Declaration

I, Carmen Cabrera-Arnau, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

The world is undergoing a rapid urbanisation process such that the majority of people now live in urban areas. In this context, it is crucial to understand the behaviour that emerges in cities as a result of complex interactions between environmental, social, economic and political factors. To improve our knowledge, different techniques are used in this thesis in order to quantitatively model how one city compares with another. Owing to the present-day ease of access to information, most of the results in the following pages have been obtained via assessment of real-world data, made available by different public organisations.

Urban scaling is used as the main modelling framework. This approach concerns the relationship between the population size of an urban area and some other urban characteristic. The work is applied to two specific topics of interest. Firstly, the amount of coverage given by the media to Mexican urban areas, before and after the 2017 Puebla earthquake, which affected several regions in Mexico. Secondly, the number of road traffic accidents per person in urban areas from several European countries for different degrees of accident severity or different definitions for the urban areas. The thesis also contains methodological contributions regarding the problem of accounting for urban areas with extremely large population in urban scaling models. Finally, this work explores the impact of the findings presented here to support the creation of new policies involving urban areas.

Impact statement

The main contribution of the research presented in this thesis is to improve our understanding of how different socioeconomic aspects vary with urban population size. This contribution is relevant considering that the world is undergoing an urbanisation process, whereby the number of people living in urban areas is increasing and many cities are growing in population.

For decades, the field of urban science has made use of mathematical models to describe or predict the behaviour of different phenomena taking place in cities. However, the rise of open data initiatives in the last few years offers new revolutionising ways to do research. This body of work takes advantage of these initiatives and delivers results which rely on both the use of mathematical models and large publicly-available data sets.

Apart from providing insightful results about the behaviour of several phenomena taking place in urban areas, this thesis also provides new definitions and methods, which could be used in applications other than the ones presented here. For example, a new index to measure the amount of media coverage in different cities is proposed, which could be used in other quantitative works about media coverage. Similarly, the methodology for the analysis of road traffic accident data could be transferred to the study of data from additional countries and to other types of event data. Finally, the thesis also contains methodological contributions regarding some statistical aspects of urban scaling models, which open new questions about the validity of this type of model.

Acknowledgements

The work presented here is the result of multiple collaborations. First and foremost, I wish to thank my supervisor, Professor Steven Bishop, for his consistent guidance, valuable contributions and generosity with time.

Thanks are also due to other people that I have worked with. Specifically, Dr Rafael Prieto Curiel who, since the beginning of my time at UCL, was willing to collaborate with me. I acknowledge working with him on research about the spread of media coverage and road traffic accident rates. Dr Humberto González Rodríguez and Dr Mara Torres Pinedo should also be mentioned here as the co-authors in my first paper on media coverage.

Dr Elsa Arcaute has provided support in a more informal capacity, by inviting me to participate in seminars and giving useful feedback on my research work. Similarly, Dr Pierpaolo Vivo and Dr Christian Boehmer have made themselves available every time I needed to talk to them.

My work relies on data sourced by different national institutions. I would like to specifically mention the UK Office for National Statistics and the Federal Statistical Office of Germany, *Statistisches Bundesamt* for answering my technical questions about their data sets.

I wish to thank UCL for their financial support to complete my research and for providing an inspiring working environment.

A special thank you to my boyfriend John for the endless and fun conversations about research and academia. And finally, my parents Miguel and Carmen, without whom none of this would have been possible.

Contents

Chapter 1

Introduction

1.1 Motivation

The world's urban population has rapidly increased in the recent decades and this increase shows no signs of diminishing. In 2020, 56.2% of the global population lived in urban areas, but as shown in Figure 1.1, this percentage is expected to keep growing in the next decades. The percentage of urban dwellers is not the same across all regions, instead, it presents variations, being as much as 82.6% in North America or 74.9% in Europe, and only 43.5% in Africa [Population Division of the UN Department of Economic and Social Affairs, 2018]. However, the emergence of hundreds of large cities, especially in Asia and Latin America [Montgomery, 2008] will likely balance these figures in the next few years.

This shift to a society of city dwellers will have a significant, but still poorly understood impact on the global environment that transcends urban boundaries [Angel et al., 2005]. On the one hand, recent studies have shown that the increasing proportion of the world's population living in cities leads to adverse consequences [Seto, Güneralp and Hutyra, 2012; Seto, Reenberg, Boone, Fragkias, Haase, Langanke, Marcotullio, Munroe, Olah and Simon, 2012; Gong et al., 2012] such as loss of biodiversity, land-cover

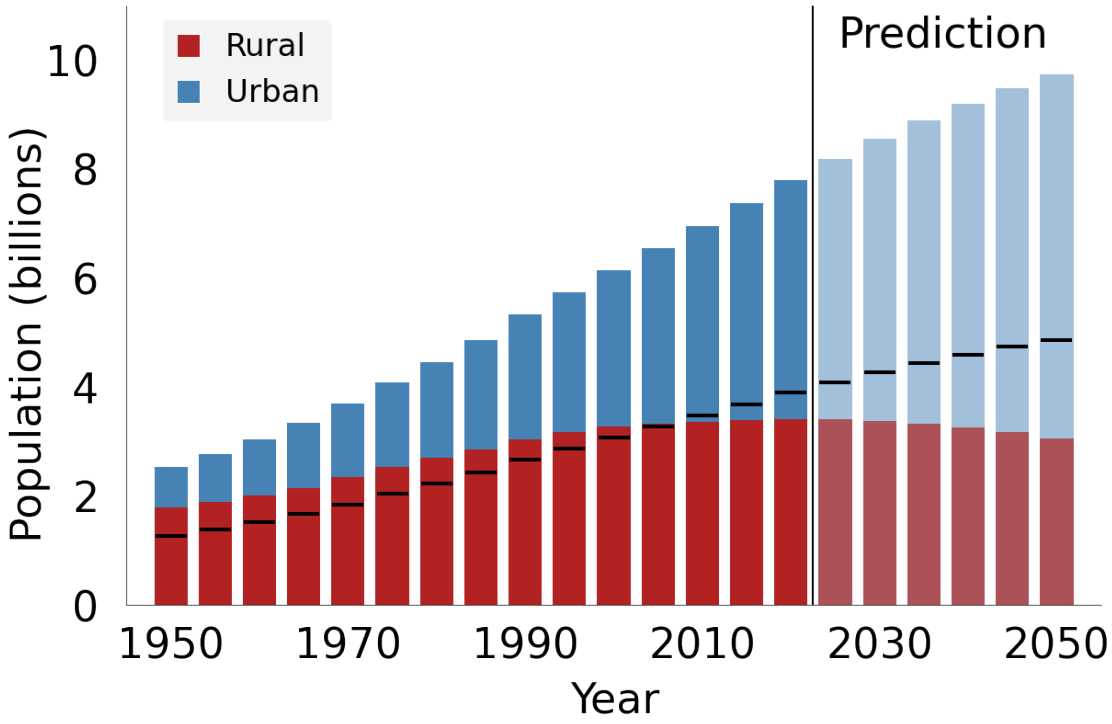


Figure 1.1: Evolution of global population, with proportion of rural and urban. For each of the depicted years, the horizontal markers correspond to half of the total population. Data retrieved from [Population Division of the UN Department of Economic and Social Affairs, 2018].

change, social disparity and deterioration of public health. On the other hand, there also exists a positive aspect of urbanisation since the effects of urban agglomeration result in increased productivity from firms [Combes et al., 2012], urban wage premium [Glaeser and Maré, 2001], improved access to healthcare [Leon, 2008] and higher concentration of highly-qualified individuals [Winters, 2011], among other aspects.

There is uncertainty, not only about the exact consequences of urbanisation, but also about its sustainability. In [Brelsford et al., 2017], an index is defined to quantify the level of development in a given region according to different measures such as access to permanent housing or sanitation. By computing the value of the index for different geographical areas, it is shown that the challenges of development are typically first addressed in

large cities, either directly via the relatively higher local taxes or indirectly via higher receipts and re-investments managed at the national level. But this development is not homogeneous and, as a consequence, severe inequalities are observed among different cities and even among different neighbourhoods.

At this point, a question arises as to whether there is an optimal population size for a city that both maximises the positive effects of urban agglomeration and hinders the negative ones. However, the choice of utility function to be optimised has been an issue of debate in the literature. For example, [Alonso, 1971] takes an economic approach to define the optimal population size for a city, depending on different priorities: i) the national government is interested exclusively in maximising total national product under conditions of labour surplus, ii) the national government is interested in maximising the total national product but in the absence of labour surplus, iii) the priority is to maximise the per capita disposable income of the inhabitants of the city itself. Richardson [Richardson, 1972], however, disagrees with Alonso's approach and defends that the fact that different priorities yield different city size optima is in contradiction with the meaning attached to the concept of 'optimal size'. Furthermore, Richardson proposes that many aspects subject to city size are not necessarily expressed in monetary terms, such as the level of pollution, noise or congestion, but they should still be taken into account. Richardson supports the idea that it is desirable to have a hierarchy of cities, but as to which hierarchy is superior, no ultimate answer can be given. More recent work, such as [Duranton and Puga, 2003], focuses instead on the derivation of the most efficient size of a city. This concept is conceived as the result of a trade-off between urban agglomeration economies and urban crowding. Following a similar approach, [Camagni et al., 2013] establishes a model of equilibrium urban size. The theoretical model is tested for several urban

areas in Europe, leading to the conclusion that there is not just a single optimal size, but many equilibrium sizes according to the presence of specific costs and advantages associated with each city.

Finding the optimal population size of a city is a problem hard to define, but it serves to illustrate that population size is a fundamental quantity in the study of cities. Firstly, describing cities or other urban settlements by their population size facilitates comparisons between them through history and across civilisations. Secondly, population size provides a very synthetic description of the relative importance of a city with respect to others, since it is often well correlated with other socioeconomic indicators [Pumain, 2000]. For these reasons, population size is considered to be 'the first dimension', i.e. the most relevant differentiation factor among a set of cities [Reiner and Parr, 1980]. A systematic knowledge of the relation between urban population size and other urban markers is therefore crucial for a successful transition to sustainability [Bettencourt et al., 2007]. Quantitative information about the effects of urbanisation can bring value in the design of policies for producing faster and more equitable development.

The main aim of this thesis is therefore to understand the effect of urban population size on the behaviour of certain phenomena taking place in cities. In particular, the interest is focused on two phenomena that have been chosen, among other reasons, due to the seriousness of their impact. Firstly, the coverage given by the media to Mexican cities before and after being hit by an earthquake. Secondly, the number of road traffic accidents in cities from different countries. In order to perform the analysis, a complex systems perspective is adopted here. The remainder of this introductory Chapter is dedicated to discuss this choice of approach and how the models from complexity science can benefit from input data. A section describing the structure of the rest of the thesis is also included.

1.2 Cities as complex systems

'The whole is greater than the sum of the parts' is the fundamental idea that, in the twentieth century, stimulated the development of a general theory of complex systems. The idea came from the realisation that phenomena of interest in many disciplines arise as a result of the interactions between lower-level entities to form a whole that displays coherent patterns and order. In the 1950s and 1960s, it became clear that cities fitted this characterisation and therefore, models from systems theory could be used to describe the mechanisms underlying the urban behaviours. Indeed, people in cities interact between themselves and the environment through the development of multiple types of land use for different economic activities. Spatial interaction is manifested by the shipment of goods for production and consumption of these economic activities, the flows of people between housing and work locations and other movements driven by the need or desire of consumers to participate in some socioeconomic activity. Furthermore, the spatial interactions are complemented by the information exchange between locations, which has increased considerably over the last decades, thanks to the boom of the Internet.

The economic activities that grant a city its identity are not unbounded. If they were, cities could thrive indefinitely and the global economy would be looking different from what it does now. Instead, cities self-regulate their own growth: the economic activities are limited by the amount of interactions that their infrastructure can accommodate. For example, central London in the UK, has reached saturation levels of traffic congestion but the tight urban space leaves little room to construct new roads. The focus is then on improving the efficiency of transport in the existing infrastructure. Self-regulation is then another key concept in complex systems theory [Batty, 2009].

While it is agreed that cities can be considered to be complex systems, there is no rigorous definition of what constitutes a complex system. Rather, complex systems theory should be regarded as a paradigm which, instead of considering that systems are centrally organised, it considers that they are structured from the bottom up and relies on the notions of emergent behaviour and self-organisation. The review article written by Ladyman [Ladyman et al., 2013] explains the features shared by objects that are commonly studied through the lenses of complex systems theory. Below, some of these features are discussed.

Emergence. The concept of emergence usually results in 'the whole is greater than the sum of the parts' and it is the essence of complex systems. The elements forming complex systems interact among themselves and with the environment, giving rise to structured behaviours. While in simple systems it is possible to infer the overall properties of the whole system by looking at the behaviour of the individual elements, this is not the case for complex systems. For this reason, the study of complex systems requires considering the objects of interest as a whole rather than just a collection of individual elements. Emergence is frequently studied via a simple model of neural networks called the Hopfield model, although in reality, this is a general model of any system formed by elements whose individual states are correlated [Bar-Yam, 1997].

Non-linearity. This characteristic is commonly seen as a failure of the principle of superposition; as Strogatz puts it, 'if you listen to your two favourite songs at the same time, you won't get double the pleasure' [Strogatz, 2000]. Non-linearity also makes cause-effect reasoning in complex systems difficult, as the direction of causality is not always clear. In the words of physicist Nigel Golenfeld 'complexity starts when causality breaks down' [Editorial, 2009]. Due to non-linearity, complex systems may display some particular behavioural patterns as they evolve, such as bi-

furcations, phase transitions or tipping points, which manifest in abrupt changes in the qualitative behaviour of their dynamics. A changing environment generally causes complex systems to evolve to different states. However, if these changes are reversed, non-linearity can prevent them from returning to the original state. This effect is known as path-dependency or hysteresis [Kapitaniak and Bishop, 1999]. Complex systems can also display chaotic behaviour as a result of non-linearity, making their exact evolution impossible to determine. Non-linearity, however, is not a necessary condition for a system to be complex. For example, models involving linear equations have been used before to study the evolution of complex systems such as a multi-species community [May, 1973] or other complex systems involving game-theoretic and quantum dynamics [MacKay, 2008]. Similarly, non-linearity is also not a sufficient condition for complexity, since there are systems formed by just one or a few elements which exhibit non-linear behaviour but would not fall in the realm of complexity science. Furthermore, systems that exhibit chaos as a result of non-linear effects, can be qualitatively indistinguishable from stochastic systems: they both have a seemingly random behaviour. However, stochastic dynamics involve, in many cases, a linear structure in their formulation. Random matrices, which are a fundamental building block of stochastic models, are a topic of interest in complexity science [MacKay, 2008].

Self-organisation and lack of central control. Complex systems display organised aggregate behaviours such as symmetry, periodicity, synchronisation or pattern formation. This ordered structure arises from uncoordinated interactions between individuals, there is no superior element in the system that coordinates all the others. Per Bak, Chao Tang and Kurt Wiesenfeld conducted some of the pioneering research on spontaneous order in complex systems [Bak et al., 1988]. In order to illustrate the idea of self-organisation, they performed simulations on a pile of sand. Their

simulations revealed that, when placing additional sand grains on the pile, it eventually reaches a stable slope: if the starting pile is too steep, the additional grains will make it collapse until it reaches the self-organised state with a stable slope; if it is too flat, the new grains will build up to also make it reach the stable slope. The shape of the sand pile remains fixed while it is at its most critically stable state. The simulations in [Bak et al., 1988] were confirmed by physical models where, instead of sand, long rice grains where used [Frette et al., 1996].

Structure by layers. Complex systems frequently display levels of structure that interact with the level above and below. Complex systems may be composed by subsystems that, in turn, have their own subsystems, and so on [Herbert, 1962]. An example of this is a single biological cell, which can be considered as a complex system formed by many molecules but at the same time, the organs are systems formed by cells. Similarly, organs can form other systems, such as the respiratory or the digestive systems, and these together form a living organism.

In the context of this thesis, urban systems can be conceptualised as complex systems structured in three layers. The first layer would be formed by the elements, either as individual urban residents or households. The second layer corresponds to the subsystems, which are the cities and other urban settlements. Urban settlements are defined by the number of elements that they contain and this measure of their size strongly differentiates them. The third layer is then the whole system of urban settlements, which interact with each other through flows of people. In addition of having a layered structure, urban systems are open, since they exchange individuals with their environment, for example migrants to and from rural areas [Pumain, 2000] and so, they also grow and change.

Numerosity. Complex systems consist of many elements. Usually, the number of elements is so large that it becomes possible and necessary to apply statistical techniques or simulations to understand their collective behaviour [Cilliers, 2002]. Numerosity is a necessary feature, however, it is not sufficient. In order to describe something as a complex system, the elements have to interact [Cilliers, 2002]. For example, an ideal gas would not be described as a complex system because the gas particles are modelled as non-interacting entities.

1.3 The role of modelling and data

Scientific models allow us to translate empirical objects into tractable formulations that can be analysed to improve our understanding of observed behaviours. In the context of social phenomena, models have been proposed to describe behaviours such as collective fear of crime [Prieto Curiel and Bishop, 2017], the dynamics of armed conflict escalation [Baudains et al., 2016], person-to-person persuasion [Smith and Curtis, 2008] or more general opinion dynamics [Deffuant et al., 2000; Weisbuch et al., 2002].

There are many modelling tactics that can be used in the application field of complex systems, however, they all share the same rationale, which is to see the collective performance of all the elements that form the system. For example, one option would be to model the behaviour of every single element and its interactions with others. Then, their collective performance can be obtained by applying statistical methods and probabilistic considerations over the ensemble of entities. This approach would be analogous to that followed in the field of statistical mechanics in physics.

Following the analogy with physics, the course of action here would correspond to that of classical thermodynamics, which is the field that motivated the development of statistical mechanics. While the main con-

cern of statistical mechanics is to explain macroscopic physical properties in terms of microscopic parameters, thermodynamics deals with the relations between macroscopic properties directly. Crucially, thermodynamics arose as an empirical field, whereas statistical mechanics has a fundamental character.

Hence, only aggregated variables regarding the objects of interest (the cities) are included in the models that appear in this thesis. 'Microscopic' measurements associated with the individual comprising parts (e.g. each person, each household, etc.) are not investigated. For example, measurements such as the total number of road traffic accidents in a city are incorporated in the models, but the number of road traffic accidents suffered by each individual residing in the city is not.

It should be noted, however, that in most thermodynamics problems, the systems under study have reached an equilibrium state. This means that, when measured, the values of the variables do not change over time or if they do, the change is so slow that the system can be regarded as being approximately in an equilibrium state. This is not the case for cities, which typically experience exponential growth rates of several per cent a year. Therefore, it is important to remark here that the application of urban scaling models in this thesis corresponds to the cross-sectional approach described in [Bettencourt et al., 2020], where population sizes and other urban properties are considered at fixed times.

It will be demonstrated how this modelling strategy that opts for a simplified version of the entities under study, still has great descriptive power.

In any case, the complex systems approach intends to capture the intricacies of reality, including emergent phenomena, feedback cycles, chaotic behaviour or cooperation and competition relations between individuals. It is for this reason that many complex systems models benefit to a great

extent from the input real-world data. In this thesis, a large amount of data is analysed. This has been possible thanks to open-data initiatives whose aim is to make data freely available to everyone without restrictions from copyright, patents or other mechanisms of control. As a result, data can be downloaded from Internet sites such as Data.gov.uk, the UK government's own open-data portal which hosts thousands of data sets.

1.4 Thesis structure

In this introductory Chapter, the motivation, overall goal and approach of this thesis have been discussed. Chapter 2 is devoted to the general methods used throughout the rest of the thesis. The focus is firstly placed on providing definitions of the main object of interest in this thesis: cities. Then, the modelling framework provided by urban scaling, which will be present in the following Chapters, is described in detail.

The following four Chapters of the thesis revolve around the analysis of different data sets and discussion of obtained results. First, in Chapter 3 the urban scaling models are used to model the relationship between population size and the amount of coverage given by the media to different Mexican urban areas. The analysis is applied at different points in time, before and after the occurrence of the 2017 Puebla earthquake, which affected a variety of Mexican urban areas. This allows us to also understand the temporal evolution of the amount of coverage.

Chapter 4 applies urban scaling models to describe the relationship between road traffic accidents taking place in urban areas from England and Wales, France and Spain and the population of these urban areas. This time, the urban areas are defined according to land use. The statistical properties of the data are taken into account to estimate the parameters in the urban scaling models. Based on the results from applying urban scaling

models, an expression for the probability of suffering a traffic accident in an urban area with a given population size is derived. The analysis also looks at the evolution in time of the number of road traffic accidents in both urban and rural areas. All the results in Chapter 4 are obtained for accidents of different degrees of severity.

A closer observation to the statistical features of the geographic and road safety data used in Chapter 4, leads to the issue of the so-called dragon-king cities. Due to their extremely large population and their out-of-ordinary features, dragon-king cities are difficult to account for in the urban scaling models. Here the issue is discussed solely in the context of urban scaling, but the problem is in reality more general, since the term 'dragon king' can refer to any type of large entities or events. In Chapter 5, different statistical methods to determine the parameters in urban scaling models are explored. The suitability of these methods to deal with dragon kings is discussed. The analysis from Chapter 5 opens a debate regarding some fundamental issues that urban scaling models are unable to address.

Chapter 6 revisits the issue of applying urban scaling models to describe the relationship between population size and road traffic accidents from different urban areas. Like in Chapter 4, the analysis includes data from England and Wales, France and Spain, but Germany is also added to the analysis. A different definition of urban area is explored, based on commuting flows. The statistical considerations discussed in Chapter 5 are taken into account for the analysis performed in Chapter 6.

While the analysis in Chapter 4 and Chapter 6 overlaps in some aspects, the approach in each Chapter is considerably different. Chapter 4 is written as an extension to one of the early research articles produced during the duration of this PhD. This early work was followed by a period dedicated to methodological research regarding the statistical aspects of urban scaling models, whose outcomes are reflected in Chapter 5. Hence,

Chapter 6 presents a more 'mature' approach to the issue of urban scaling models in the context of road safety. On the one hand, the analysis from Chapter 6 allows us to understand how the estimation of parameters may be affected by a different definition of urban areas. On the other hand, the results include new visualisations and insights that have only been obtained through experience. Finally, Chapter 6 places special focus on the behaviour of urban areas belonging to two categories: those that have dragon-king features and those that do not.

Chapters 3, 4, 5 and 6 use different technical tools and reach different conclusions. Nevertheless, they share a common structure. The four Chapters have an introductory section that highlights the relevance of the problem under study. Since the results corresponding to Chapters 3, 4 and 6 are derived from real-world data sets, each contain a section where the definitions and possible complications corresponding to the respective data sets are given. Chapter 5 bases its conclusions on synthetic data, so it also contains a section describing how this synthetic data can be produced. Every Chapter contains sections dedicated to the results and discussion.

The final Chapter of the thesis gathers some general remarks. Firstly, the results of the work are summarised. Then, the limitations of the methods and conclusions are discussed, using the experience gained throughout the thesis. The final Chapter also considers the implications of the results in other fields or in policy-making scenarios. Finally, guidelines and ideas for future research are provided.

1.5 The use of the term 'road accident'

A note on terminology should be given here regarding the use of the term 'road accident' or variations of it such as 'road traffic accident', 'traffic accident' or simply 'accident'. In this thesis, these terms were chosen to

refer to incidents that take place in public roads where at least one person is injured and which involve at least one vehicle. This choice is in line with the terminology used by institutions such as `gov.uk` or the UK Office for National Statistics.

Even though the term 'road accident' is widespread, it is important to remark here that many groups recommend replacing it with other terms such as 'road traffic collision' or 'road traffic crash'. The reason for this recommendation is that the word accident can imply that the incident is unintentional or that there is no-one to blame, but the reality is that many accidents are the consequence of reckless driving behaviour, such as drink-driving or excessive driving speed.

Therefore, it should be clarified that the term 'road accident' will be kept here, however, no 'accidental' connotation should be attributed to its meaning.

Chapter 2

Methods

2.1 Introduction

Essentially, the aim of this thesis is to understand how the prevalence of certain social phenomena in cities might be affected by urban population size. In order to achieve the aim, a definition of what constitutes a city needs to be established. However, there is no single way of establishing the boundaries of cities and different criteria are often chosen according to the type of analysis to be performed. The first section of this Chapter is dedicated to establish how cities will be defined to study the different events of interest in this thesis.

Once the city boundaries are determined and the data is collected, models that capture the urban behaviours can be applied. This Chapter describes the different avenues for urban modelling that have been explored as part of the research process during the writing of the thesis. However, special emphasis is placed on urban scaling models, due to their suitability to model the relation between urban population size and the incidence of different events.

2.2 Defining urban areas

In the past, the size of a city was limited by how far we could walk or ride a horse and cities rarely reached a population of more than 1 million. The industrial revolution facilitated the use of faster and more efficient transport, hence changing the morphology and size of cities. In the current period, the transition from energy to information is destined to physically alter our cities once again. While the thriving infrastructures and information technologies are driving an increasing need for face-to-face interactions, they also facilitate remote work and virtual meetings. As a result, our cities are growing large, integrated cores and, at the same time, they are sprawling at ever-lower densities in their peripheries [Batty and Ferguson, 2011]. Defining a city boundary is more challenging than ever before and, given the global reach of information, these boundaries are becoming less relevant. However, in order to better understand how urbanisation is likely to affect our lives, establishing what does at what does not constitute a city is a priority. Taking these considerations into account, cities and towns that have traditionally been regarded as different entities, may be classified as just one nowadays. For this reason, instead of city or town, the term urban area will be used henceforth.

There is no single way of establishing the boundaries of urban areas and different criteria are explored in this thesis. Firstly, the Mexican Metropolitan Zones are a convenient choice of urban areas for Chapter 3. However, these are only defined for populations of at least 100,000. Hence, a different definition is used for other urban areas that are not listed as Metropolitan Zones but that are still of interest for the analysis. For the study of road traffic accidents in Chapter 4, a criterion based on land use is deemed appropriate, as the environmental features of roads play a very significant role in conditioning the occurrence of an accident. But as pointed out by

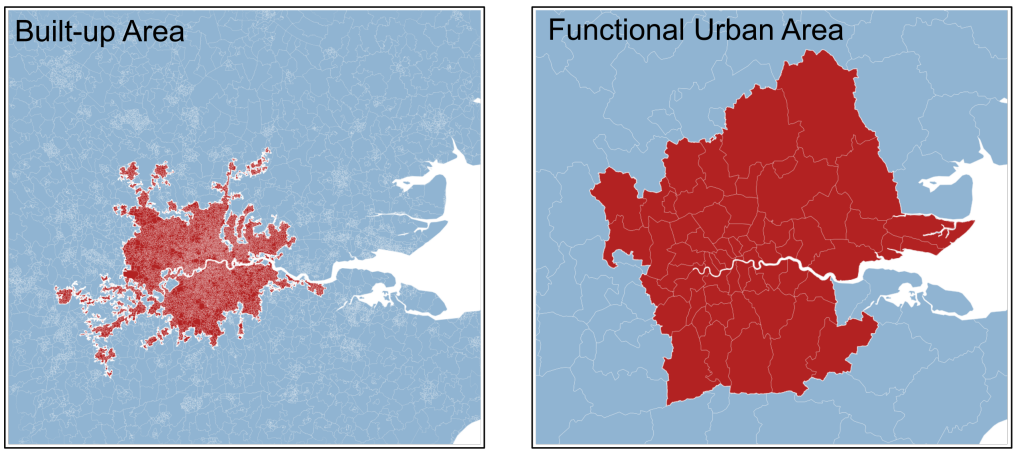


Figure 2.1: The urban area corresponding to London, England. On the left panel, the highlighted area is known as London's Built-up Area. It is defined as a cluster of Lower Layer Super Output Area divisions that satisfy a set of land-use criteria. The highlighted area on the right panel is known as London's Functional Urban Area and it is defined as a cluster of local authority units that satisfy a set of criteria regarding commuting flows.

Batty and Ferguson in [Batty and Ferguson, 2011], it is the exchange of information and interactions what determines the cohesion of cities nowadays so, like in [Arcaute et al., 2015], a classification based on commuting flows is also used to analyse road traffic accidents in Chapter 6. Figure 2.1 illustrates how the boundaries of a city can differ when defined according to different criteria. Specifically, it depicts the urban area corresponding to London defined by land use and by commuting flows. In the next sections, details and references about these methods are provided.

It should be mentioned that each country analysed in this thesis makes its data available for different types of geographical hierarchies. In Mexico, data is available for the Mexican municipalities, which are the second-level administrative units, with the Mexican states being the first-level administrative units. Population and road traffic accident data are available by the

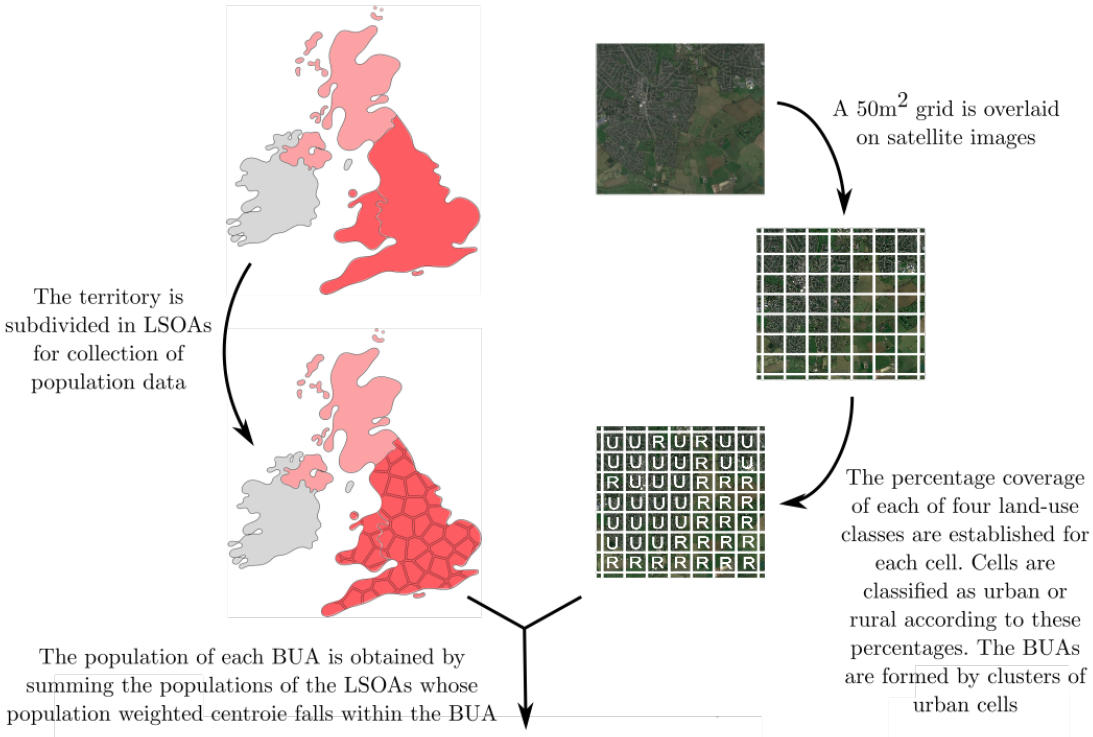
local authority unit in France, Germany and Spain (*commune*, *gemeinde* and *municipio* respectively). In the case of England and Wales, data is available for lower level geographic hierarchies known as Lower Layer Super Output Areas (LSOAs), designed specifically to improve the reporting of small area statistics.

As explained by Pumain in [Pumain, 2004], for some administrative or political purposes, it may be relevant to consider each municipality, local authority or LSOA as a separate and autonomous entity. However, for a geographical or systemic study, a better definition is to aggregate within a single urban agglomeration all urban geographical hierarchies which are contiguous and whose urbanisation usually results from the historical growth around an older urban centre. The latter definition is recommended by the UN for the production of urban statistics and it is the one that will be taken in this thesis.

Therefore, urban areas in each country are defined here as groups of geographical hierarchies for which data is available. Then, the production of statistics about an urban area can be done simply based on the geographical hierarchies that are classified as part of that urban area. Figure 2.2 displays an example of the process whereby data is assigned to an urban area. In this case, the total population of the built-up areas (BUAs) from England and Wales is being computed as the aggregated data corresponding to the LSOAs that are within each BUA's boundaries.

2.2.1 Urban areas in Mexico

In Chapter 3, the media coverage given to Mexican urban areas of different sizes is analysed. For the purposes of this analysis, information about the population size of each urban area is needed. Hence, it is also necessary to define the boundaries of the urban areas. Some of the urban areas under consideration are large metropolis that span over several *municipios* or



BUA_Name	BUA_Code	Population
Greater London BUA	E34004707	10072820
Greater Manchester BUA	E34005054	2581872
West Midlands BUA	E34005001	2476649
West Yorkshire BUA	E34004684	1806287
Liverpool BUA	E34004801	869343
⋮	⋮	⋮

Figure 2.2: Diagram of the process to assign population data, collected for the Lower Layer Super Output Areas (LSOAs), to the built-up areas (BUAs) from England and Wales.

municipalities, which are the second-level administrative units in Mexico. When this is the case, their boundaries are set to be those of the Mexican metropolitan zones (MZs), which are established by the Secretariat of Agrarian, Land, and Urban Development (SEDATU), the National Population Council (CONAPO) and the National Institute of Statistics and Geography (INEGI). The methodology to establish the MZs can only be found in Spanish in [Secretaría de Desarrollo Agrario Territorial y Urbano and Consejo Nacional de Población and Instituto Nacional de Estadística y Geografía, 2015]. Below, the main steps of this process are summarised. In the analysis, other urban areas that do no satisfy the criteria to be classified as a MZ are also considered. When this is the case, their population size is set to be that of the corresponding municipality where they are located, according to the population figures published by INEGI [Instituto Nacional de Estadística y Geografía, 2015].

Metropolitan zones

The MZs are groups of second-level administrative units, known as municipalities in Mexico. The municipalities forming a MZ are of two types:

1. The inner municipalities belonging to the core city. They must have the following characteristics:

    (a) They share an inter-municipal conurbation with population equal or larger than 100,000, defined as a continuous built-up area that spans over two or more urban localities from different municipalities with urban characteristics. A locality is defined as a settlement with one or more dwellings, which may be inhabited or not. A locality is said to be urban if it has a population of at least 2,500 and rural otherwise. The methodology does not specify what is meant by 'built-up'.

(b) They contain localities with 100,000 or more residents, which show a high level of physical and functional integration with neighbouring urban municipalities. Functional integration is determined according to commuting flows.

(c) They contain localities of population size 500,000 or more, or localities which are state capitals, as long as these do not belong to another MZ already.

(d) They contain localities of population size 200,000 or more which are located at a maximum distance of 20 km from the Northern or Southern borders or which are located on the coastline.

2. The outer municipalities, which are not part of the inter-municipal conurbation shared by the inner municipalities, but that display urban characteristics and maintain a high level of functional integrity with the core municipalities. They must fulfill the following statistical and geographical criteria:

(a) Their most populous urban locality is no further than 15 km away from the core city and is accessible via a double lane road.

(b) At least 15% of their active population (aged 15 to 70) works in the inner municipalities and/or at least 10% of the people working in the outer municipalities have their place of residence in the inner municipalities.

(c) At least 75% of the active population works on the secondary or tertiary sector.

(d) The average population density is more than 20 people per hectare.

The outer municipalities can also be defined according to other urban planning and policy criteria. If this is the case, the municipalities must also fulfill criteria 2.(b), 2.(c), 2.(d) and/or be recognised by the federal, state and municipal governments as part of the corresponding MZ for urban development and territorial organisation purposes.

2.2.2 Urban areas based on a land-use criterion

England and Wales

The built-up areas (BUAs) in England and Wales (E&W) are geographical units, which correspond to cities, towns and other urban settlements. Their physical boundaries are generated by firstly subdividing the territory in a 50 m^2 grid based on the British National Grid System. Then, based on satellite imaging, the proportion of territory corresponding to each of the four established classes of land use (man-made, natural, agricultural or water) is determined for each cell. Cells are then classified as urban or rural according to these percentages. All the contiguous cells that are considered of urban land use are grouped into urban polygons. If an urban polygon is larger than 200,000 m^2, it is said to form a BUA; if two or more polygons are less than 200 m apart and their joint area is above 200,000 m^2, they are considered to be the same BUA; if there are non-urban cells inside an urban polygon whose total area is above 200,000 m^2 corresponding, for example, to a park, these non-urban cells are also classified as part of the BUA. Full details about this classification process can be found in [Department for Transport, 2017*a*].

However, data is published in the UK at the Lower Layer Super Output Area (LSOA) level and, as mentioned above, the urban areas from the countries analysed in this thesis are defined as groups of geographical hierarchies for which data is available. In order to match the grid cells with

the LSOAs, a lookup table [Office for National Statistics, Open Geography Portal, 2011] can be used that for each BUA, lists all the LSOAs whose population weighted centroid (PWC) lies within the BUA.

France

For the purposes of this work, the so-called *unités urbaines* (UUs) are considered to be the urban areas in France. Similar to E&W, the definition of UU relies on the construction of patches of built-up land. This is done through the analysis of BD TOPO, a data base that provides 3D vector descriptions of the elements covering France's land. All buildings for residential purposes as well as other public, industrial and commercial spaces are considered for the construction of built-up patches. The built-up patches used for the definition of UUs must contain no gaps of more than 200 m between buildings. Special considerations are taken for bodies of water that are located within a built-up patch or that divide a continuous stretch of built-up land into regions.

Additionally, these built-up patches must have a population of at least 2,000. The population of these patches is obtained from the administrative dataset Fideli, which contains demographic information on dwellings and individuals in France, and from GEOSTAT 2011, which provides population data in a 1 km grid that extends across 29 EU member countries.

But, as in Mexico and E&W, the boundaries of the built-up areas must be matched with geographical hierarchies of practical relevance. In the case of France, these are known as *communes*. Hence, a UU is formed by a set of communes which either:

1. Contain entirely a built-up patch with a population of at least 2,000.

2. Contain only a fraction of a built-up patch, but more than 50% of the *commune*'s population is concentrated there.

3. Contain only a fraction of a built-up patch, but more than 2,000 people in the *commune* live there, even if this is less than 50% of the total population.

Minor adjustments are then performed to adjust some of the UUs obtained with the process described above. Details for these adjustments as well as a more extended explanation of the process can be found in French in [Institut National de la Statistique et des Études Économiques, 2018*b*].

Spain

Unlike in E&W and France, Spain uses population statistics as the basis to establish its urban areas and only considers land-use data at a later stage in the process. The methodology, which can be found online in Spanish [Ministerio de Fomento, 2018*a*], mentions that land-use data is indeed considered in the definition of the urban areas, but the exact details regarding how it is used are not provided.

The Spanish territory is split into *municipios* or municipalities, which are administrative divisions similar to the French *communes*, although generally with larger populations and areas. The Spanish urban areas or *Áreas urbanas* (AUs) can be formed by one or more municipalities and they are of two types:

1. Great urban areas. Generally, they are formed by at least one municipality with a population of at least 50,000. All the municipalities included in a great urban area have a population of at least 1,000.

2. Small urban areas. This type can be split into two subcategories:

   (a) On the one hand, AUs with a population size between 20,000 and 50,000 which are not included in group 1.

(b) On the other hand, AUs corresponding to municipalities with a population size between 5,000 and 20,000. From all the municipalities whose population falls in this range, only some of them are considered to be AUs. Among those are the municipalities with a population of over 10,000 in the municipality nucleus. The Spanish documentation for the definition of UAs does not specify how this nucleus is determined, however, it suggests that a land-use criterion is used for this purpose.

2.2.3 Urban areas based on commuting flows

In Chapter 6, the incidence of road traffic accidents is explored for urban areas defined according to commuting flows. Given that all the countries explored in Chapter 6 are European, the functional urban areas (FUAs) established by Eurostat [Eurostat, 2021*a*], which is a Directorate-General of the European Commission, are used as the urban areas. The construction of FUAs consists of a two-step process whereby 'cities' are defined first and FUAs are established using the cities as a starting point. The full methodology can be found in [Eurostat, 2021*b*], however, for ease of comparison with the urban areas described in the previous sections, the basic procedure is provided here. Firstly, the so-called 'cities' need to be determined according to the following steps:

1. A population grid of 1 km^2 is constructed. All grid cells with more than 1,500 inhabitants are selected.

2. The selected grid cells are clustered and any gaps withing a cluster are filled. Then, only the clusters with a population of at least 50,000 are kept and referred to as urban centers.

3. Then, the urban centers are matched with the local authority units (LAUs), which are the geographical hierarchies of practical relevance

in each country, known as districts or unitary authorities in England, principal areas in Wales, *communes* in France, *gemeinden* in Germany and *municipios* in Spain. If at least half of a LAU's population is inside the urban centers, then this LAU is selected as a candidate to become part of the city.

4. The candidate LAUs are filtered ensuring that the resulting city has a link to the political level, at least 50% of its population lives in a urban centre and at least 75% of the population of the urban centre lives in a city.

Once the cities are established, some further steps need to be taken to define the FUAs:

- If 15% of the active population of a city work in a different city, these two cities are treated as the same FUA.
- All the LAUs where 15% of the active population works in a city are selected and added to the corresponding FUA. If this is the case for more than one city, the LAU is assigned to the FUA of the city with the largest population.
- LAUs that are not contiguous to the rest of their corresponding FUA are dropped. LAUs that have not been selected but are fully surrounded by a single FUA are included in the FUA.

2.3 Modelling urban areas

As already discussed in Chapter 1, the global population has already shifted from being mostly rural to being mostly urban over the last few decades. The aim of this work is then to improve our understanding of the effect that city population size has on the behaviour of phenomena taking place in urban areas. A combination of modelling techniques and a large amount of input data are used to achieve this aim.

2.3.1 Background

Modelling the behaviour of urban areas accurately has become a priority since the urbanisation process shows no signs of ceasing. While economics and regional science were the first disciplines to tackle the problem of modelling urban areas, the rise of agent-based models and other computational techniques in the 1950s and 1960s was crucial for understanding urban areas.

There are lots of aspects of urban areas that can be modelled, but in this thesis, the focus is on urban scaling models, due to their suitability to describe relations between the population size and other variables of interest. Below, a more detailed description of this type of model is given.

Other relevant models are also included below: models for urban growth and the distribution of urban population size, spatial interaction models and network theory models. While these other models have not been used in this thesis as prominently as urban scaling models, they have been explored as part of the research process. Hence, it is deemed appropriate to include them here for completion and also, to encourage their use as alternative approaches to the problems addressed in the thesis. For an exhaustive review of urban models, see [Barthélemy, 2019].

2.3.2 Urban scaling models

When it comes to quantifying different aspects of cities, simple per capita measures are most commonly used. However, these assume implicitly that urban characteristics increase linearly with population size. This assumption is not entirely correct as it ignores the inherent non-linear nature of the organisation and dynamics of cities as they grow.

Following the tradition inherited from allometry theory [Haldane, 1985; Thompson, 2014] —which studies the relationship between body size of an organism and other features such as shape, anatomy or physiology— it is typically hypothesised that environmental, economic and social properties of a city scale as a power law of the city population size [Bettencourt et al., 2010] so that if X is the population of a city and Y is an urban indicator, then:

$$Y(X) = \alpha X^{\beta}, \tag{2.1}$$

where the scaling exponent β is, in general, different from 1, and α is a proportionality constant.

In the context of allometry theory, assuming the form of model in equation 2.1 to fit to data, it has been found before that larger warmed-blooded animals are more metabolically efficient than the smaller ones, since their basal metabolism grows sublinearly with body weight [Kleiber, 1932]. This result is known as Kleiber's Law, depicted in Figure 2.3, where the horizontal axis represents the body weight of a mammal and the vertical axis is the metabolic rate measured in calories per kilogram of body weight. If both axis are represented in logarithmic scale, it can be seen, perhaps surprisingly, that the data lies on an almost perfect straight line of slope $\frac{3}{4}$.

Analogously, the economic productivity of a city varies with its population size according to $\beta = 1.15$ [Bettencourt et al., 2010], i.e. it increases systematically by 2.21 times its value with every doubling of a city's pop-

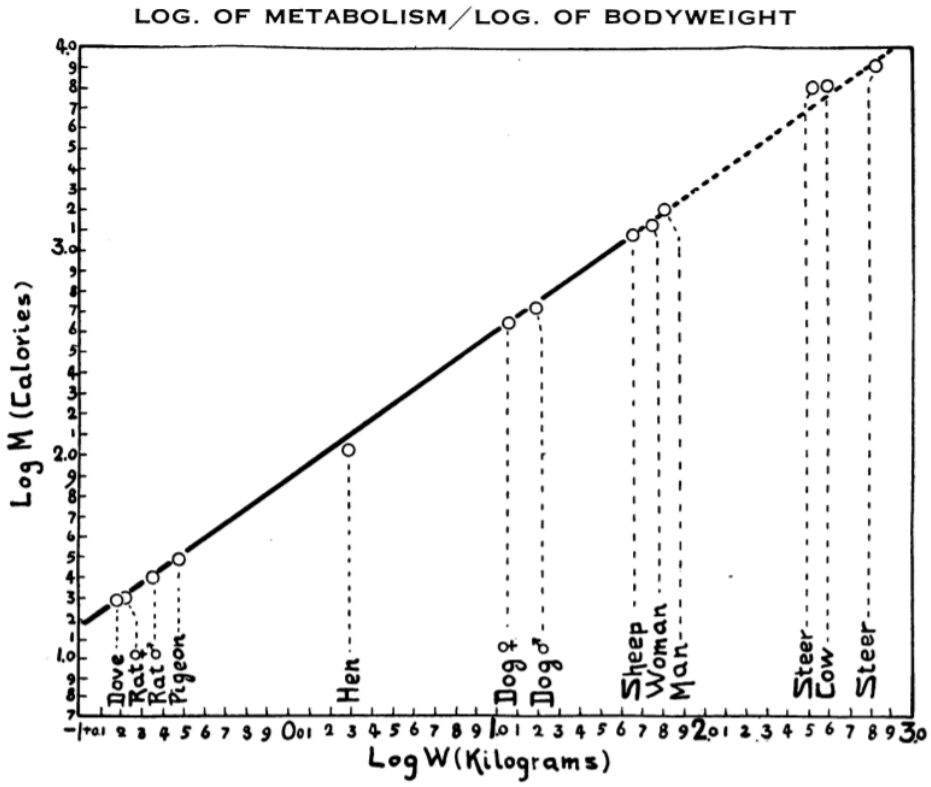


Figure 2.3: Representation of Kleiber's Law, as an example of an allometric relation. The law states that for the majority of mammals, their metabolic rate scales to the $\frac{3}{4}$ power of their mass. The Figure is from Kleiber's original 1932 article [Kleiber, 1932]. The two data points with the label 'steer' correspond to groups of steers of different sizes.

ulation. The walking speed [Bornstein and Bornstein, 2007], the criminal activity [Glaeser and Sacerdote, 1999], the CO2 emissions [Fragkias et al., 2013], the average number of contacts and communication activity [Schläpfer et al., 2014], the economic diversification [Youn et al., 2016], the road length distribution [Strano et al., 2017], the number of people migrating to a city [Prieto Curiel, Pappalardo, Gabrielli and Bishop, 2018], the amount of media coverage received by a city [Prieto Curiel et al., 2019] or the number of road traffic accidents [Cabrera-Arnau et al., 2020], have all been found to scale as a power law with city size.

Apart from providing an analogy with allometry theory, using power laws to model urban indicators provides with a simple, straight-forward way of classifying a set of data points as linear, superlinear or sublinear, depending on whether the value of the scaling exponent β is equal, larger or smaller than 1. This value determines whether larger cities are more efficient or productive than the smaller counterparts with respect to a certain urban characteristic, or in other words, whether that characteristic follows an economy of scale [Marshall, 2013]. However, a value of β larger than 1 does not necessarily mean more efficiency or productivity, instead, the value needs to be interpreted according to the urban indicator being analysed. For example, if β was less than 1 in the case where Y is the number of petrol stations per person, this would mean that indeed, cities become more efficient with respect to the number of petrol stations as they grow in population size. But if β was less than 1 for Y being the number of patents, this would mean that as cities grow, they do not become more productive with regards to the number of patents.

Scaling models allow the long-standing problem of how to rank cities to be addressed in a meaningful manner. Through the observation of scaling behaviour, it is possible to predict, in an approximate way, the expected average characteristics that a city of a given size should manifest. What is more, deviations from urban scaling models become sometimes the most interesting pieces of information for both policy and scientific analyses, as they are usually the result of local characteristics that make a city exceptional with respect to its peers [Bettencourt et al., 2010].

Table 2.1 is based on a similar table found in [Bettencourt et al., 2007] and gathers values of the scaling exponent β estimated for different data sets from USA, Germany and China by several authors. It also includes the scaling exponents estimated in this thesis, whose respective rows have been highlighted in red.

This thesis explores, first and foremost, different applications of urban scaling models, but it also investigates the weaknesses of the models. The reader of this work should then keep in mind that the different Chapters of the thesis have been written, roughly, in chronological order. Hence, Chapter 3 is the starting point for the application of urban scaling models from a somewhat naive standpoint. As the thesis progresses, other applications of urban scaling models are explored and some methodological aspects are analysed. The experience accumulated through the different Chapters leads to an increasingly critical view of urban scaling models. This view, shared in Chapter 7, is in line with the view of Shalizi in [Shalizi, 2011].

2.3.3 Models for urban growth and the distribution of urban population size

It is generally accepted that the large disparity in the observed population size of urban areas, which varies across several orders of magnitude, is captured by Zipf's empirical law. This law is closely related to the probability distribution of finding an urban area with a given population size. Understanding the mechanisms of population growth that lead to such distributional patters is among the most fundamental aspects of the study of urban areas. For example, in 1931, Gibrat proposed the rule of proportional growth which, when applied to urban areas, states that the population growth rate is proportional to the current population size [Gibrat, 1931]. However, it can be shown that Gibrat's rule is not fully consistent with Zip's law and other urban growth models have been since proposed in order to capture the same behaviour of the empirical results. One of the most remarkable approaches is Gabaix's [Gabaix, 1999], which is based on Gibrat's but with the constraint that small urban areas cannot shrink to zero.

Table 2.1: Exponents with 95% confidence intervals for urban indicators found in [Bettencourt et al., 2007], corresponding to different countries and years. Highlighted in red are the results found in this thesis. The confidence intervals marked with an asterisk have been computed based on a Gaussian approximation.

Variable	β	**95 % CI**
Gasoline stations US, 2001	0.77	[0.74, 0.81]
Gasoline sales US, 2001	0.79	[0.73, 0.80]
Road surface Germany, 2002	0.83	[0.74, 0.92]
Length of electrical cables Germany, 2002	0.87	[0.82, 0.92]
Accidents in functional urban areas Germany, 2018	0.92	[0.86, 0.97]
Accidents in functional urban areas England & Wales, 2018	0.99	[0.93, 1.04]
Household electrical consumption Germany, 2002	1.00	[0.94, 1.06]
Total employment US, 2001	1.01	[0.99, 1.02]
Household water consumption China, 2002	1.01	[0.89, 1.11]
Accidents in functional urban areas Spain, 2015	1.03	[0.93, 1.14]
Household electrical consumption China, 2002	1.05	[0.89, 1.22]
Total electrical consumption Germany, 2002	1.07	[1.03, 1.11]
Total bank deposits US, 1996	1.08	[1.03, 1.11]
Accidents in built-up urban areas France, 20018-18	1.10	[1.09, 1.11]
Accidents in built-up urban areas England & Wales, 2008-18	1.11	[1.10, 1.12]
Total wages US, 2002	1.12	[1.09, 1.13]
Accidents in functional urban areas France, 2018	1.13	[0.96, 1.27]
Supercreative employment US, 2003	1.15	[1.11, 1.18]
Accidents in built-up urban areas Spain, 2015	1.16	[1.15, 1.16]
GDP China, 2002	1.15	[1.06, 1.23]
New patents US, 2001	1.27	[1.25, 1.29]
Private R&D employment US, 2002	1.34	[1.29, 1.39]
Media coverage post-earthquake Mexico, 2017	1.60	[1.40, 1.80]*
Media coverage pre-earthquake Mexico, 20017	1.70	[1.66, 1.74]*

Chapter 5 elaborates on urban growth models and the issue of the distribution of urban population size, since these areas are in close relation with the concept of dragon king. Additionally, Pumain gives an extensive review of modelling approaches to these problems in [Pumain, 2000].

2.3.4 Spatial interaction models

While models for population growth focus on the evolution of a single urban area over time, another aspect of interest is the link between these patterns with the large-scale behavior described by Zipf's law. In order to understand this link, models for the global dynamics of a system are needed. In particular, it is necessary to understand the flows of people, information and goods between cities. A variety of models have been proposed to model flows in the social sciences, such as the diffusion model introduced by Bouchaud and Mézard in the context of wealth distribution [Bouchaud and Mézard, 2000]. However, in this thesis the focus is on spatial interaction models, since they lie in the heartland of quantitative geography.

Spatial interaction can be defined as the flow of people, commodities, capital or information from an origin location to a destination. The term thus includes movements such as migration, travel-to-work, shopping, recreation, commodity flows, capital flows, communication flows, airline passenger traffic, the choice of health care services, and even the attendance at events such as conferences, cultural events and sport events [Haynes and Fotheringham, 1984]. Usually, each agent involved in the movement trades off the benefit of moving to a different location with the costs that are necessary to overcome the spatial separation between the origin and destination [Fischer, 2006].

Mathematically, spatial interaction models consist of discrete origin/destination pairs, which can be represented in a matrix where the rows and columns are related to the origin and destination locations respectively.

Such matrix is known as the origin/destination or OD matrix. The actual value of each entry in the OD matrix corresponds to the volume of people, goods or other commodities moving between each origin and destination. The most general formulation of spatial interaction models is as follows:

$$T_{ij} = f(V_i, W_j, S_{ij}), \tag{2.2}$$

where T_{ij} is the entry corresponding to row i and column j and represents the volume of the flow between origin i and destination j. The right hand side of equation (2.2) indicates that T_{ij} is a function of the attributes V_i of origin i, the attributes W_j of destination j and the attributes S_{ij} of the separation between origin i and destination j. The attribute variables corresponding to the origin and destination, V_i and W_j, often have a socioeconomic nature and represent aspects such as population, number of jobs available, industrial output or gross domestic product. The attributes of the separation between origin and destination Sij are usually related to the distance, transport costs, or travel time.

There exist several formulations of equation (2.2), but that corresponding to gravity models remains to be the cornerstone of spatial interaction models. Gravity models owe their name to the fact that they were built in analogy with Isaac Newton's model for gravitational law. The first gravity models are usually attributed to Carey [Carey, 1858] and Ravenstein [Ravenstein, 1885]. However, Odlyzko [Odlyzko, 2015] traces their origin back to the 1846 publication by Belgian civil engineer Henri-Guillaume Desart. In this publication, Desart produced an analysis of an extensive and unique dataset for passenger travel on Belgian railways and he also came up with a first version of a gravity model. Odlyzko argues that, "had the validity and value of gravity models been recognized properly, the investment losses of [the great Railway Mania in Britain] could have been lessened,

and more efficient rail systems in Britain and many other countries would have been built".

Newtonian gravity is based on the idea that the attraction between two objects is proportional to their mass and inversely proportional to the (square of) the distance between them. Consequently, the general formulation of spatial interaction models can be adapted to reflect this basic principle. In their most basic form, gravity models for spatial interaction are given by

$$T_{ij} = k\frac{V_i W_j}{S_{ij}} \tag{2.3}$$

where k is a proportionality constant related to the rate of the flow (e.g. k will be higher if the flow T_{ij} is considered over the period of a year than if it is considered over a month) and S_{ij} is simply the physical distance between origin i and destination j.

While simple, the above formulation of gravity models is rigid. A more flexible formulation is usually given by

$$T_{ij} = k\frac{V_i^{\alpha} W_j^{\gamma}}{S_{ij}^{\beta}}, \tag{2.4}$$

where the new parameters α and γ are known as 'emissivity' and 'attractiveness' or potential to generate or attract movements; and β is a parameter that corresponds to the efficiency of the transport system between origin and destination. Again, S_{ij} is usually taken to be the physical distance between locations i and j. Estimating the value of these parameters, however, is a significant challenge in the use of spatial interaction models, since testing the validity of the estimations requires large amounts of mobility data which is not always easy to obtain.

Gravity models were originated as an attempt to use Newtonian physics to explain social science phenomena. This Newtonian physics approach, however, was criticised for its confusing social science interpretation. One of the most influential attempts to detach gravity models from Newtonian physics was due to Alan Wilson [Wilson, 1967]. Wilson used statistical mechanics to propose a theoretical justification based on the principle of entropy maximisation. This approach to the problem allowed for aggregate interaction to be treated as a basic estimation problem in information theory as opposed to a thermodynamics problem. However, many social scientists were still concerned about Wilson's approach being yet another social physics analog. Since Wilson's 1967 publication, more efforts have been placed on removing gravity models from the social physics framework and hence, giving them a purely probabilistic interpretation [Haynes and Fotheringham, 1984].

Spatial interaction models can be used to model interurban mobility as well as the small-scale behaviour of cities. However, it should be mentioned here that there exist models that have been specially conceived to describe the internal distribution of the population in urban areas. An example of this is the Alonso–Muth–Mills model, considered to be one of the pioneering works in urban economics [Brueckner, 1987]. In this modelling approach, all the individuals seek to maximise the same utility function and this eventually determines their location within an urban area. Krugman's work is another example of a model for the distribution of people and economic activity in specific regions within the same urban area [Krugman, 1996].

2.3.5 Network theory

As complex systems, cities display highly non-trivial interactions between their socioeconomic and infrastructural elements. These interactions are very diverse and take place at different levels and scales. For example, at a

local scale, people interact with each other, forming personal relationships that may be affected and may affect the geography of the city were they live. Locations in a city may also interact with each other, through the transport network and the topology of the roads. These interactions are not bounded to a single city, but they can also take place at a larger scale between cities. The raise of the Internet facilitates virtual interactions and flow of information between cities. Furthermore, good road connections, railway or air travel may keep cities in close connection.

The interactions between the elements of a complex system can be conceptualised and analysed through network theoretic considerations. Mathematically, a network or graph G is defined as an ordered pair $G = (V, E)$ where:

- V is the set of vertices, which represent the elements that form the system.
- $E \subseteq \{\{x, y\} \mid x, y \in V\}$ is the set of edges, which are pairs of vertices. An edge between two vertices indicates the presence of an interaction between the corresponding elements of the system.

By including additional constraints on the set of edges, different types of networks can be defined. For example, directed networks are those where the direction of the edges matter or simple networks are those that do not allow for edges from one vertex to itself.

Based on this definition of a network, a variety of network metrics can be calculated to understand the behaviour of the system being modelled. Some of these metrics are: the size of the network (number of vertices or edges), the density (ratio between number of present edges to the total number of possible edges), the degree distribution (distribution of the number of edges connected to each vertex), the shortest path between two vertices (minimum number of edges to connect from one vertex to another), different

measures of centrality of a vertex or an edge (these metrics quantify the 'importance' of a vertex or an edge), etc. Detailed explanations of these concepts and more can be found in [Newman, 2010].

Network models have been used before in the context of cities and the applications are very diverse. For example, the authors in [Zhong et al., 2014] construct a network model based on a data set from Singapore's automatic smart card fare collection system used in public transport. Each vertex in the network represents a station and the edges denote the possibility of travel between any two stations. The edges are weighted according to the volume of travel, which is the number of trips made. The authors then apply several network theoretic calculations in order to obtain an overall view of travel demand, detect urban centers and hubs, and uncover socioeconomic clusters defined as neighborhoods and their borders. By projecting the results of the network analysis back onto geographical space, the spatial structure of urban movements is revealed, showing that the population's activity is separated in different areas which differ from the existing administrative ones. In addition, the authors find that Singapore is developing rapidly towards a polycentric urban form, where new subcenters and communities are emerging largely in line with the city's master plan.

Another example of an application of network models in the urban context is shown in [Barthélemy et al., 2013], where the authors analyse, over a period of more than 200 years, the evolution of different network metrics corresponding to the road network of Paris. Results show that these metrics evolve smoothly, suggesting that most changes in the street network are the result of 'self-organisation', that is, of naturally occurring, continuous, local growth processes. The smooth evolution of most metrics takes place despite the significant perturbations that happened around the mid-1800s, when Napoleon II commissioned the architect Georges-Eugène Haussmann to modernise Paris by building large squares and avenues connecting points

of interest and by improving the traffic flow and the circulation of army troops. The transformative changes of the Haussmann period manifest in the spatial reorganisation of the most central vertices of the street network.

Although not directly related to urban modelling, one further application of network theory that should be mentioned here due to its relevance to the Covid-19 pandemic, is shown in [Colizza et al., 2006]. The authors of this publication analyse the complete worldwide air transportation network and find that its properties are responsible for the global pattern of emerging diseases. Their analysis shows that large-scale mathematical models that take fully into account the complexity of the transportation network can be used to obtain forecasts of emergent disease outbreaks. Additionally, they provide a tool to quantify the predictability of epidemic patterns based on the computation of confidence intervals in epidemic forecast and in the risk analysis of containment scenarios.

Network theory provides a versatile framework which, as demonstrated through the applications mentioned above, has had a major impact in the development of urban modelling. The applications, however, span far beyond the realm of urban modelling and concern aspects such as the growth dynamics of complex systems subject to network interactions [Barabási and Albert, 1999; Barrat et al., 2004] or the characteristics of network building blocks, known as motifs, which could uncover structural design principles of networks [Milo et al., 2002].

Chapter 3

Relationship between media coverage and population size of Mexican urban areas

3.1 Introduction

The work presented in this Chapter has been covered in the research paper entitled 'Temporal and spatial analysis of the media spotlight' [Prieto Curiel et al., 2019], to which the author of this thesis contributed.

From all the events which occur daily, only a few are deemed to be newsworthy enough to be reported as news in traditional print newspapers or online [Harcup and O'Neill, 2001]. The stories which are picked up by the press usually have special attributes, such as unexpectedness, major negative consequences, an effect on the social elite, violent content, eye-catching pictures or a widespread impact [Chermak and Gruenewald, 2006; Galtung and Ruge, 1965].

Media stories are the outcome of complex interactions between the different elements that are involved in the publication process. Firstly, editors of different media outlets set a general agenda and then coordinate with journalists to decide what is newsworthy. However, the audience is selective in their media choices [Berelson, 1952] and their attention span is limited [Simon, 1971] so that collectively, they decide what they want to consume [Morley, 2003]. Feedback from readers needs then to be considered in order to create media content matching their interest. Traditionally, this feedback was gleaned from the audience's choices of content over time, but nowadays, new technologies have the potential to provide information about the audience's reaction almost instantly. Although this is an area of debate, it might be considered that the audience itself 'manages the news' by maintaining or losing interest in a given subject [Vosoughi et al., 2018]. Any significant discrepancies between reality and what is portrayed by the media may then be attributed to the audience's interests. Therefore, the analysis of media content needs to take into account the collective behaviour, feedback cycles and relations between content creators and consumers. For this reason, the approach of complex systems that is taken throughout this thesis, can be beneficial to understand patterns in the amount of coverage given by the media to different events or topics.

The aim of this Chapter is to use different data analysis tools and modelling techniques to understand some of the patterns displayed by the amount of media coverage given to different locations over time. Firstly, the temporal evolution of the amount of media coverage given to a specific event is analysed. Secondly, the effect of population size on the amount of coverage that the media gives to urban areas of different population sizes is assessed. In order to do this, urban scaling models, which were described in Chapter 2, are used as the main modelling tool. The findings of this Chapter are based on real data from a particular case study regarding the

amount of media coverage before and after the occurrence of the Puebla earthquake, which hit Mexico on the 19th of September of 2017. The earthquake, hereafter referred to as S19, was mainly felt in the central part of Mexico, although it affected several urban areas of different population size. With a magnitude of 7.1 Mw, the S19 event was in the media spotlight for several weeks.

It should be noted here that none of the authors in [Prieto Curiel et al., 2019] are experts in the field of media studies. Consequently, the reader should keep in mind that the results from this Chapter do not belong to this field of research. Instead, the results should be regarded as an application of urban scaling models and other data analysis tools to a specific data set.

3.1.1 Temporal media coverage

In terms of the temporal aspects, the public attention given to various cultural products follows a consistent pattern over time [Candia et al., 2018; Coman, 2018]. In particular, the attention that the audience places on any specific event or topic covered by the media can be categorised into four stages: the pre-problem stage, a discovery stage, a stage of gradual decline in attention and a post-problem stage [Downs, 1972]. This attention cycle is closely related to the coverage that the media gives to news related to climate change [McDonald, 2009]; the rise and fall of the anti-nuclear movement [Joppke, 1991]; terrorism and travel safety [Hall, 2002]; and organisational changes in the structures of government [Peters and Hogwood, 1985], among other examples.

However, the analysis of temporal aspects of the coverage given by the media is challenging, as it often relates to topics, such as terrorism [Petersen, 2009] or climate change [McDonald, 2009], rather than specific events. There are studies about the evolution of the coverage of events on online platforms [Wu and Huberman, 2007] and social media [Weng

et al., 2012; Yang and Leskovec, 2011]. But this popularity, instead of being measured as the amount of coverage given by the news providers, is seen from the point of view of the readers' reaction, through the 'upvotes', 'shares of memes' or 'hashtags'.

The temporal distribution of the amount of media coverage is studied here in the case of the S19 event, whose consequences were severe enough to capture the audience's and the media's attention for a long period. The fact that the S19 event stayed on the spotlight for several weeks allows us to discern some structure in the evolution of the media coverage.

3.1.2 Spatial media coverage

Interest in a particular event might reduce as the physical distance or, in a more generalised way, any cultural or social 'distance' from the event increases [Tobler, 1970]. It is then expected that most of the coverage given by the media is mainly focused on activities and events which are closer to where their audience happens to be. Media can thus help to uncover the strength of the relations between different regions [Yuan, Y. and Liu, Y. and Wei, G., 2017]. But, what is the amount of media coverage given to events which are at a similar distance? Are certain locations more newsworthy?

In order to answer this question, the media coverage given to various events occurring in different locations can be compared. However, this analysis might reveal differences between the events themselves rather than the social behaviour viewed through the lens of the media. Instead of considering several events, the idea here is to investigate the behaviour of the media before and after a single event: the S19 earthquake. Since this thesis is particularly concerned with the consequences of urban population size, the analysis will consist in applying the modelling framework provided by urban scaling models, discussed in section 2.2.2 from Chapter 2.

3.2 Data and methods

A common way to analyse the focus of media is to measure the area in printed newspapers and magazines (or the minutes on radio or TV) devoted to a particular story. Using this approach, for instance, it was discovered that newspapers concentrate on crimes which have a sexual or a violent component [Ditton and Duffy, 1983; Ditton et al., 2004]. In particular, it was found that in the USA, 30% of crime stories in the newspapers are about homicides but in reality, only 0.02% of all crimes suffered involve a murder [Liska and Baccaglini, 1990]. It has also been shown that the amount of media coverage to climate change is higher in carbon dependent countries with commitments under the Kyoto Protocol [Schmidt et al., 2013]. By using the same approach, a decay in the coverage of terrorism news given by a newspaper between 2000 and 2007 was detected [Petersen, 2009].

Over recent years, however, many traditional media outlets such as newspapers or magazines have dramatically changed their methods of delivery. For example, now, less than 14% of The Guardian's audience actually reads the printed edition of that newspaper [PressGazette, 2017]. The main media channels with an informative focus, have moved to online streams and hence, the ways to analyse the focus of the media have also changed. Therefore, one possible strategy is to count the number of different URLs published by e.g. a newspaper. To help with this analysis, there are resources for searching links published by different online media outlets, such as the GDELT Project [Leetaru and Schrodt, 2013], so that outputs from different sites can be analysed to reveal items related to different topics, for example, to climate change [Olteanu et al., 2015]. However, during major events, different URLs are not necessarily associated to different news items; for instance, an entry corresponding to one

particular URL can be updated several times a day as the events unfold. It then becomes hard to measure the media coverage based solely on the number of URLs. Furthermore, identifying all the URLs published by a website can also be challenging, as they are usually non-trivially divided into sections, subsections, columns and more.

Due to the ease of access to Twitter data, the strategy followed here is to take the tweets published by several newspapers' Twitter accounts about a particular subject as a proxy for the coverage of the subject given by those newspapers. Twitter data has been used before [Amato et al., 2017; Kounadi et al., 2015; Pak and Paroubek, 2010] to detect flows of national politics [Ausserhofer and Maireder, 2013], crime hotspots [Malleson and Andresen, 2015], exposure to cross-ideological content [Himelboim et al., 2013], activism [Xu et al., 2014], etc. and to manage environmental natural hazards or diseases [Cvetojevic, S. and Hochmair, H. H., 2018].

This approach requires an assumption: every update to the corresponding newspaper's website is usually announced via a related post on the associated Twitter account. If several tweets regarding the same issue are posted by the same account, this indicates the newspaper considers it to be important. If there are no tweets about a certain piece of news, it is likely to be less relevant for the newspaper. The reasoning supporting this assumption is that, with little effort (compared to producing the news update), mass media outlets can reach a broader audience by promoting their updates on their social media accounts. Online newspapers have then an incentive to promote their contents on Twitter. It is also assumed that, by taking into account the online newspapers with largest audiences, a significant proportion of the national written news output is being considered. The online newspapers selected for this work have different approaches (commercial and public) and target audiences, so a further assumption can be made that they are representative of the overall written news activity.

3.2.1 Identifying media channels

There are a large number of mass media outlets, but only 19 are taken into account in this study. These are the most popular Mexican online newspapers. Their popularity is measured as the size of their audience. These newspapers are assumed to be representative of the overall news activity produced by the national mass media.

The coverage given by the 19 most popular Mexican online newspapers is measured through their Twitter activity, which is taken as a proxy for the newspapers' contents. The newspapers' Twitter accounts are then sorted in decreasing order of popularity according to their number of followers. For account A_i, with $i = 1, 2, ..., n$ and $n = 19$, the number of Twitter followers is denoted as F_i. In total, the 19 accounts have 41.3 million followers, although it should be noted that a Twitter user might follow more than one newspaper, there being only around 9 million Twitter users in Mexico (whose population is close to 130 million) in 2017.

It is important to note that there is some variability among the chosen newspapers, all gathered in Table 3.1: some have a commercial purpose, while others focus more on informing citizens. For example, *Televisa* is part of a large consortium which also has many TV channels, and *OnceTV* is broadcast by a University with subsidised resources. However, the analysis in the following sections shows that the similar trends are observed in all the outlets, with similar patterns of decay and scaling parameters. Thus, the results are general, rather than specific for only one particular type of outlet.

3.2.2 Collecting tweets from selected accounts

In order to collect the tweets, the Twitter API 'Get Tweet timelines' [Twitter, 2017] is used. Although 'Get Tweet timelines' limits queries to only

Table 3.1: Choice of newspapers, sorted by descending number of followers on Twitter at the time of the analysis (2017).

	Newspaper	**Twitter account**	**Type**	**Followers (thousands)**
1	Aristegui Noticias	@AristeguiOnline	commercial	7,742
2	El Universal	@El_Universal_Mx	commercial	4,668
3	Proceso	@revistaproceso	commercial	4,595
4	Milenio	@Milenio	commercial	4,018
5	Sopitas	@sopitas	commercial	2,773
6	Televisa	@NTelevisa_com	commercial	2,710
7	Reforma	@Reforma	commercial	2,460
8	La Jornada	@lajornadaonline	commercial	2,022
9	Animal Político	@Pajaropolitico	commercial	1,722
10	Excelsior	@Excelsior	commercial	1,505
11	Sin Embargo	@SinEmbargoMX	commercial	1,145
12	Pictoline	@pictoline	commercial	1,097
13	El Financiero	@ElFinanciero_Mx	commercial	1,016
14	Publimetro	@PublimetroMX	commercial	689
15	El Economista	@eleconomista	commercial	350
16	UnoTV	@UnoNoticias	commercial	321
17	SDP	@sdpnoticias	commercial	319
18	La Silla Rota	@la_sillarota	commercial	287
19	Once Noticias	@OnceNoticiasTV	public	68

the 3,200 most recent tweets from one account, in this case, the retrieval of tweets is carried out periodically, so tweets corresponding to a sufficiently long period are gathered.

To establish a baseline, 18,083 tweets are collected for a full week before the S19 earthquake. Since there was an earthquake of smaller magnitude on the 7th of September, the baseline week is selected to be that between the 29th August and the 4th September 2017. Then, 73,389 tweets are collected between the 19th September and the 17th October 2017, corresponding to the four weeks starting from the day of the S19 earthquake. In total, 91,472 tweets are collected from the 19 accounts.

3.2.3 Urban areas in the media

For the purposes of this analysis, the 15 Mexican urban areas in Table 3.2 are considered. These urban areas are selected based on two different criteria: their population has to be more than 20,000 and the local measure of intensity for the S19 earthquake has to be moderate or higher using the Modified Mercalli Intensity scale. On this scale, moderate corresponds to grade 5. The Modified Mercalli Intensity scale quantifies the local measure of intensity according to the response from people and buildings rather than the raw excitation force. Hence, the selected urban areas are not necessarily the closest ones to the epicentre (according to the US Geological Service Report, the S19 earthquake, USGS Event ID: us2000ar20, was centred 1 km ESE of Ayutla), but the ones that were most affected. The map in Figure 3.1 depicts the location of the urban areas considered here, the epicentre of the earthquake and the contours corresponding to different grades on the Modified Mercalli Intensity scale.

If these urban areas are included in the list of Mexican metropolitan zones (see section 2.1.1 in Chapter 2 for detailed definition), their population is set to be that of the corresponding MZ. Otherwise, the population is set to be that of the corresponding municipality, established by the National Institute of Statistics and Geography (INEGI) [Instituto Nacional de Estadística y Geografía, 2015]. Even though there are yearly population projections for the MZs, the population of the *municipios* is only published by INEGI every five years, either as part of the census, or as part of the Intercensal Survey. The last published data available for all the urban areas corresponds to the year 2015.

For each urban area, the population size is denoted by X_j, where now $j = 1, 2, ...m$ and $m = 15$. The urban areas are ordered by descending population size so that $X_1 = 21,275,109$ corresponds to the metropolitan

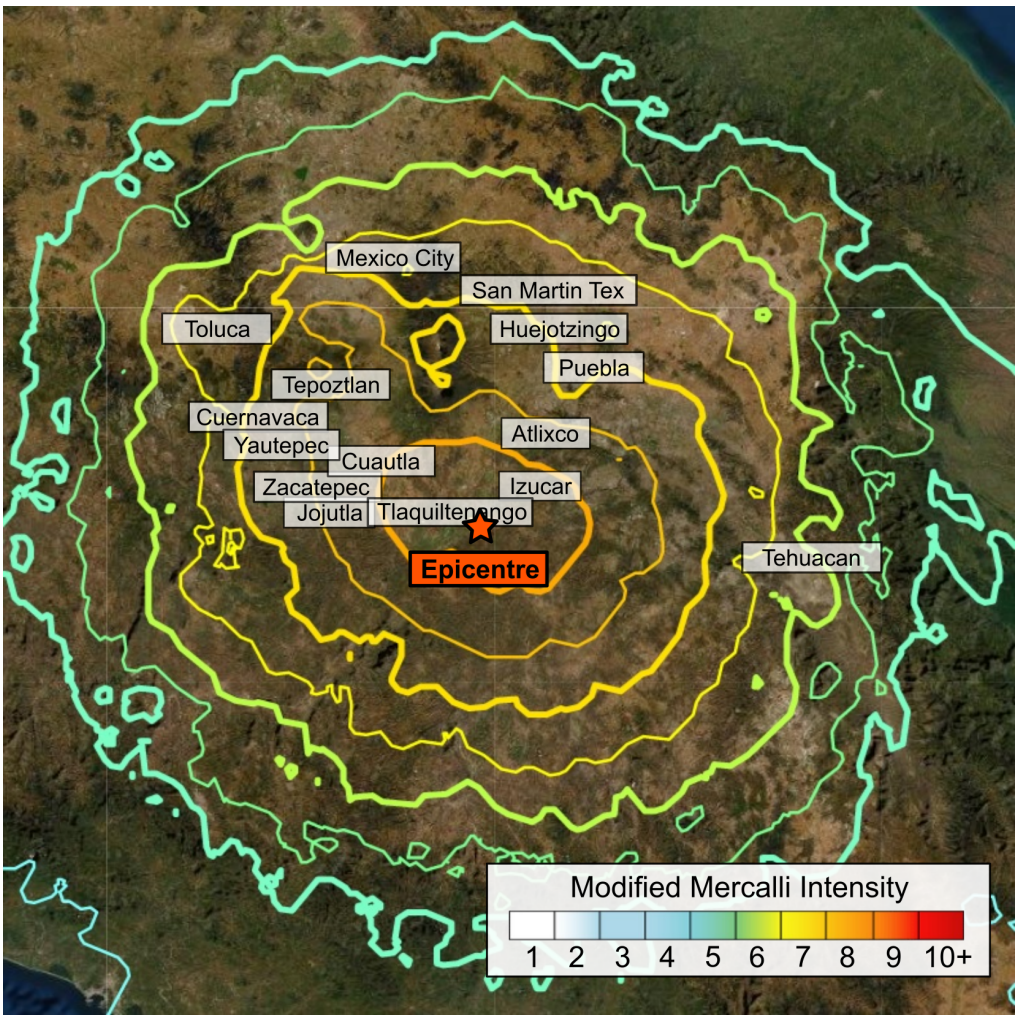


Figure 3.1: Map of the region affected by the S19 earthquake. The map shows the location of the urban areas considered in the analysis, the epicentre of the earthquake and the contours corresponding to different grades on the Modified Mercalli Intensity scale.

zone of Mexico City and $X_{15} = 33,844$ corresponds to the municipality of Tlaquiltenango, in the state of Morelos.

It is important to note that even though Twitter penetration can be higher in bigger cities, the results presented here are not invalidated by this observation. Assuming that the Twitter content is similar to the content in other platforms belonging to the same newspaper, if media channels decide

to publish more news items about the big urban areas due to the higher number of followers there, then this is considered to be part of the scaling effects [Bokányi et al., 2019].

Table 3.2: Choice of urban areas, sorted by descending population size.

Name	State	Population	Metropolitan zone
Mexico City (CDMX)	Mexico City	21,275,109	YES
Puebla	Puebla	2,994,147	YES
Toluca	Estado Mexico	2,207,581	YES
Cuernavaca	Morelos	1,003,174	YES
Cuautla	Morelos	476,170	YES
Tehuacán	Puebla	340,644	YES
San Martín Texmelucán	Puebla	152,051	NO
Atlixco	Puebla	134,364	NO
Yautepec	Morelos	102,690	NO
Izúcar	Puebla	77,601	NO
Huejotzingo	Puebla	73,771	NO
Jojutla	Morelos	57,121	NO
Tepoztlán	Morelos	46,946	NO
Zacatepec	Morelos	36,159	NO
Tlaquiltenango	Morelos	33,844	NO

3.2.4 Classifying a tweet as being related to a specific urban area or to the earthquake

The collected tweets can be classified [Dodds et al., 2011] as being related to the earthquake, an urban area, or both, depending on whether or not they contain terms associated with these subjects. To do this, two lists of terms are created. One contains terms related to the earthquake, the other list contains terms related to the urban areas, such as their names, names of city mayors and names of significant places or buildings. This last aspect is particularly important in the classification of news items related to Mexico City, as tweets often refer to specific locations within the urban area. The two lists can be found in Tables 3.3 and 3.4.

Table 3.3: List of earthquake-related terms.

Terms				
19S	CDMX	Emergencia	Magnitud	Riesgo
19sverificado	Chiapas	Epicentro	Mexicoestadepie	Roma
7S	Colapsadas	Escombros	MexicoUnido	S19
Acopio	Colapsados	Fallecidos	Morelos	S7
Afectadas	Colapso	Fonaden	Obregon	Sacude
Afectados	Colegio	Fonden	Oscilatorio	Sismica
Albergue	Condesa	Frida	Polanco	Sismo
Alvaro	Damnificados	FridaSofia	PrayForMexico	Terremoto
AlvaroObregon	Dañada	FuerzaCDMX	Puebla	Totolapan
AO	Daño	FuerzaMexico	Rebsamen	Trepidatorio
Apoyo	Demolicion	FuerzaMorelos	Reconstruccion	Verificado
Ayuda	Derrumbe	FuerzaPuebla	Reconstruir	Viveres
AyudaCDMX	Donaciones	Guerrero	Rescate	Viviendas
Brigadistas	Eartquake	Jojutla	Rescatistas	Voluntarios

Table 3.4: List of terms related to urban areas.

Term	Urban area	Term	Urban area
Atlixco	Atlixco	Metro	CDMX
Condesa	CMDX	Venustiano	CDMX
Roma	CDMX	Carranza	CDMX
Polanco	CDMX	Hiram	CDMX
Rebsamen	CDMX	Almeida	CDMX
Frida	CDMX	Cuautla	Cuautla
AO286	CDMX	Cuernavaca	Cuernavaca
Alvaro	CDMX	Huejotzingo	Huejotzingo
Obregon	CDMX	Izucar	Izucar
CDMX	CDMX	Jojutla	Jojutla
Mancera	CDMX	Puebla	Puebla
C5	CDMX	RMV	Puebla
Patricio	CDMX	Rafael	Puebla
Sanz	CDMX	Moreno	Puebla
Valle	CDMX	Cholula	Puebla
Delegacion(es)	CDMX	San Martin Tex.	San Martin Tex.
Benito Juarez	CDMX	Tehuacan	Tehuacan
Delegacion BJ	CDMX	Tepoztlan	Tepoztlan
Cuauhtemoc	CDMX	Tlaquiltenango	Tlaquiltenango
Monreal	CDMX	Toluca	Toluca
Aeropuerto	CDMX	Yautepec	Yautepec
AICM	CDMX	Zacatepec	Zacatepec

As these lists of predefined terms may not be exhaustive, the tweets that do not match any of the terms are manually examined and classified a posteriori. This ensures no false negatives in the classification process. The false positives are not controlled, but their number may be reduced due to the context in which the tweets are generated: it is assumed that a tweet matching 'earthquake' or 'reconstruction' during the week after an earthquake is related to the natural disaster, particularly since only tweets posted by newspapers are analysed. From the 73,389 tweets posted during the four-week period after the earthquake, more than one-third contained information related to the earthquake.

None of the tweets that are used for the analysis share their location, therefore, tweet classification is purely made by checking if the tweets contain any terms related to any of the urban areas. If instead, individual users who felt the earthquake had been considered, their location would have been relevant. For example, in a study that classified the media users according to their behaviour [Zelenkauskaite and Balduccini, 2017], their IP addresses were used to identify the location.

3.2.5 Media coverage index (MCI)

In order to quantify the amount of media coverage given to each of the 15 urban areas by the selected Twitter accounts, the media coverage index (MCI) is defined. Let T_i be the number of tweets posted by Twitter account A_i, with $i = 1, ..., n$. Let T_{ij} be the number of tweets posted by account A_i with a reference to urban area j, with $j = 1, ..., m$. When the goal is to quantify the media coverage given to an urban area because of a certain topic, like the earthquake in this case, then T_{ij} must be taken as the number of tweets published by account A_i which contain a reference to both urban area j and the topic in question.

Some accounts post more often than others, although there tends to be an overlap in the content of the tweets when they are very frequent. In order to counter this bias, the proportion tweets T_{ij}/T_i related to urban area j (and possibly a topic) out of all the tweets T_i posted by account A_i is considered. The MCI Y_j for urban area j is then defined as the weighted average of the quantity T_{ij}/T_i over the index i

$$Y_j = \frac{1}{F^{(n)}} \sum_{i=1}^{n} F_i \frac{T_{ij}}{T_i}. \tag{3.1}$$

where the weights are determined by the number of followers F_i of each account.

In equation (3.1), the quantity $F^{(n)} = \sum_{i=1}^{n} F_i$ and therefore, the factor $1/F^{(n)}$ simply rescales the MCI so it is dimensionless and lies between 0 and 1. Although the focus here is on the main Mexican newspapers' Twitter accounts, the definition of the MCI is general and could be used for other countries and other types of media. Furthermore, the value of the MCI is stable with respect to the number of Twitter accounts under consideration, in the sense that its value only experiences small changes when the number of Twitter accounts n is altered. The stability of the MCI is shown below.

Stability of the media coverage index

Let $Y_j(n)$ be the MCI for urban area j, where the number n of Twitter accounts under consideration is now made explicit. This can be rewritten in terms of $Y_j(n-1)$:

$$Y_j(n) = \frac{1}{F^{(n)}} \sum_{i=1}^{n} F_i \frac{T_{ij}}{T_i} = \frac{F^{(n-1)}}{F^{(n)}} Y_j(n-1) + \frac{F_n}{F^{(n)}} \frac{T_{nj}}{T_n}. \tag{3.2}$$

The MCI will be stable with respect to the number of Twitter accounts n if $Y_j(n-1)$ does not differ significantly from $Y_j(n)$.

In the extreme case when the account A_n does not post any tweets related to urban area j, $T_{nj} = 0$ and therefore, $Y_j(n) = \frac{F^{(n-1)}}{F^{(n)}} Y_j(n-1)$. The value of $Y_j(n)$ is then bounded from below by $\frac{F^{(n-1)}}{F^{(n)}} Y_j(n-1)$. The other extreme case takes place when all the tweets posted by account A_n refer to urban area j, then $T_{nj}/T_n = 1$. This results in

$$\frac{F^{(n-1)}}{F^{(n)}} Y_j(n-1) \leq Y_j(n) \leq \frac{F^{(n-1)}}{F^{(n)}} Y_j(n-1) + \frac{F_n}{F^{(n)}}. \tag{3.3}$$

Since $F^{(n)} = F^{(n-1)} + F_n$ and $F^{(n-1)} >> F_n$ when n is large enough, the quantity $\frac{F^{(n-1)}}{F^{(n)}}$ in equation (3.3) is close to 1, and $\frac{F_n}{F^{(n)}}$ is close to 0. The lower bound for $Y_j(n)$ is then $\frac{F^{(n-1)}}{F^{(n)}} Y_j(n-1) \approx Y_j(n-1)$.

For the upper bound, there is one realistic assumption that can be made: if A_n places all its attention on urban area j, then it is likely that all the other accounts $A_1, \ldots, A_{n-1}$ also give most of their attention to urban area j. Therefore, the term $\frac{F_n}{F^{(n)}}$ will be small compared to $\frac{F^{(n-1)}}{F^{(n)}} Y_j(n-1) \approx Y_j(n-1)$, as $Y_j(n-1)$ will be close to 1. Following this reasoning, the more accounts considered (so the bigger n is), the lesser the effect that A_n has on $Y_j(n)$ and the closer $Y_j(n)$ is to $Y_j(n-1)$.

To illustrate the stability, the MCI corresponding to Mexico City during the first week after the earthquake can be obtained from the 19 most followed Twitter accounts belonging to Mexican newspapers. The result is 0.594436. If the least popular account is removed, then the MCI has a value of 0.594441, which is a change of less than 0.001%. Even if the most popular from the 19 accounts was to be removed, the MCI would be 0.628648, and the change would be 5.442%.

3.3 Results

3.3.1 Time evolution of the media coverage

On a midweek day from the baseline week prior to the earthquake, the 19 media accounts posted from 2,584 to 3,042 tweets; on a weekend day from the same week, they published from 1,759 to 2,257 tweets. There are also fluctuations within a 24-hour period, with fewer tweets released at night. On the day immediately after the earthquake, the number of tweets increased to 3,316, and to 4,415 the following day. However, the number of posts returned to the usual values within three days as seen in Figure 3.2, which overlays the baseline week and the week immediately after the earthquake.

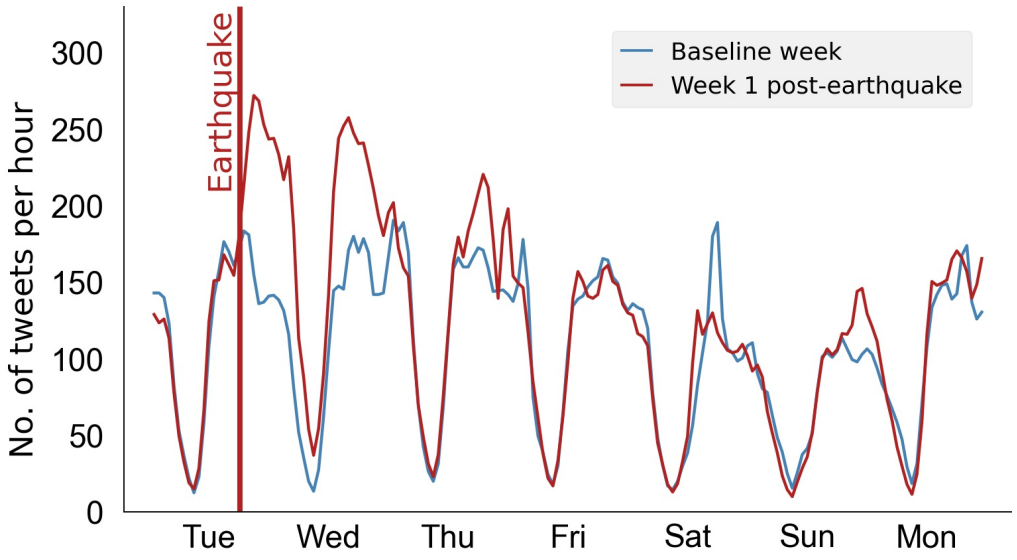


Figure 3.2: Evolution of the average number of tweets per hour published by the 19 selected online newspapers. Comparing a baseline week prior to the earthquake and the week starting from the day of the earthquake.

Figure 3.3 shows the average number of tweets per day published by the 19 selected online newspapers over the four-week period starting on the day of the earthquake. It also shows the average number of all the tweets per day published by the same accounts over the same period. The Figure suggests that almost immediately after the earthquake, a large proportion of tweets is related to the earthquake. However, a few days later, the media interest in the earthquake dies out.

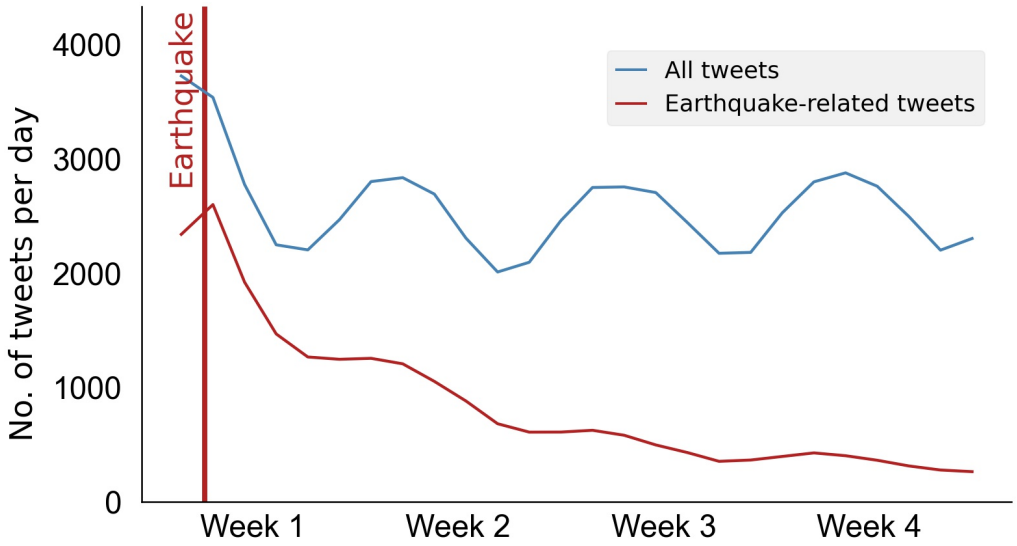


Figure 3.3: Evolution of the average number of tweets per day and earthquake-related tweets published by the 19 selected online newspapers over the four week-period starting on the earthquake day. A 2-day rolling average has been applied to smooth the curves.

To model the decay of the media interest in the earthquake, let $a(t)$ be the proportion of tweets devoted to the earthquake at time t, measured in days. Then, if $t = 0$ is set to be the moment when the earthquake struck, $t = 1$ would represent 24 hours after the earthquake. Hence, for $t < 0$, $a(t) = 0$. Then, $a(t)$ can be interpreted as the probability that a random tweet published at time t was related to the earthquake.

An assumption can be made such that for $t > 0$, $a(t)$ decreases exponentially:

$$a(t) = a_0 e^{-\psi t}. \tag{3.4}$$

The value a_0 represents the initial proportion of news devoted to the earthquake. The amount $e^{-\psi}$ represents the proportion of tweets related to the earthquake with respect to the previous day. The model parameters a_0 and ψ can be estimated from data using the method of non-linear least-squares. The estimated values are denoted $\hat{a}_0$ and $\hat{\psi}$, and their values are $\hat{a}_0 = 0.739 \pm 0.016$ and $\hat{\psi} = 0.089 \pm 0.003$, were the errors correspond to one standard deviation. The curve corresponding to this model is included in Figure 3.4. Hence, according to the exponential model in equation (3.4), the initial proportion of tweets related to the earthquake is about 74% and the number of tweets related to the earthquake on a given day is about 91% of the day before.

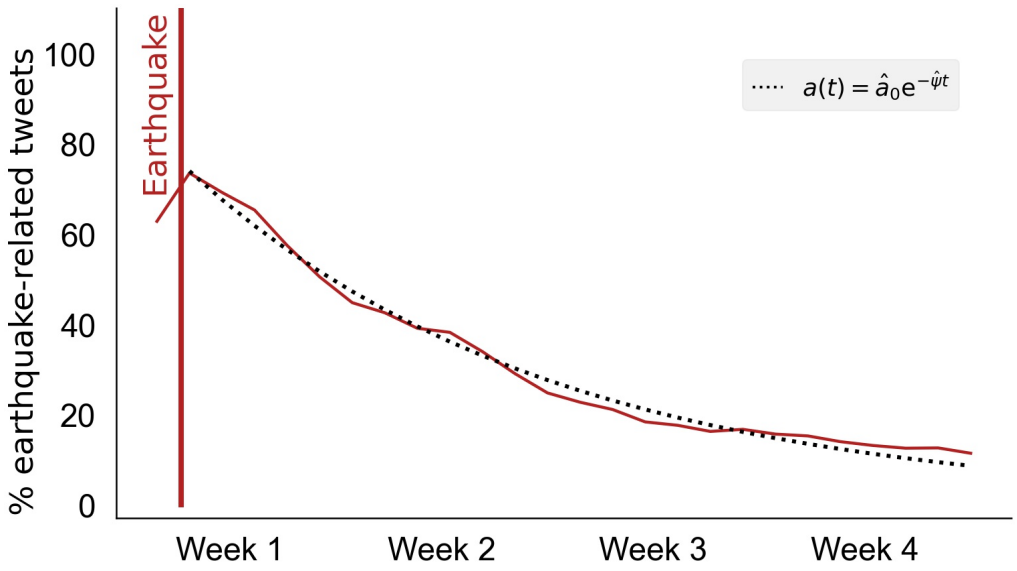


Figure 3.4: Evolution of the average percentage of tweets per day that mention the earthquake, published by the 19 selected online newspapers over the four-week period starting on the day of the earthquake.

3.3.2 Location of the spotlight

During the baseline week, for every tweet that mentions an urban area with around 3 million inhabitants, 40 tweets are found related to Mexico City, which is just above 7 times this size. Therefore, on a per capita basis, the number of tweets that refer to Mexico City is 5.6 times larger than the number of tweets that refer to an urban area with 3 million people. This suggests that the amount of media coverage across urban areas follows superlinear growth with respect to population size. However, just after the earthquake, smaller urban areas are more frequently mentioned in the tweets published by the online newspapers considered here.

To formally model the relation between the amount of media coverage given to the Mexican urban areas and their population size, an urban scaling model of the following type is assumed

$$Y = \alpha X^{\beta}, \tag{3.5}$$

where in this case, Y is the MCI corresponding to an urban area, and X is the population size of each urban area. The proportionality constant $\alpha > 0$ and the scaling exponent $\beta > 0$ are the model parameters. The value of these parameters can be estimated from the data and denoted by $\hat{\alpha}$ and $\hat{\beta}$ respectively. If $\hat{\beta} > 1$, the MCI increases superlinearly with population size, while if $\hat{\beta} < 1$, the MCI increases sublinearly. If the data reveals a value of $\hat{\beta}$ close to 1, then the MCI is not affected by the population size.

The scaling model parameters can be estimated as those that best fit the data according to the method of non-linear least squares. The estimated scaling parameters can be found in Table 3.5, corresponding to data from the baseline week, the first week starting on the day of the earthquake and the week 4 after the earthquake. The error for each estimated parameter corresponds to one standard deviation.

Table 3.5: Estimated values of the scaling model parameters for data from the baseline week and weeks 1 and 4 after the earthquake.

	Scaling constant $\hat{\alpha}$	**Scaling exponent** $\hat{\beta}$
Baseline week	$(7.555 \pm 0.008) \times 10^{-15}$	1.7 ± 0.02
Week 1	$(0.753 \pm 1.341) \times 10^{-12}$	1.60 ± 0.10
Week 4	$(3.069 \pm 0.022) \times 10^{-16}$	1.77 ± 0.09

Data for the coverage placed on each urban area during a baseline week before the earthquake gives a scaling exponent of $\hat{\beta}_{W0} = 1.79 \pm 0.02$, hence implying a superlinear scaling relation between the media coverage index and the population size (see Figure 3.5). For the data corresponding to the first week after the earthquake, the parameter is found to be $\hat{\beta}_{W1} = 1.60 \pm 0.10$. Therefore, although the scaling behaviour remains to be superlinear, it is not as strong as that observed during the baseline week. The scaling relation is shown to return to its baseline behaviour by week four after the earthquake, when $\hat{\beta}_{W4} = 1.77 \pm 0.09$. The values of the parameters have been estimated via the method of non-linear least squares.

3.4 Conclusions

It has been found that both the producers and consumers of the most popular online Mexican newspapers follow emergent collective behaviours. On the one hand, the amount of coverage that they give to an earthquake event that affected a range of Mexican urban areas, decays over time. On the other hand, the amount of coverage given to particular urban areas depends on their population size, with this dependency being different before and after the earthquake event.

To quantitatively measure the media coverage given to an urban area and/or a specific subject, a media coverage index has been defined. Here, the metric is applied to the 19 most popular online newspapers in Mexico,

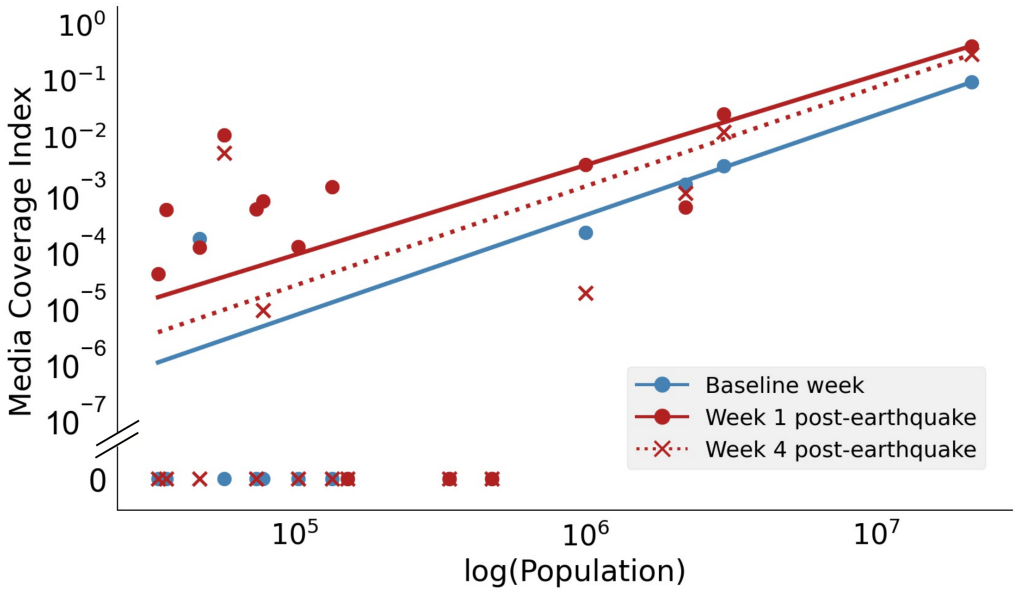


Figure 3.5: Media coverage index for urban areas of different size, computed for the baseline week, the week starting on the day of the earthquake and on the fourth week after the earthquake. Both axis are in logarithmic scale. During a typical pre-earthquake week, the media coverage index per person is estimated to scale superlinearly, meaning that large urban areas receive more media coverage per person. However, during the immediate post-earthquake period, it is observed that the superlinear scaling relation is not as strong, since the media coverage given to smaller urban areas increases. Four weeks after the earthquake, the estimated scaling exponent has almost returned to its baseline value.

with their popularity being determined by the size of their audience. In particular, the amount of coverage given by each outlet is measured by counting the number of tweets related to the subject that are posted via the Twitter accounts corresponding to the newspapers. The assumption is that what they decide to post on their Twitter accounts is a reflection of what they publish on their websites. The media coverage index places different weights to each Twitter account according to the number of followers, which is assumed to reflect the size of each newspaper's audience.

3.4.1 Rapid decay of the media coverage

Assuming an exponential model for the decay of the media coverage over time, it is estimated that the amount of coverage given to the earthquake on a given day is 91% of that corresponding to the previous day. Similar to cultural products, such as music or academic papers [Candia et al., 2018; Coman, 2018], media coverage has its own distinctive decay.

Although there might be other specific aspects which determine how much coverage is given to subjects other than the earthquake, a similar behaviour is expected for the evolution of this coverage. There is an initial peak in the amount of coverage, but as time passes, the coverage decreases exponentially. The rate of decay, denoted by ψ, may depend on factors such as geographical proximity or whether sports, stars or celebrities are involved. The earthquake was a particularly significant event and therefore, there was a large amount of coverage post-earthquake. It is likely that other events start with a lower initial peak.

3.4.2 Coverage of the media is not homogeneously distributed

The earthquake in Mexico provided a special event to quantitatively measure the spatial distribution of the media coverage. The earthquake affected a large region, including a number of urban areas and yet, the media spotlight on different locations varied.

Results show that in the absence of the earthquake there is usually a superlinear relation between the amount of media coverage (measured by the media coverage index) given to urban areas and their population size. Before the earthquake, the superlinear scaling parameter of $\beta = 1.7$ is well above 1, meaning that large urban are seemingly more newsworthy than the smaller ones. Immediately after the earthquake, the scaling is

still superlinear, but the effect is not as strong. This implies that small urban areas such as Jojutla, which previously were rarely mentioned in the news, received considerably more attention. This change in the spatial distribution of the media coverage is only observed in the short-term: as the general interest in the earthquake decays, the spatial distribution of the media coverage also comes back to its usual form, i.e. four weeks after the earthquake the superlinear scaling behaviour of the media coverage with respect to population size returns to its baseline value.

3.4.3 Limitations of the current study and ideas for future work

In this work, tweets are classified as being related to an urban area and/or the earthquake by performing successive automated look-ups of the terms included in the lists from Tables 3.3 and 3.4. It is relatively quick to manually check that this method of classification is very accurate, with very few amendments needed. However, if a bigger database was to be used, more advanced language processing methods would be needed for the classification.

One limitation regarding the definition of the media coverage index is that it does not take into account the impact of the event under study on the different urban areas. For example, in the case of the earthquake, Tlaquiltenango scored an 8 on the Modified Mercalli Intensity whereas Tehuacán scored a 6. An improved version of the media coverage index would normalise the amount of media attention with respect to the impact of the earthquake event at each location, but this is left for future work.

It should be noted that, as discussed later in this thesis, extremely populous urban areas, sometimes referred to as dragon kings [Sornette, 2009], can be considered as statistical outliers due to their special socioeconomic

role in the urban dynamics of a country. Hence, they could be affecting the underlying scaling models followed by the rest of urban areas under consideration [Arcaute et al., 2015]. However, in this case, the results after the earthquake are compared against the baseline week before the earthquake, so it is possible to see how the media coverage is distributed across cities also before the event.

The trends detected here, based on the analysis of the Twitter accounts belonging to online newspapers, are likely to be reflected also in traditional printed media and other news outlets, including those with a regional or global outlook. The methodology suggested here can be applied to other regions of the world and to other equally suitable events.

Chapter 4

Relationship between road accidents and population size in built-up areas from England and Wales, France and Spain

4.1 Introduction

The work presented in this Chapter has been partially covered in the research paper entitled 'Uncovering the behaviour of road accidents in urban areas' [Cabrera-Arnau et al., 2020].

On the 17th of August of 1896, Bridget Driscoll was at Crystal Palace, in South-East London, attending a religious event. Unfortunately, Mrs Driscoll made history that day for being the first person to die as a consequence of a car crash in Great Britain. The driver, Arthur Edsall, claimed to be driving his Roger-Benz at only 4 miles per hour when he struck Mrs Driscoll, and that he had rung his bell and shouted. However, he had started driving only three weeks before, there was no license requirement

at that time and he had been given no instruction as to which side of the road to keep to [BBC, 2010].

Since 1896, we have come a long way. The application of laws in line with the so-called 'best practice' on behavioural risk factors —such as speed, drink-driving and failing to use motorcycle helmets, seat-belts and child restraints— have positioned the UK among the top five countries with regards to the lowest mortality rates on the roads. With the aim of reducing to zero the number of people killed or seriously injured on the roads, many countries have now incorporated Vision Zero [Vision Zero, 2020], a safety strategy initiated in Sweden in the 90s in the framework of the Safe System approach.

The promising trends towards this goal are displayed on Figure 4.1, where the number of deaths per capita has been plotted for France, Great Britain and Spain, which are the regions explored in this Chapter, as well as Germany, which will be also studied in Chapter 6. The data to generate this Figure is available for download from the year 1994 until 2017 [OECD, 2020]. In the case of Spain, only data until 2015 can be found. It can be seen that the incorporation of new road safety measures, such as fines for using mobile phones while driving or the introduction of a penalty point system for the driving record, are usually accompanied by drops in the number of deaths due to road accidents. It should be mentioned that Great Britain's remarkably low numbers of deaths, compared to other countries, over the time period displayed in Figure 4.1 are perhaps the result of the country's early adoption of road safety rules (e.g. the penalty points system was in place before 1994). For all countries, the period spanning from 2007 to 2010 also saw unusual reductions almost entirely due to the economic recession [Parliamentary Advisory Council for Transport Safety, 2019].

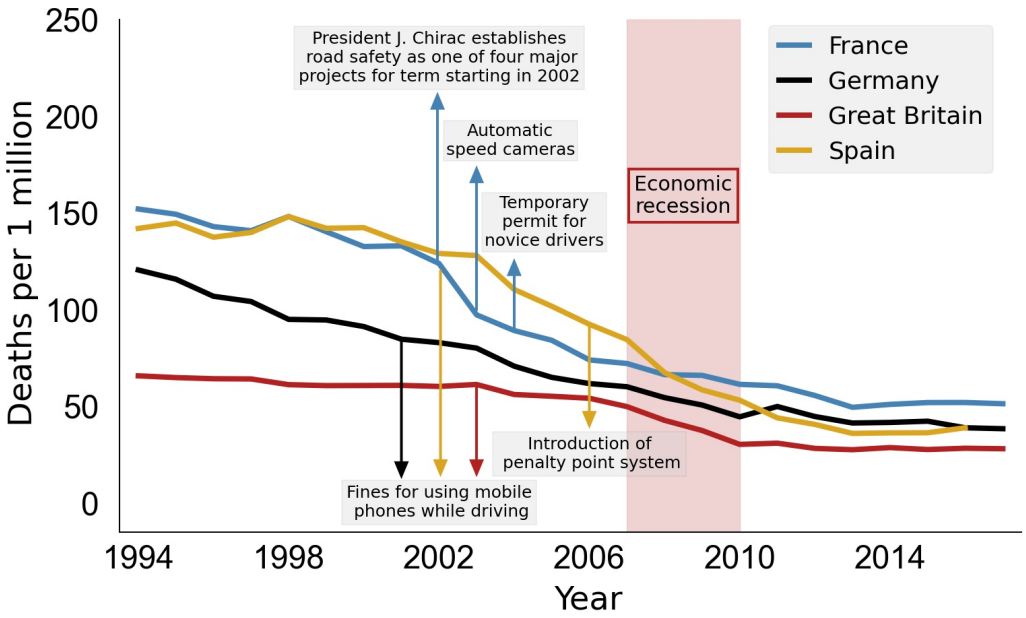


Figure 4.1: Evolution of the number of deaths per 1 million inhabitants due to road accidents, for the France, Germany, Great Britain and Spain.

The trends observed in Figure 4.1 are, in reality, the result of the interaction between many factors such as the population size of a given location, its road infrastructure, its economy, the average quality of the vehicles, the vehicle usage and people's driving attitude. The diagram in Figure 4.2, based on [Alirezaei et al., 2017], is a depiction of the non-trivial causal relationships between all of these factors. This type of representation, known as causal loop diagram, is typically used in order to understand the non-linear behaviour of complex systems. By convention, positive and negative reinforcement interactions are symbolised with the $+$ and $-$ signs respectively. For example, as seen in Figure 4.2, a good economy allows for the road infrastructure to be kept in good condition, and so, the number of accidents can also be kept to a lower level; however, a good economy also promotes more vehicle usage, which might result in excessive use of the infrastructure, increasing then the risk of accident.

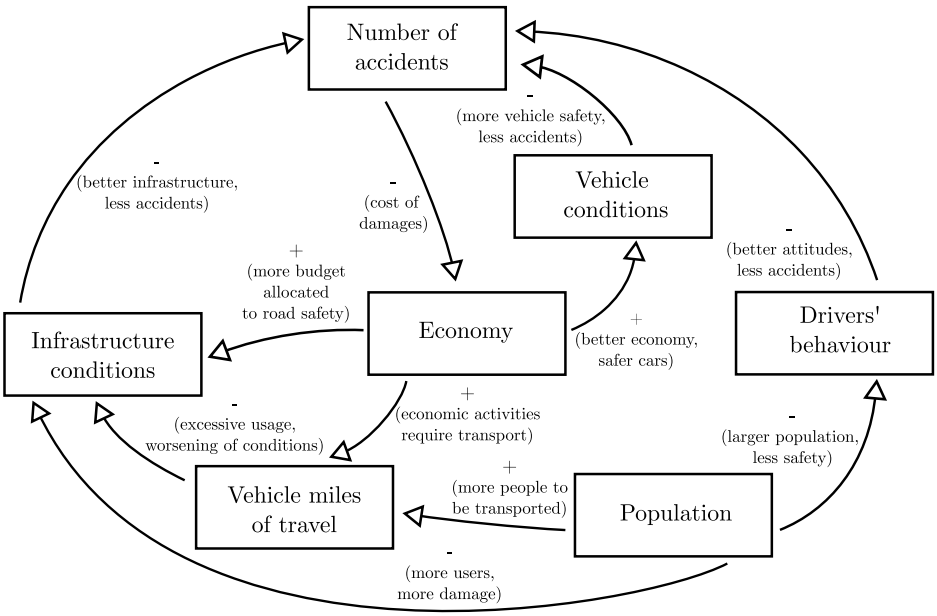


Figure 4.2: Causal loop diagram for the main factors involved in the number of accidents in a given population.

The number of accidents in a given location can then be considered a result of the complex interactions between many elements, as was also the case for the media coverage, discussed in Chapter 3. These interactions are really much more intricate than even suggested by Figure 4.2, which has been narrowed down to the 'economy - population - road safety' nexus. A complex systems approach is taken here, since it allows to capture the high level of integration between the different elements that influence the number of road accidents. Particularly, as in Chapter 3, urban scaling models are used to describe the relation between the population and the number of accidents occurring in urban areas. This modelling framework, which was introduced in section 2.2.2 of Chapter 2, is applied to analyse data corresponding to road accidents of different degrees of severity occurring in urban areas from three regions: England and Wales, France and Spain. Urban scaling models are also used here to derive an expression

for the probability of suffering accidents of a certain degree of severity in an urban area, given the population size of this urban area. The urban areas considered in this Chapter are defined based on a land-use criterion. The full description of these urban areas for the three regions of interest can be found in section 2.1.2 of Chapter 2. In this Chapter, the temporal distribution of the road accidents in urban and rural areas is also explored.

4.1.1 Road accidents: the current picture

Despite the encouraging trends observed in some countries, the Global Status Report on Road Safety 2018 published by the World Health Organization states that, currently, more than 1.35 million people die each year on the World's roads [World Health Organization, 2018].

The number of fatalities relative to the size of the World's population has stabilised in recent years, however, as shown in Figure 4.3 road accidents are still the leading cause for deaths among young people aged between 5 and 29, and the eighth cause for all the age groups, above HIV/AIDS, tuberculosis and diarrhoeal diseases. In addition to deaths on the roads, about 50 million people suffer non-fatal road injuries as well as other indirect health consequences each year [World Health Organization, 2018]. Road accidents pose a serious problem for the economy, especially in low- and middle-income countries, where the death rates due to road injuries are three times higher than in high-income countries.

The number of motor vehicles in the world has risen from 0.85 billion in the year 2000 to 2.1 billion in 2016 [World Health Organization, 2018], leading to an increased exposure to traffic for most people. This motorisation has grown hand in hand with urbanisation [Population Division of the UN Department of Economic and Social Affairs, 2018]. The quantitative understanding of the effects that urban organisation and dynamics have on road accidents is thus key for a successful transition to sustainability.

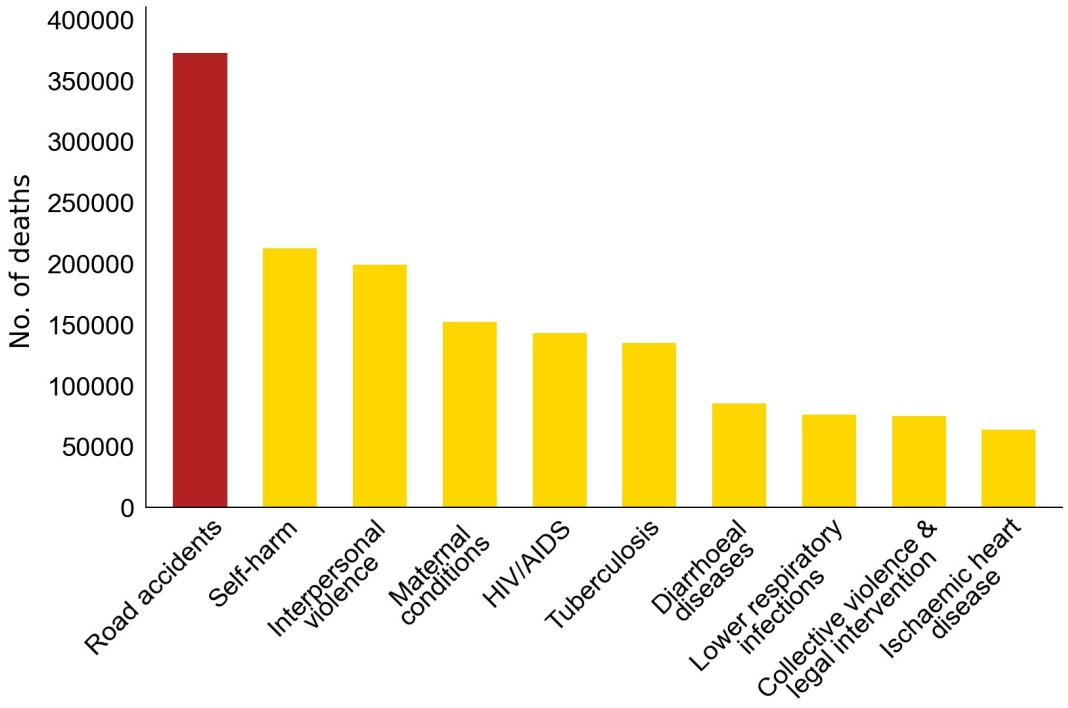


Figure 4.3: Leading causes of death globally among young people aged between 5 and 29, with road accidents being the first. Data retrieved from [World Health Organization, 2018].

Here, the focus is on the temporal and spatial patterns in the frequency of road accidents of different degrees of severity. Specifically, the interest is in understanding the weekly fluctuations of road accidents in urban areas and how they compare to the fluctuations in rural areas. Among the urban areas, it is investigated whether their population size plays a role in determining the number of accidents. Similarly, a model for the likelihood of suffering an accident in an urban area and its dependence with population size is proposed.

Regarding their distribution over time, road accidents have been shown to follow some patterns. For example, in [Erdogan et al., 2008], the authors use kernel density analysis and repeatability analysis to identify accident hotspots in Turkey. Then, by studying the accident frequencies on a monthly and a daily basis, they find that most fatal accidents tend to occur

in the summer and over the days forming the weekend (Friday, Saturday and Sunday), especially around midnight. In [Anderson, 2009], by using a similar hotspot-detection analysis, Anderson finds fifteen types of accident hotspots in London, United Kingdom, many of them corresponding to weekend or late night driving. The authors in [Kumar and Toshniwal, 2016] also find that high frequency accident locations register the higher number of accidents during the late evening hours.

Rather than looking for accident hotspots, here the data is aggregated by the hour of the day and the time of the week, revealing the evolution in time of the road accidents with different degrees of severity reported in England and Wales, in France and in Spain. In road safety research, it is common to include data corresponding to more than one year, especially for the time-series analysis, as the number of road accidents and casualties in only one year may be strongly affected by local, temporary conditions (e.g. existing work zones, infrastructure projects, traffic arrangements or even one major accident with multiple casualties). In England and Wales and France, complete and reliable road safety data are available for several decades, so the analysis is performed on data corresponding to the period from 2008 to 2018. In the case of Spain, data is only available from 2008 to 2015. Furthermore, different time series are generated for different degrees of accident severity as well as different type of road —rural and urban.

Although road accidents are a global concern, there are still unanswered questions about how the number of road traffic accidents is spread across different locations. When it comes to urban areas, the following question arises: do locations with a larger population size register more accidents? This question does not have a straightforward answer since there is more dense traffic in larger cities [Louf and Barthelemy, 2014], but more traffic may cause lower average driving speeds and therefore, fewer severe accidents [Shefer and Rietveld, 1997; Theofilatos, 2017; Xu et al., 2013].

Whitelegg is perhaps one of the pioneering authors to address the question by taking a geographic approach to road traffic accidents in Great Britain. Since his 1987 publication [Whitelegg, 1987], the precision and availability of road safety data has improved considerably. Since then, some studies have focused on identifying locations of high accident frequency (see [Erdogan et al., 2008; Anderson, 2009; Kumar and Toshniwal, 2016]). For example, the authors in [Prieto Curiel, González Ramírez and Bishop, 2018] introduce a metric to measure the concentration of road traffic accidents based on a mixture model applied to the counts of accidents over a discretised space. Their analysis is performed in London's urban roads and on Mexican motorways, showing that about 5% of the road junctions are the site of 50% of the accidents, while around 80% of the road junctions expect close to zero accidents.

There are, however, fewer works about the relation between the population size of a given location and the associated number of accidents. Here, the framework of urban scaling, which was already discussed in Chapter 2, is used to model the relationship between the population size X of an urban area and the number of accidents Y occurring in that urban area over a certain period of time. The main hypothesis of urban scaling is that X and Y are related according to:

$$Y(X) = \alpha X^{\beta}, \tag{4.1}$$

where the parameter β is the scaling exponent and α is a proportionality constant, with these both parameters to be estimated from data.

In 1949, before the application of urban scaling models became popularised, Smeed [Smeed, 1949] proposed a rule of the following form

$$\frac{D}{X} = \gamma\Big(\frac{V}{X}\Big)^{\delta}, \tag{4.2}$$

that relates the yearly number of deaths by road traffic accident D, the size of the population X, and the number of registered vehicles V in a certain country. The constants γ and δ are evaluated by fitting this model to data. When data referring to various countries was considered, the parameters in (4.2) were estimated to be, approximately, $\hat{\gamma} \approx 0.0003$ and $\hat{\delta} \approx 1/3$. This rule, known as Smeed's law, was revised by Adams in 1987 [Adams, 1987], who proposed using vehicle miles instead of the number of vehicles per person as a measure of exposure to traffic. Smeed's law was further revised by Andreassen in two publications from 1985 [Andreassen, 1985] and 1991 [Andreassen, 1991], where he discussed that the values of the parameters γ and δ given by Smeed are not correct since they arise as the result of a spurious correlation of the variables D/X and V/X, as both of these variables contain the population X in the denominator. Andreassen also recommended considering the number of accidents and their severity instead of just the number of deaths, especially when the results are to be used for accident reduction purposes.

The estimated value of the scaling exponent β in equation (4.1) is found here for accidents of different degrees of severity by fitting this equation to the data, instead of equation (4.2). Furthermore, based on the same mathematical formulation, an expression for the likelihood of suffering road traffic accidents in an urban area of a certain population size during a given period of time is also derived.

4.2 Data and methods

As it has already been discussed in Chapter 2, due to the increasing opportunities for interactions enhanced by the thriving technologies and infrastructures, urban areas are no longer constrained to concrete physical boundaries, but instead sprawl across the landscape to accommodate the

interactions and flows that determine the urban network. However, for the purposes of the analysis in this Chapter, it is necessary to establish the extent of the urban areas of interest. By doing this, it is then possible to produce, unequivocally, population and road traffic accident counts.

4.2.1 Geographic data

A classification of urban areas based on land-use is deemed appropriate for this Chapter, since the environmental features of urban and rural roads are relevant factors in determining the frequency of road traffic accidents. This classification is explained in detail in Chapter 2, for England and Wales, France and Spain, which are the three territories considered here.

Even though the entity 'England and Wales' is not a country, it will be referred as such in this Chapter for simplicity.

4.2.2 Road safety data

As it is the case for geographic data, different countries also record their data regarding road traffic accidents in different ways, which are described below. For each territory, the recorded location of each road traffic accident is used to decide whether it happened in one of the urban areas described above, or in a rural area.

As it is explained below, the degree of severity of an accident in each of the three territories under consideration is determined by the most severely injured victim. Hence, for example, a fatal accident involves at least one death and, as a consequence, the number of deaths due to road traffic accidents in a given location is, in general, higher than the number of fatal accidents.

England and Wales

The databases 'Road Safety Data - Accidents' corresponding to the years from 2008 to 2018 contain all the road traffic accidents that were reported in Great Britain (England, Wales and Scotland) over this period and can be found for download on the `https://data.gov.uk/` website.

Due to the fact that geographical data for England and Wales (E&W) and for Scotland and Northern Ireland are held separately, only the road traffic accidents that occurred in English and Welsh urban and rural roads during the period 2008-2018 are considered here. Table 4.1 gathers the number of accidents in E&W according to their severity (fatal, serious or minor) and the type of area (urban or rural) where they occurred.

Table 4.1: Number of road traffic accidents during the period 2008-2018 in England and Wales according to their severity and the type of area where they occurred.

	Urban			**Rural**			
	Fatal	**Serious**	**Minor**	**Fatal**	**Serious**	**Minor**	**Total**
2008	773	12,257	88,896	1,321	8,628	46,380	158,255
2009	686	11,609	85,997	1,172	8,396	43,965	151,825
2010	539	11,067	82,369	1,003	7,666	41,463	144,107
2011	584	11,634	81,442	1,037	7,681	39,119	141,497
2012	570	11,757	77,547	911	7,427	37,682	135,894
2013	493	10,716	73,044	951	7,383	36,309	128,896
2014	550	11,198	79,090	926	7,937	37,344	137,045
2015	552	10,870	75,546	910	7,700	35,569	131,147
2016	536	11,751	71,573	957	8,544	34,880	128,268
2017	581	13,120	69,284	954	8,037	30,872	122,848
2018	592	13,533	64,561	927	8,251	28,317	116,181
Total	6,483	129,512	849,349	11,069	87,650	411,900	1,495,963

In this work, the accidents are classified in E&W as urban if i) they are registered as urban in the Road Safety databases (see definition of urban/rural for road safety purposes here [Department for Transport, 2017*a*]) and ii) the registered lower layer super output area (LSOA) of their loca-

tion corresponds to a BUA with more than 15,000 people, according to the lookup table retrieved from [Office for National Statistics, Open Geography Portal, 2011]. The accidents that are registered as rural in the Road Safety databases or occur outside a BUA or within BUAs of population size smaller than 15,000 are classified as rural for the purposes of this analysis.

In Great Britain, an accident's severity is established according to that of the most severely injured casualty. Human casualties are officially classified as fatal, serious or slight. Throughout this work, however, accidents where the most severe injury is slight, will be referred to as 'minor accidents'. Fatal casualties are those where the sustained injuries cause death less than 30 days after the accident (confirmed suicides are excluded). Injuries where the victim needs to be detained in hospital as an 'in-patient' are deemed to be serious. Fractures, concussions, internal injuries, crushings, burns (excluding friction burns), severe cuts, severe general shock requiring medical treatment and injuries causing death 30 or more days after the accident are also considered to be serious even if hospitalisation is not involved. Finally, slight injuries, such as a sprain (including neck whiplash injury), bruising, cuts or minor shocks, are those that have a minor character and sometimes they do not even require medical treatment [Department for Transport, 2017*a*].

France

For metropolitan France (i.e. all French territories that are within the European continent), detailed data about the road traffic accidents reported to the order forces involving personal injury during the period spanning from 2008 to 2018 can be found in [Observatoire National Interministériel de la Sécurité Routière, 2018]. Table 4.2 gathers the number of accidents in France according to their severity and the type of area where they occurred.

Table 4.2: Number of road traffic accidents during the period 2008-2018 in France according to their severity and the type of area where they occurred.

	Urban			Rural			
	Fatal	**Serious**	**Minor**	**Fatal**	**Serious**	**Minor**	**Total**
2008	603	11,323	28,031	1,304	13,212	11,293	65,766
2009	577	9,882	25,873	1,025	11,727	13,041	62,125
2010	579	9,821	25,806	1,068	11,169	11,141	59,584
2011	555	9,716	24,561	1,023	10,879	10,563	57,297
2012	520	9,047	23,002	2,028	9,394	8,987	52,978
2013	539	8,731	20,753	965	8,835	9,822	49,645
2014	499	8,583	21,449	955	9,674	9,902	51,062
2015	500	8,526	19,367	1,181	12,599	8,110	50,283
2016	550	8,449	18,650	1,272	10,812	10,750	50,483
2017	531	8,854	18,420	923	11,482	11,937	52,147
2018	501	6,520	19,777	1,054	8,071	13,668	49,591
Total	5,954	99,452	245,689	12,798	117,854	119,214	600,961

French accidents are classified as urban if i) they are registered as *en agglomération* in the Road Safety databases and ii) the registered *commune* of their location corresponds to an UU with more than 15,000 people, according to the lookup table retrieved from [Institut National de la Statistique et des Études Économiques, 2018*a*]. The accidents that are registered as rural in the Road Safety databases or occur outside a UU or within UUs of population size smaller than 15,000 are classified as rural for the purposes of this analysis.

It should be noted that, in France, only reported accidents requiring medical care are included in the data sets, while in E&W, all reported accidents are included. As in E&W, an accident is classified as fatal if at least one of the people involved dies within 30 days after the accident. The accident is serious if the most severely injured victim requires hospitalisation for more than 24 hours. And finally, the accident is minor when the most severely injured victim requires medical care but no hospitalisation lasting more than 24 hours.

Spain

Statistical data regarding the road traffic accidents with injuries in Spain can be found at [Dirección General de Tráfico, 2015]. However, data with the required amount of detail is only available for the period 2008-2015.

Table 4.3: Number of accidents during the period 2008-2015 in Spain according to their severity and the type of area where they occurred.

	Urban			**Rural**			
	Fatal	**Serious**	**Minor**	**Fatal**	**Serious**	**Minor**	**Total**
2008	1,004	8,149	55,380	1,063	4,416	11,995	82,007
2009	895	6,981	53,105	937	3,520	11,869	77,307
2010	812	5,936	53,154	892	3,157	11,085	75,036
2011	771	5,810	52,450	827	2,887	10,146	72,891
2012	772	5,512	53,776	699	2,459	10,484	73,702
2013	684	5,641	58,990	592	2,298	11,431	79,636
2014	725	5,205	59,539	617	2,090	11,046	79,222
2015	694	5,296	68,079	648	2,078	10,937	87,732
Total	6,357	48,530	454,473	6,275	22,905	88,993	627,533

For the analysis of the temporal distribution of the road traffic accidents, accidents in Spain are classified as urban if the registered *municipio* of their location corresponds to an urban area with more than 15,000 people, according to the lookup table retrieved from [Ministerio de Fomento, 2018*b*], or as rural otherwise. Since the Great and Small Urban Areas in Spain have very different characteristics, only the Great Urban Areas or GAUs, are considered for the study of the relationship between road traffic accident incidence and population size of urban areas: in particular, the accidents that occur inside a GAU with a population size larger than 50,000 are classified as urban.

Like in France, only reported accidents requiring medical care are included in the Spanish data. Accidents are classified as fatal if at least one of the people involved dies within 30 days after the accident; as serious

if the most severely injured victim requires hospitalisation for more than 24 hours and as minor if the most severely injured victim requires medical care but no hospitalisation lasting more than 24 hours.

4.3 Results

4.3.1 Temporal distribution of the road traffic accident incidence

A clearer understanding of the incidence of road traffic accidents of different severity levels in urban areas and how this compares to the incidence in rural areas, can be gained by visualising the data. In Figures 4.4, 4.5 and 4.6 the weekly evolution of the number of accidents of fatal, serious and minor severity respectively are displayed. The accidents corresponding to E&W and France occurred from the year 2008 until 2018 and in Spain, from the year 2008 until 2015 are displayed. Accident frequency by the hour and day of the week is represented in the vertical axis and a simple 2-hour window moving mean calculation is applied in all cases to smooth the graphs. The discretisation of the time dimension by the hour and week day, represented in the horizontal axis, is chosen for display over others (for example by day and month) as it produces more insightful temporal patterns.

Based on Figure 4.4, rural areas in E&W and France experience more fatalities than urban areas. This is not the case for Spain, where it should be recalled that the definition of urban area is more relaxed than in other countries. In the three territories, fatal accidents seem to increase on Friday and Saturday nights, which are also the most popular times for alcohol consumption [Kuntsche and Cooper, 2010; Lac et al., 2016].

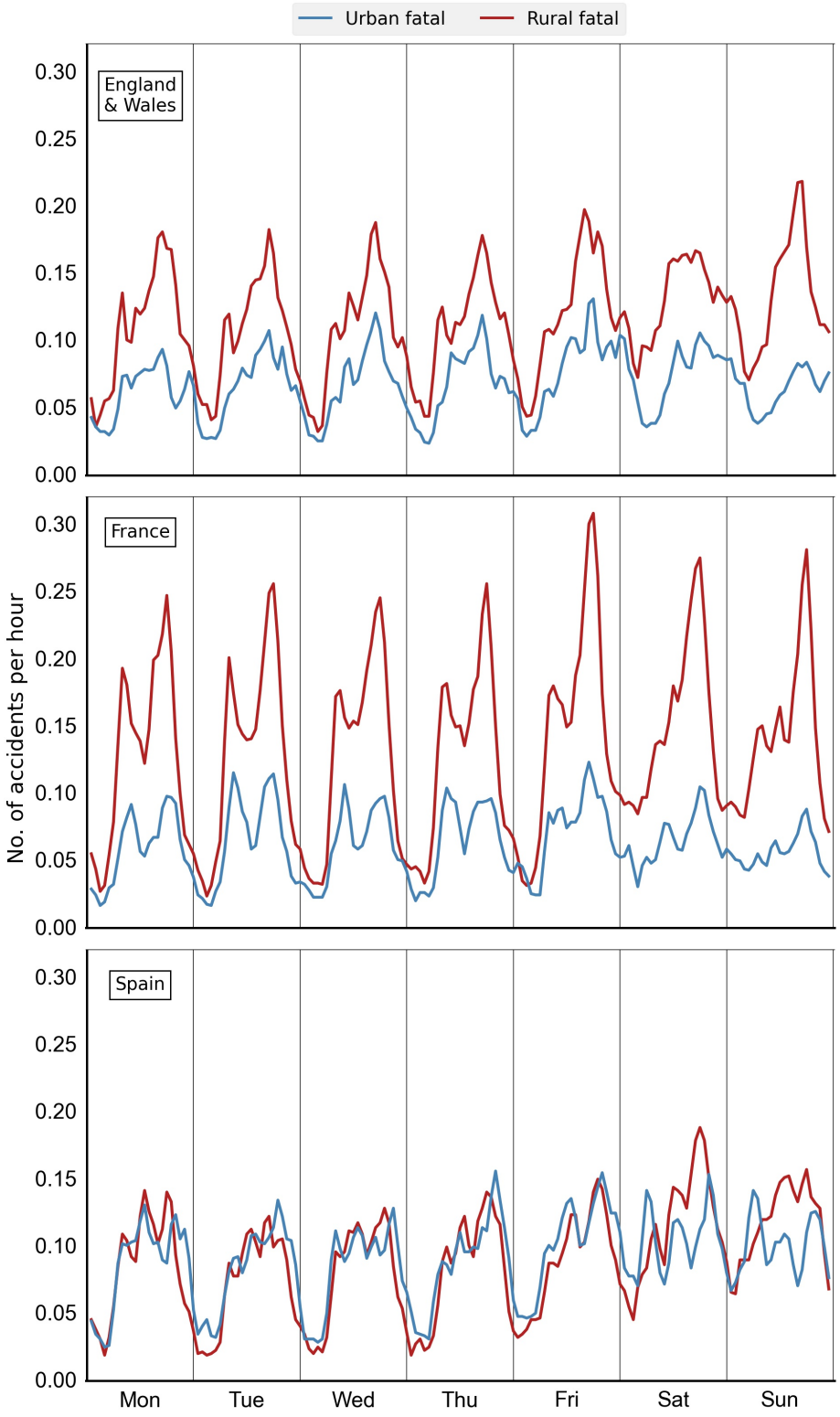


Figure 4.4: Evolution of the number of fatal accidents per hour throughout the week in E&W (top row), France (middle row) and Spain (bottom row).

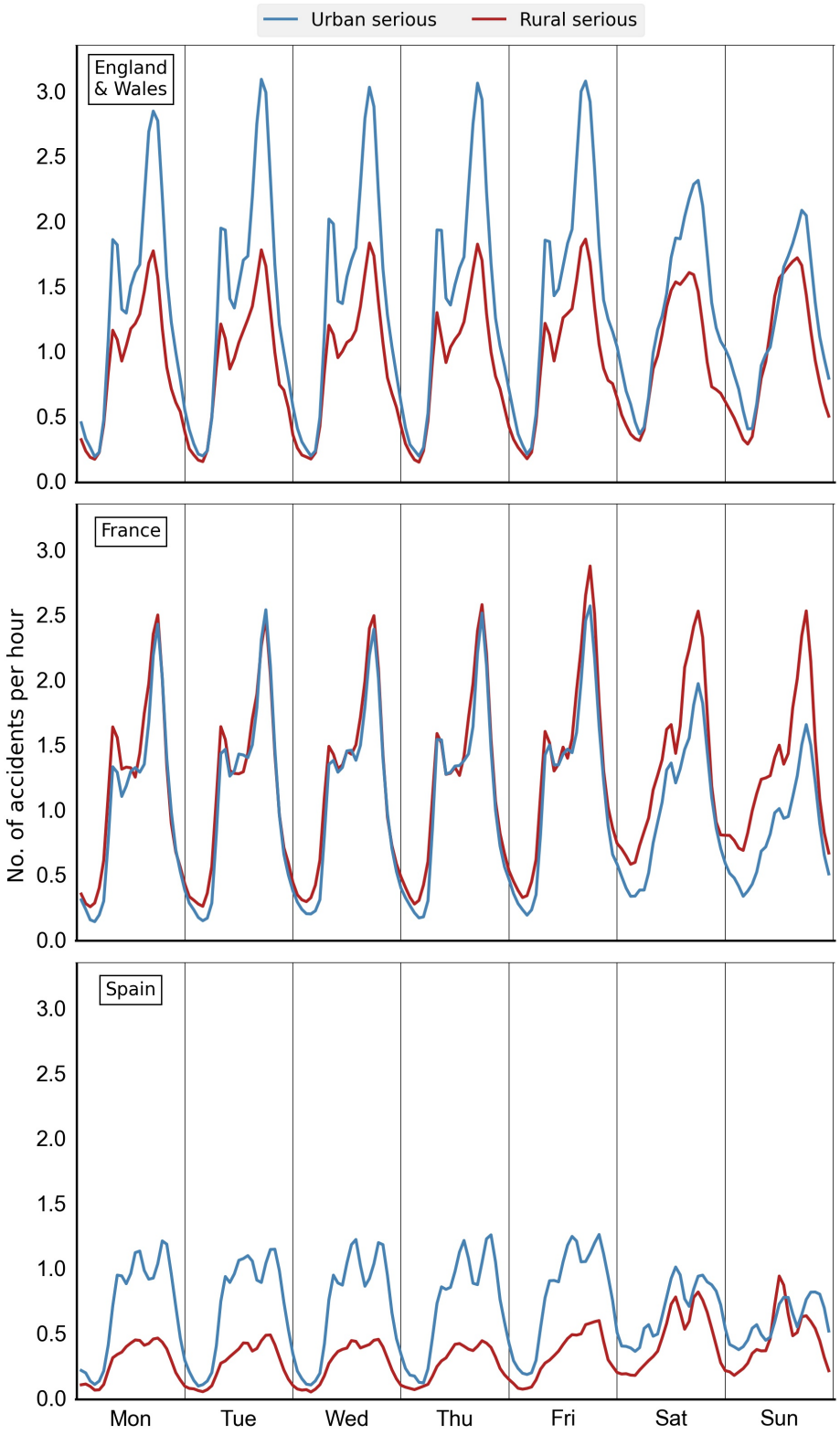


Figure 4.5: Evolution of the number of serious accidents per hour throughout the week in E&W (top row), France (middle row) and Spain (bottom row).

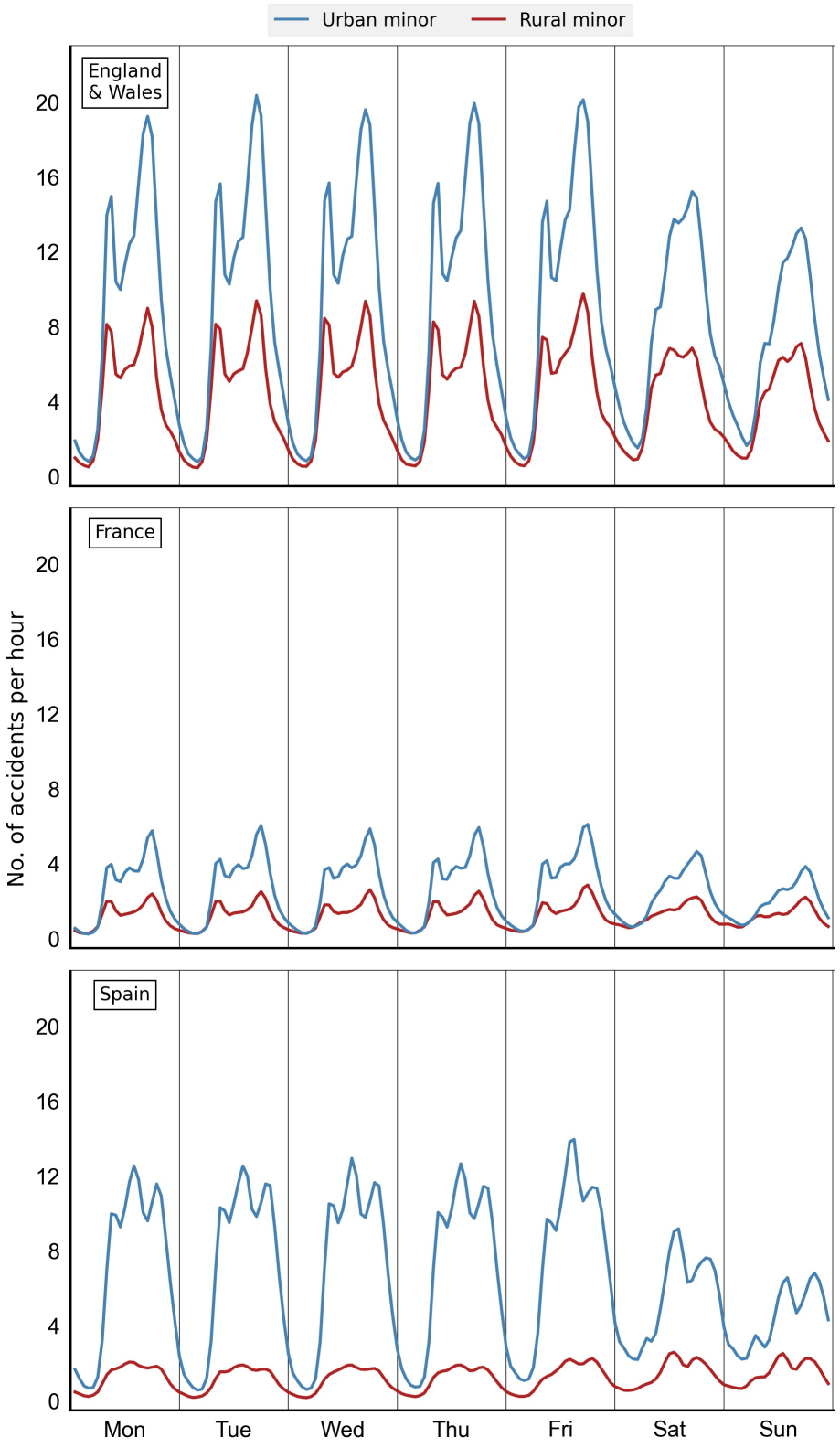


Figure 4.6: Evolution of the number of minor accidents per hour throughout the week in E&W (top row), France (middle row) and Spain (bottom row).

Figure 4.5 reveals that the occurrence of serious accidents, follows a clear bimodal distribution with daily periodicity during the week days in rural and urban areas from E&W and in rural areas from France, with each of the peaks corresponding to the morning and evening rush hours. In Spain and, to a lesser extent, also in France, the traffic accidents occurring in urban areas follow a trimodal daily periodicity during the week days, reflecting the customary longer lunch breaks in these countries, which allow some workers to commute giving rise to an afternoon rush hour. The larger size of the peaks corresponding to the evening rush hour in E&W and in France is something to note and it might be due to factors such as a more congested traffic flow, higher levels of stress and exhaustion in the drivers after a day of work, reduced visibility during the evening or a combination of these.

During the weekends, the distribution becomes closer to unimodal in E&W and France and closer to bimodal in Spain, reflecting the fact that less people are travelling to work. Even though Spain's population (46.94 M as of 2019) is smaller than that of E&W (59.44 M) or France (67.06 M), Figure 4.5 shows a much lower number of serious accidents for Spain than for the other two countries. This discrepancy in numbers is possibly due to underreporting.

In Figure 4.6, minor accidents exhibit patterns that are similar to those corresponding to serious accidents for the three countries. As in Figure 4.5, there are large discrepancies in the total number of accidents, showing that E&W reports a lot more minor accidents than France and Spain, where only minor accidents requiring medical care are reported.

Remarkably, Figures 4.5 and 4.6 show that both serious and minor accidents have a much higher incidence in urban than in rural areas.

4.3.2 Relationship between road traffic accident incidence and population size

The total number of traffic accidents in each urban area in a given time period is count data. For this reason, the data for the number of accidents is assumed to follow a Poisson distribution and fitted using a Poisson regression to equation (4.1). The same statistical analysis is applied across all degrees of severity. The maximum likelihood estimates for the parameters, with 95% confidence intervals are displayed in Table 4.4, where the subscripts F, S, M have been added to indicate whether the parameters have been computed for accidents of fatal, serious or minor severity, and T for accidents of all degrees of severity. These estimates are obtained with traffic accident data corresponding to the years 2008-2018 for E&W and France and 2008-2015 for Spain; the populations are those corresponding to the middle point of each period: the 2013 and the 2011 population estimates respectively.

Table 4.4: Maximum likelihood estimates for the parameters corresponding to the urban scaling models describing the relation between the number of accidents of different degrees of severity in urban areas and their population sizes.

		E&W	**France**	**Spain**
Fatal	$\log \alpha_F$	-9.925 ± 0.087	-7.501 ± 0.061	-5.845 ± 0.094
	β_F	1.076 ± 0.006	0.959 ± 0.005	0.836 ± 0.007
Serious	$\log \alpha_S$	-6.813 ± 0.019	-6.035 ± 0.015	-5.819 ± 0.033
	β_S	1.067 ± 0.001	1.063 ± 0.001	0.991 ± 0.002
Minor	$\log \alpha_M$	-5.556 ± 0.008	-5.818 ± 0.010	-6.128 ± 0.011
	β_M	1.113 ± 0.001	1.114 ± 0.001	1.178 ± 0.001
All severity levels	$\log \alpha_T$	-5.321 ± 0.007	-5.230 ± 0.008	-5.719 ± 0.010
	β_T	1.107 ± 0.001	1.097 ± 0.001	1.156 ± 0.001

The results for the estimated values of the parameters in Table 4.4 suggest that the total number of accidents increases superlinearly with population size. However, if the number of accidents is broken down by severity level, it is observed that, while the number of accidents increases with population size, it tends to increase faster for lower levels of severity. Specifically, minor accidents in E&W, France and Spain present a superlinear behaviour whereby the number of accidents in urban areas grows faster than proportionally with population size. While in E&W, serious and fatal accidents display a weaker superlinear behaviour (but still superlinear), in France, only serious accidents grow superlinearly and in Spain, neither serious nor fatal accidents grow superlinearly with population size.

These results are displayed in Figure 4.7, where the number of fatal and serious accidents have been represented against the number of minor accidents for E&W, France and Spain. Each dot is a data point corresponding to an urban area and its size depicts the population size of the urban area that it represents. The yellow and blue lines have been obtained with the values of the estimated parameters and, when written in terms of the number of minor accidents Y_M, they can be expressed as $\alpha_S\left(\frac{Y_M}{\alpha_M}\right)^{\frac{\beta_S}{\beta_M}}$ or $\alpha_F\left(\frac{Y_M}{\alpha_M}\right)^{\frac{\beta_F}{\beta_M}}$ respectively. The lines are computed using a range of values for X which are comprised between the minimum and maximum populations registered in each territory. It is consistently observed that the number of accidents for smaller urban areas is overestimated by the scaling model.

4.3.3 Probability of a traffic accident of a given degree of severity

In the previous section, the number of accidents in urban areas is modelled as a power-law of their population size. But, given an accident, does the

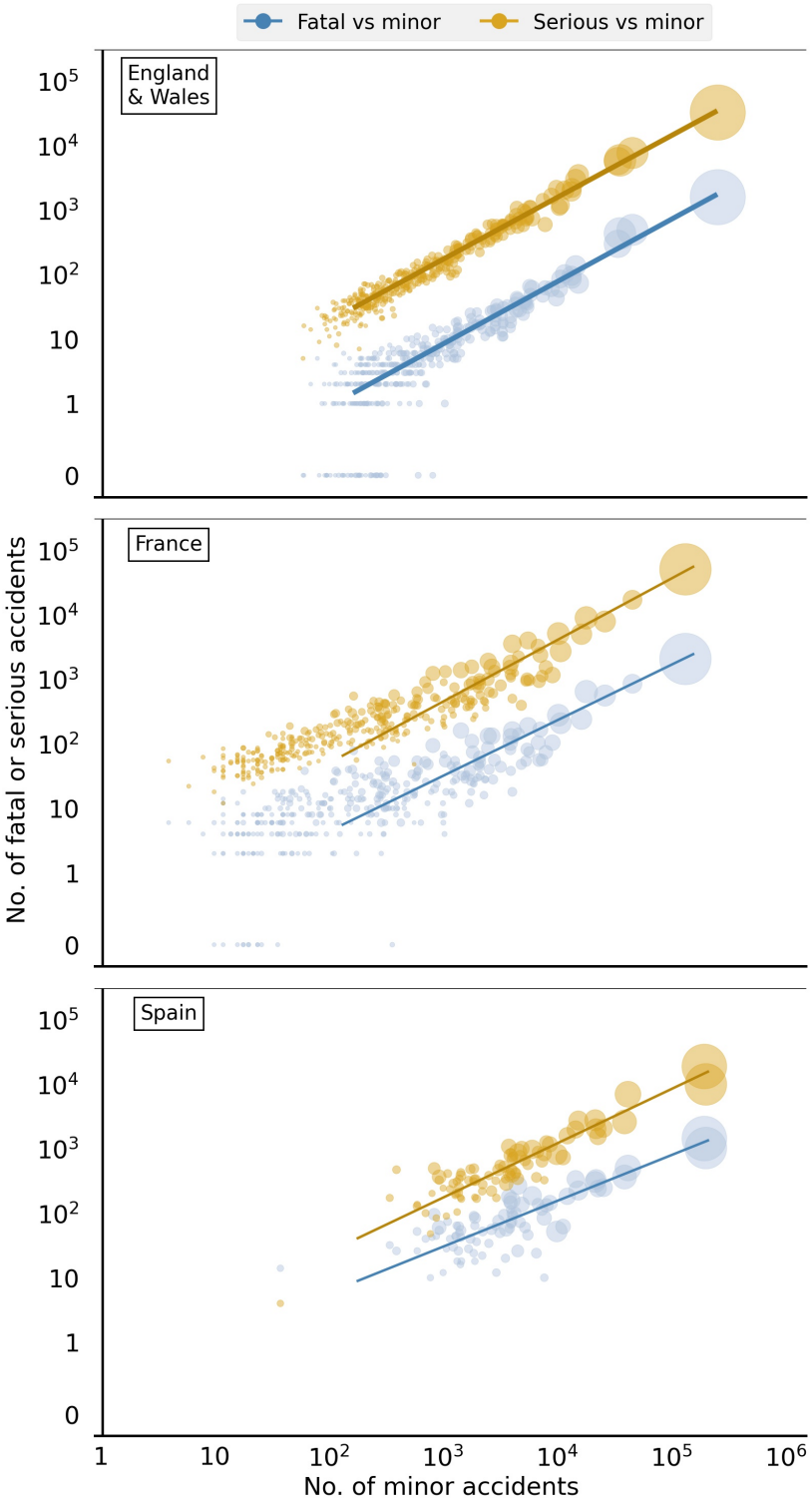


Figure 4.7: No. of serious accidents against no. of minor accidents (yellow) and no. of fatal accidents against no. of minor accidents (blue) for E&W, France and Spain. The size of the dots size represents the population sizes corresponding to each urban area. The lines are obtained with the parameters gathered in Table 4.4.

probability that it is of minor, serious or fatal severity also vary with the population size of the urban areas?

Consider the probability mass function over the random variable A, which corresponds to the number of accidents that an individual residing in an urban area of population $X = x$ suffers in a given period of time. If $Y = y$ is the total number of accidents occurring in the urban area in the same period of time, this probability mass function is a binomial distribution of the type:

$$\Pr(A = a|X = x, Y = y) = \binom{y}{a} p^a (1-p)^{y-a}, \tag{4.3}$$

where p is the probability that a certain individual in the urban area under consideration suffers a accidents. Since the population size is x, then, assuming all individuals have the same probability of suffering an accident, $p = 1/x$. According to equation (4.3), the probability that a given individual suffers at least one accident is given by:

$$\Pr(A > 0|X = x, Y = y) = 1 - \Pr(A = 0) = 1 - (1-p)^y \approx yp, \tag{4.4}$$

where the approximation in the last step can be taken when $p << 1$. For the urban areas taken into account here, all the populations are above 15,000, so p is smaller or equal to 6.667×10^{-5}, and therefore, the approximation is justified. Using equation (4.1) to write Y in terms of X and the fact that $p = 1/x$:

$$\Pr(A > 0|X = x, Y = y) \approx \alpha x^{\beta - 1}. \tag{4.5}$$

This equation can also be applied to accidents of a particular degree of severity. It then follows that the probability that given an accident, the

accident is, for example, fatal, is the conditional:

$$\Pr(F|A>0; X=x, Y=y) = \frac{\Pr(F\cap A)}{\Pr(A)} = \frac{\alpha_F}{\alpha_T} x^{\beta_F-\beta_T}, \qquad (4.6)$$

where A denotes the event of an accident; F, the event of a fatal accident; α_F and β_F, the parameters from equation (4.1) when only fatal accidents are considered; and finally, α_T and β_T, the parameters when accidents from all degrees of severity are considered. The probability that given an accident, the accident is serious or minor can be obtained analogously.

According to the estimated values for the parameters shown in Table 4.4, the probability that an individual suffers at least a fatal accident (an individual is said to suffer a fatal accident if there was at least one fatal injury in the vehicle) in an urban area from E&W during the period 2008-2018 increases sublinearly with the population size of the urban area since the exponent in equation (4.5) for fatal accidents is $\beta_F - 1 = 0.076 \pm 0.006$. In France and Spain, however, the probability decreases sublinearly with population size, as $-1 < \beta_F - 1 < 0$ in those cases. Similar results can be observed for serious accidents, except in this case, the probability of suffering at least a serious accident in France during the period 2008-2018 also increases sublinearly with population size. For minor accidents, the probability in the three territories increases sublinearly.

From equation (4.6), it can be learned that the probability that, given an accident in an urban area, the accident is fatal, decreases sublinearly with population size in E&W, France and Spain, since in the three territories the quantity $\beta_F - \beta_T$ is between -1 and 0. Therefore, even though the probability that suffering at least an accident during the period 2008-2018 becomes higher as the population size of an urban area increases, the probability that this accident is fatal decreases with the population size in the case of E&W. Serious accidents follow a similar trend. However,

the probability that a given accident is minor increases sublinearly with population size as $0 < \beta_M - \beta_T < 1$ in the three territories.

4.4 Discussion

Environmental factors and their influence on drivers' behaviour, combine to give rise to specific trends in the incidence of road accidents. Through the generation of time series regarding the frequency of road traffic accidents occurring in England and Wales (E&W) and France from the year 2008 to the year 2018 and in Spain from the year 2008 to the year 2015, it can be learned that accidents leading to serious and minor injuries occur more often on urban roads than on rural roads and they produce clear patterns throughout the week: on weekdays, there are, depending on the country, two or three daily peaks of different height in the number of accidents in both urban and rural roads, the lower ones corresponding to the morning and afternoon rush hours, and the higher one, to the evening rush hour; at weekends, there are less accidents than on weekdays and the peaks are no longer present; during the night, the number of serious and minor accidents is relatively very low. As it has already been found before in the literature, e.g [Zwerling et al., 2005], it is observed here that fatal accidents tend to take place in rural areas with a higher frequency and they are more evenly spread throughout the day. In addition, fewer accidents occur at night, except possibly on Friday and Saturday night. It is known that the following factors contribute to the frequency of fatal accidents being higher in rural than in urban areas: the characteristics of the rural crashes (more likely to be head-on crashes and single vehicle accidents with stationary objects [Zwerling et al., 2005]), rural drivers' demographics (larger proportion of older drivers in rural areas, with increased fragility and higher chances of dying [Tefft, 2017]), their typical behaviours (their travelling speeds in rural areas may be greater [Zwerling et al., 2005], they are less likely to use seat belts [Beck et al., 2017], etc.) or the difficulty of obtaining timely medical assistance on rural roads [Gonzalez et al., 2006; Byrne et al., 2019]. It is

estimated that, approximately 20% of fatal accidents in E&W during 2017 were drink-drive accidents [Department for Transport, 2017*b*]. This could explain the rise in the number of fatal accidents on Friday and Saturday nights, since alcohol consumption is higher on the weekends [Kuntsche and Cooper, 2010; Lac et al., 2016].

Turning the attention to urban areas, an approximate power-law scaling behaviour between the number of urban road accidents and the population size of the urban areas where they take place is assumed. The total number of accidents at each location scales faster than linearly. This superlinear effect is stronger the lower the degree of severity of the accidents, i.e. minor accidents can be modelled with a scaling exponent that has a higher value than for fatal and serious accidents. Generally speaking, urban areas in E&W, France and Spain do not present economies of scale in terms the number of road accidents. However, France displays a sublinear behaviour for fatal accidents and so does Spain for both serious and fatal accidents.

It is also observed that the probability of suffering at least one accident of any degree of severity in an urban area over the period 2008-2018 increases sublinearly with population size in E&W. In France, this is also true except for fatal accidents, in which case, the probability decreases sublinearly. In Spain, the probability of suffering at least one accident or, specifically, a minor accident in the period 2008-2015 increases sublinearly; if the accident is serious or fatal, this probability decreases sublinearly instead. The probability that a given accident is fatal or serious decreases sublinearly and the probability that it is minor increases sublinearly in the three territories. That is, on a per capita basis, more populated urban areas are more prone to accidents, but are less deadly than the less populated counterparts.

But, why are these scaling behaviours observed? It is a known fact that, as the population size of an urban settlement increases, the road surface

also increases but it does so sublinearly [Bettencourt et al., 2007]. It is therefore expected that, due to this reduction of space, the traffic congestion delay —which could perhaps be classified as an environmental factor— increases superlinearly, as shown in [Louf and Barthelemy, 2014]. Whether traffic congestion has an impact on the frequency of road accidents still remains as an open question, with different conclusions reached by different studies. For example, [Wang et al., 2009] suggest that traffic congestion has little or no impact on the frequency of road accidents, although their results are constrained to the M25 orbital London motorway. Others conclude that congestion could lead to a reduction in the number of fatalities [Shefer and Rietveld, 1997], but again, this result is restricted to highways and only considers fatal accidents. In [Theofilatos, 2017] and [Xu et al., 2013], it is shown that variations in traffic congestion levels significantly influence accident occurrence although they have a generally mixed influence on accident severity: low severity accidents tend to occur in congested traffic flow conditions whereas severe and fatal accidents occur more often when the traffic is uncongested and when there are large differences in speed between adjacent lanes. These conclusions seem to agree with the result obtained here with regards to fatal accidents: the majority of fatal accidents occur in rural areas and the power-law relation with population size generally presents a lower scaling exponent than for accidents of minor severity.

Even if traffic congestion did not have a direct impact on the frequency of road accidents, previous works show that stress levels from drivers would be higher when driving in highly congested traffic conditions [Hennessy and Wiesenthal, 1997; Wener and Evans, 2011] and their satisfaction levels would be worse due to an increase in the travelling time [Higgins et al., 2018]. Furthermore, drivers' stress levels are influenced not only by aspects related to the driving context —such as traffic congestion— but by

a myriad of situational and personal factors [Hennessy et al., 2000] that seem to be enhanced in cities. Several studies [Trivedi et al., 2008; Srivastava, 2009; Rishi and Khuntia, 2012] point towards urban areas as a possible cause of certain mental health disorders, with anxiety, depression and socioeconomic stress among them. A poor mental health status along with other psychological states related to sleep, fatigue, alertness, physical activity, emotional situation, etc. have been demonstrated to be a risk factor for road accidents [Taylor and Dorn, 2006; Lagarde et al., 2004; Simon and Corbett, 1996]. Therefore, an increase in traffic congestion as the population size of an urban area gets larger contributes to the drivers' higher level of stress and, together with many other urban factors, could plausibly cause the observed superlinear scaling laws. A fully comprehensive analysis of the causes of the different scaling behaviours for accidents of different severity requires cross-disciplinary work from experts in urbanism, psychologists, engineers and policy makers, since road accidents are the result of highly complex interactions of environmental, driver, vehicle, socioeconomic and legislative factors.

In the light of the generally observed superlinear scaling behaviour of road accidents, especially those of minor severity, multiple ways of proceeding can be proposed. One is that urbanisation and road accident prevention strategies should now be conceived with a special focus on population distribution among the different cities within a country. Another is that road accident prevention strategies should also focus on understanding the causes of urban road accidents.

It is important to note that notwithstanding its mathematical simplicity, scaling theory comes with flaws. Indeed, it is known that scaling exponents fluctuate considerably when they are different from one, depending, for example, on how the boundaries of the different urban areas are defined [Arcaute et al., 2015]. The comparison between the results presented here

for different types of area and different degrees of severity, can still prove to be insightful as long as all the assumptions about the data and the modelling techniques are taken into account for their interpretation. Some of the drawbacks of the use of scaling laws are explored in the next Chapter, with a particular emphasis on the issue of dragon-king urban areas, which due to their unusually large population are hard to account for statistically speaking.

Chapter 5

Dragon kings and urban scaling models

5.1 Introduction

The work presented in this Chapter has been partially covered in the research paper entitled 'The effect of dragon kings on the estimation of scaling law parameters' [Cabrera-Arnau et al., 2020].

In the previous Chapters, it has been demonstrated how a certain characteristic Y of an urban area can be mathematically related to its population size X through an urban scaling model of the form

$$Y(X) = \alpha X^{\beta}, \tag{5.1}$$

where β is the scaling exponent, which is generally different from 1, and α is a proportionality constant.

Urban scaling models allow us to classify a set of data points as linear, superlinear or sublinear, depending on whether the scaling exponent is equal, larger or smaller than 1. The estimated value of the scaling exponent, denoted by $\hat{\beta}$, provides information about whether urban areas become

more efficient or productive with respect to a certain characteristic X as population size increases [Marshall, 2013].

However, it is challenging to quantify reliably from data the value of the scaling exponents and their estimates usually come with caveats [Depersin and Barthelemy, 2018; Molinero and Thurner, 2019; Ribeiro and Queiros-Condé, 2017; Cottineau et al., 2017]. For example, Arcaute *et al.* [Arcaute et al., 2015] raise the issue that, on the one hand, cities with extremely large populations such as London in the UK, tend to behave differently from the rest of cities within the urban system due to their socioeconomic role; therefore, rather than comparing these with cities within the same country's urban system, they should possibly be analysed relative to other cities with extremely large populations and separately to the rest of cities within an urban system. On the other hand, these extremely populous cities act as hubs, making the urban system highly integrated. From this point of view, all the cities in the urban system should be analysed together despite the fact that the value of the estimated scaling parameters might be dominated by the unusual behaviour of extremely populous cities.

For example, the urban area corresponding to London has a population size of more than 10 million people. The second largest urban area in England and Wales is that corresponding to Manchester, with a population of around 3 million people. The population of London is thus approximately three times larger than the population of Manchester, hence granting London a very special role within the urban system of England and Wales. With its large population, London also outperforms other urban areas in the same region with regards to culture and socioeconomic power. Following up on the road accidents topic discussed in the previous Chapter, London's built-up area registered 25,200 road accidents in 2013, a figure at least seven times larger than the 3,000 road accidents registered in Manchester.

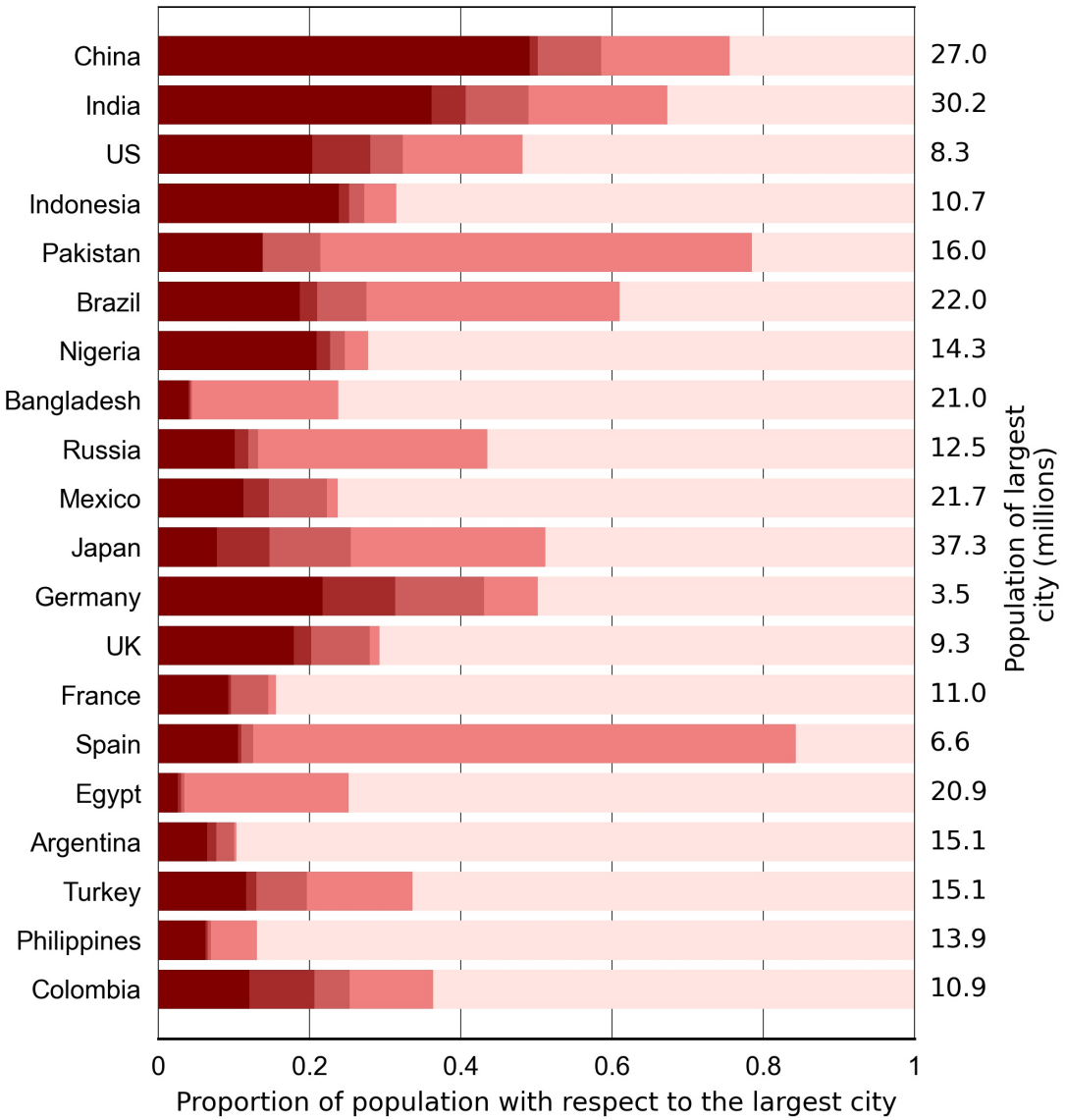


Figure 5.1: Population sizes of the top five most populous cities. The length of the five bars corresponding to each country represents the proportion of the population of the top five most populous cities relative to the population of the largest city in each country. Data retrieved from [United Nations, Department of Economic and Social Affairs, Population Division, 2018].

Places like London have traditionally been referred to as primate cities. The term was introduced by Jefferson in [Jefferson, 1939], where he talks about the amplifying mechanisms dictating the growth of large urban areas: 'once a city is larger than any other in its country, this mere fact gives

it an impetus that cannot affect any other city, and it draws away from all of them in character as well as in size . . . It becomes a *primate city*'. Jefferson points out the fact that the existence of primate cities is ubiquitous across different countries.

Indeed, Figure 5.1, shows the proportion of population of the top five largest cities in 20 countries, in relation to the largest city in each of these countries. In 13 of the 20 countries, the second largest city is at most half the size of the top largest and in 19 of the 20, the third largest is half the size of the top largest, showing that most countries have one or two cities that stand out because of their population size and possibly, with regards to other aspects.

For example, Argentina, stands out as one of the countries with the largest relative difference of population size between the first and second most populous cities. The largest city in the country is Buenos Aires, with a population of 15.1 million. The second city is Córdoba, which is about 0.1 the size of Buenos Aires, with a population of 1.57 million. The third city, Rosario, is also about 0.1 the size of Buenos Aires and has a population of 1.53 million. The fourth city is Mendoza, 0.08 the size of Buenos Aires and a population of 1.17 million. Finally, the fifth city is San Miguel de Tucuman, with 0.07 the size of Buenos Aires and a population of 0.99.

In other countries, the difference in populations between the largest and second largest cities is not so large. For example, the largest city in Spain is Madrid, with 6.6 million, and the second largest is Barcelona, with 5.6 million, which is 0.85 times the population of Madrid. The fact that there are two very large cities can be due to different reasons. In the case of Spain, the country has historically been divided into different kingdoms, each with its own capital city, and it was not until the late 15th century that the first big unification of these kingdoms took place. Even though

there is not a unique primate city, both Madrid and Barcelona are relatively much larger than the third largest city, Valencia, with just under 1 million people and about 0.13 the size of Madrid. As well as being the top two largest cities in the country, Madrid and Barcelona are also centres for socioeconomic power and international relations.

Rather than considering the extremely populous urban areas as a separate category of cities, in this Chapter the aim is to understand how their inclusion might affect the estimated values of the scaling parameters. This approach is philosophically in line with the complex systems perspective by which systems formed by many interacting parts need to be considered as a whole in order to properly understand their behaviour. Indeed, if Greater London was taken out from the urban system corresponding to the United Kingdom, the result of any analysis would be somewhat incomplete. In order to specify what is meant by extremely populous cities, the statistical concept of dragon king [Sornette, 2009] is introduced here. A 'dragon king' is defined as an event that, due to its large size, is statistically and mechanistically considered to be an outlier of the underlying heavy-tailed probability distribution followed by the rest of events. Sornette shows that dragon kings can be present in the distribution of city population sizes which, as is the case for many other systems in nature, follows a heavy-tailed probability distribution.

5.1.1 Heavy-tailed distributions

The size distribution of many events in nature is characterised by the presence of few extremely large events and a large number of smaller events. For example, over the course of a century, thousands of low-magnitude earthquakes will be recorded, however, there will probably be at most one earthquake of magnitude 8 or above. Similarly, there are only one or two

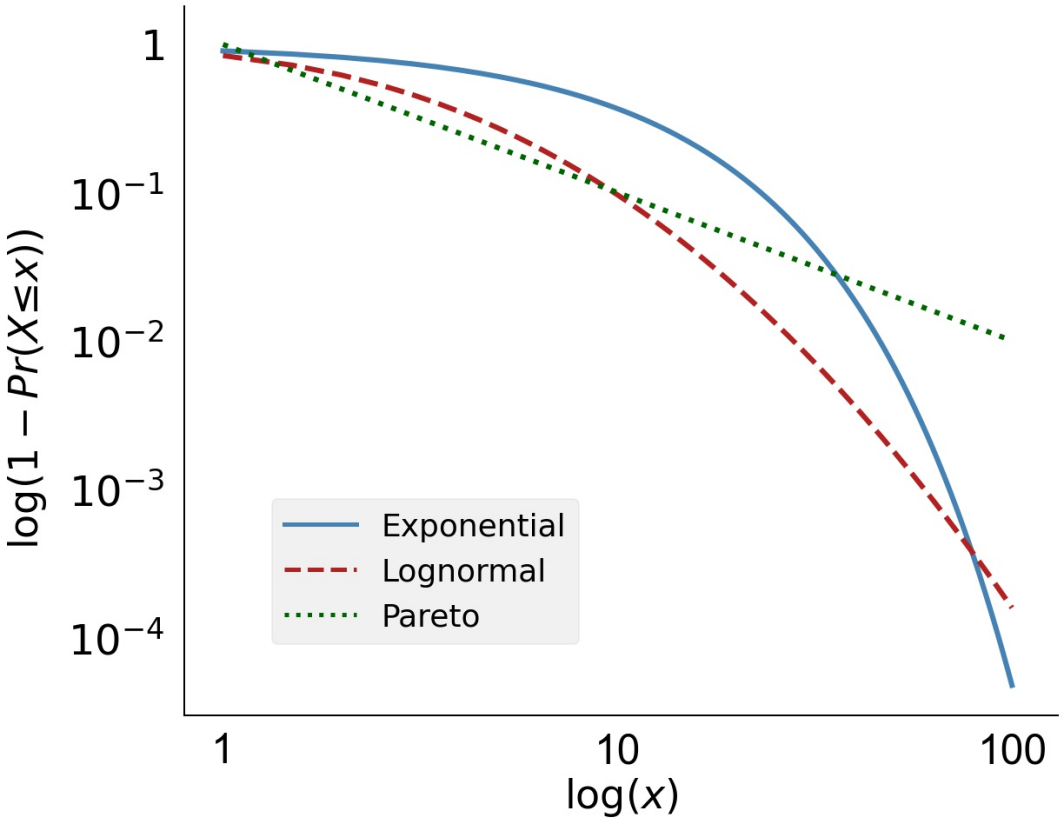


Figure 5.2: Representation of two heavy-tailed probability distributions, the power-law and the lognormal, and an exponential probability distribution. The vertical axis corresponds to $1 - \Pr(X \leq x)$, also known as survival function. Both axes are in logarithmic scale.

very large cities in a country, whereas there are lots of smaller ones. This distribution of city sizes was first identified by Le Maître [Le Maître, 1682] and then formalised by Auerbach [Auerbach, 1931].

Heavy-tailed probability distributions of event sizes are often used to account for this type of size distribution mathematically. The random variable X is said to follow a heavy-tailed distribution if the decay of the complementary cumulative distribution function $\Pr(X > x)$, also known as survival function or simply tail distribution, is slower than exponential, that is

$$\lim_{x \to \infty} \mathrm{e}^{tx} \Pr(X > x) = \infty \tag{5.2}$$

for all $t > 0$. Figure 5.2 illustrates the sub-exponential decay of heavy-tailed distributions. In particular, two types of heavy-tailed distributions have been represented together with an exponential probability distribution for comparison, whose cumulative distribution function is given by

$$\Pr(X \leq x) = 1 - \mathrm{e}^{-\lambda x}, \tag{5.3}$$

with $\lambda > 0$ being the parameter of the distribution, which has been set to 0.1 for the plot. The first type of heavy-tailed distribution is a Pareto, whose cumulative distribution function (CDF) has the following form

$$\Pr(X \leq x) = \begin{cases} 1 - \left(\frac{k}{x}\right)^a & x \geq k \\ 0 & x < k \end{cases} \tag{5.4}$$

where k and a are parameters whose values have been both set to 1 in Figure 5.2. The other type, with a decay slower than exponential but faster than that corresponding to a Pareto distribution, is a lognormal distribution, with a CDF of the form

$$\Pr(X \leq x) = \frac{1}{2} + \frac{1}{2}\,\mathrm{erf}\Big(\frac{\log x - \mu}{\sqrt{2}\sigma}\Big), \tag{5.5}$$

where erf is the error function and μ and σ are the distribution parameters, both set to 1. If the random variable X represents the size of an event and it is heavy-tailed distributed, then large events would be rare, but perfectly possible events.

5.1.2 Generating mechanisms and dragon kings

In [Pumain, 2000], Pumain reviews several models for the growth of urban population sizes. These models have the ability to demonstrate how

different growth processes could lead to several heavy-tailed distributions of city population sizes such as a Pareto [Steindl, 1965; Simon, 1955] or a lognormal distribution [Gibrat, 1931]. Most of the models for growth are stochastic and even though they have different specifications, they are all based on the same kind of main hypothesis. For example, Gibrat's law of proportionate growth (see [Sutton, 1997] for details) is a stochastic model based on the following two principles:

- At each time interval, cities experience a population growth which is proportional to their current population size. Equivalently, the growth rates of cities at each time interval are independent of their population sizes.

- The distribution of growth rates among cities is independent from one time interval to another.

If cities grow according to Gibrat's law, their distribution of sizes tends to a lognormal probability distribution, given by equation (5.5). For the upper tail, i.e. the part of the distribution containing the larger values, the lognormal is very close to a Pareto distribution of the form given by equation (5.4). This explains why Zipf's law [Zipf, 1949], which is a particular distribution with power-law behaviour for discrete data, has also been repeatedly claimed to hold in the case of the distribution of urban population sizes (e.g. [Gabaix, 1999]). Zipf's law is usually known as the 'rank-size rule' and shows that, when plotted on logarithmic scale, the population size of a city is a linear function with slope -1 of its rank in the urban hierarchy. While the Pareto distribution is a continuous probability distribution, Zipf's law can be derived from it by binning the data into ranks. Sornette, who introduced the concept of dragon king, verifies once again that the empirical distribution of the population size of the urban areas in a country is, in general, compatible with Zipf's law [Sornette, 2009].

Zipf's law is accepted as the simplest reference for the distribution of urban population sizes. However, empirical studies show that the actual slope of the rank-size rule can vary a lot. The empirical observations led De Vries [De Vries, 1990] to raise some issues related to the conceptual interpretation of Zipf's law. De Vries warns about "the use of a largely arbitrary norm (the rank-size rule) and the confusion over measurement techniques" and about the fact that Zipf's law is also subject to "the problem of arbitrariness in the delimitation of regions". De Vries' warnings are manifested in Cottineau's 2017 meta-analysis [Cottineau, 2017], where she investigates both technical and topical factors influencing the value of the slope of Zipf's law. She finds that 40% of the variation in the estimated values of the slope can be attributed to technical factors such as the way in which the urban areas are defined or the method used for the estimation of the slope. Additionally, among the topical factors, she points out that those related to the urban process are more relevant in determining the distribution of population sizes than those related to economic development.

The lognormal or the Pareto distributions can then be regarded as very general models for the distribution of urban population sizes, but caution should be exercised when interpreting particular deviations from these models. Especially in the case of long and heavy-tailed distributions, there can be large similarities between observations and a slight better fit of a statistical model should not be a reason to believing more in it [Pumain, 2000]. In fact, as shown in [Cottineau, 2017], deviations from Gibrat's model are frequent, making the non-spatial stochastic model of Gibrat unsuitable to accurately account for empirical distributions of urban population sizes, which are usually subject to clear spatial patterns such as migrations. If Gibrat's model is not suitable to explain the generative mechanism driving urban hierarchies, it follows that any probability distribution that would emerge from this process could also be the 'wrong' one.

Sornette's work shows yet another possible way in which observations might deviate from Zipf's law. In many cases, the population size of the most populous urban area, or sometimes more than one populous urban areas in a country, is actually much larger than expected if Zipf's law were to be satisfied for the whole range of population sizes. Sornette concludes that the observations for the population size of the largest urban areas in a country are outliers of the heavy-tailed distribution used to describe the rest of the urban areas, be it a lognormal or a distribution with power-law behaviour like Zipf's law or Pareto. Lahèrre and Sornette [Laherrère and Sornette, 1998] coined the term 'dragon kings' to refer to these meaningful outliers. Sornette argues that the use of the word 'king' emphasises the fact that these are not the usual outliers, which are frequently discarded as errors due to their supposed spurious nature. Instead, these are important events, which are beyond the extrapolation of the fat tail distribution of the rest of events from the same sample. The word 'dragon' stresses that the events are a "completely different kind of animal, beyond the normal, with extraordinary characteristics, and whose presence, if confirmed, has profound significance".

In the urban context, the statistical concept of dragon king corresponds to that of primate cities, once defined by Jefferson in a more heuristic manner [Jefferson, 1939]. Jefferson remarks that when a city becomes larger than others, this gives it a certain status that, historically, results in amplifying mechanisms for its own growth, distinguishing it from all the other cities. This view is consistent with the statistical one: the fact that dragon kings are outliers suggests that, contrary to the rest of smaller urban areas, they must not satisfy Gibrat's law for growth, but a different growing mechanism. There are several effects which could give some advantage to the largest cities some advantage in the growth process. Firstly, the large population size of a dragon-king city increases the probability of appear-

ance and use of innovations, which will eventually attract more people. Furthermore, because more people live in the largest cities, there are more interactions with the rest of the urban network and so, the largest cities may capture the innovations which come from elsewhere [Pumain, 2000].

The presence of dragon kings in data related to the random variable X, which can represent population size or any other heavy-tailed distributed quantity, can be detected through the application of statistical tests. For example, Pisarenko and Sornette [Pisarenko and Sornette, 2012] introduce two tests that work when X is assumed to follow a particular type of heavy-tailed distribution: the power-law probability distribution. Similarly, Janczura and Weron [Janczura and Weron, 2012] propose a more generalised test, where X can be compatible with any heavy-tailed probability distribution. However, in all these instances, the underlying distribution of X has to be known. Although further statistical tests such as that proposed by Clauset *et al.* [Clauset et al., 2009] can help decide whether a data sample is compatible with a given hypothesised distribution, they cannot guarantee whether the data is actually drawn from that distribution. Identifying the true underlying probability distribution of X is therefore challenging and it is particularly so when the distribution is believed to have a heavy tail. This is partly due to the fact that the number of data points in the upper tail of empirical heavy-tailed distributions is often scarce.

These complications are captured in the successive disagreements in the literature. As mentioned before, some authors such as [Gabaix, 1999] and [Sornette, 2009] claim that the distribution of city population sizes follows Zipf's law, which is a type of power-law. However, Eeckhout finds that this distribution is lognormal for the population size of the US urban areas according to the 2000 census [Eeckhout, 2004], in agreement with the expected distribution for Gibrat's law. Levy, however, argues that,

while the lognormal distribution models correctly most of the observations, the upper tail of the empirical distribution is not accounted for, and is much better described by a power-law Zipfian distribution [Levy, 2009]. To this, Eeckhout replies that Levy's use of log-log plots is unsubstantiated [Eeckhout, 2009]. In [Malevergne et al., 2011], the authors attempt to close the debate by introducing a statistical test between lognormal or power-law distributions, concluding that, in the case of the urban areas in the US, the null hypothesis of a lognormal distribution should be rejected for the upper tail in favour of a power-law.

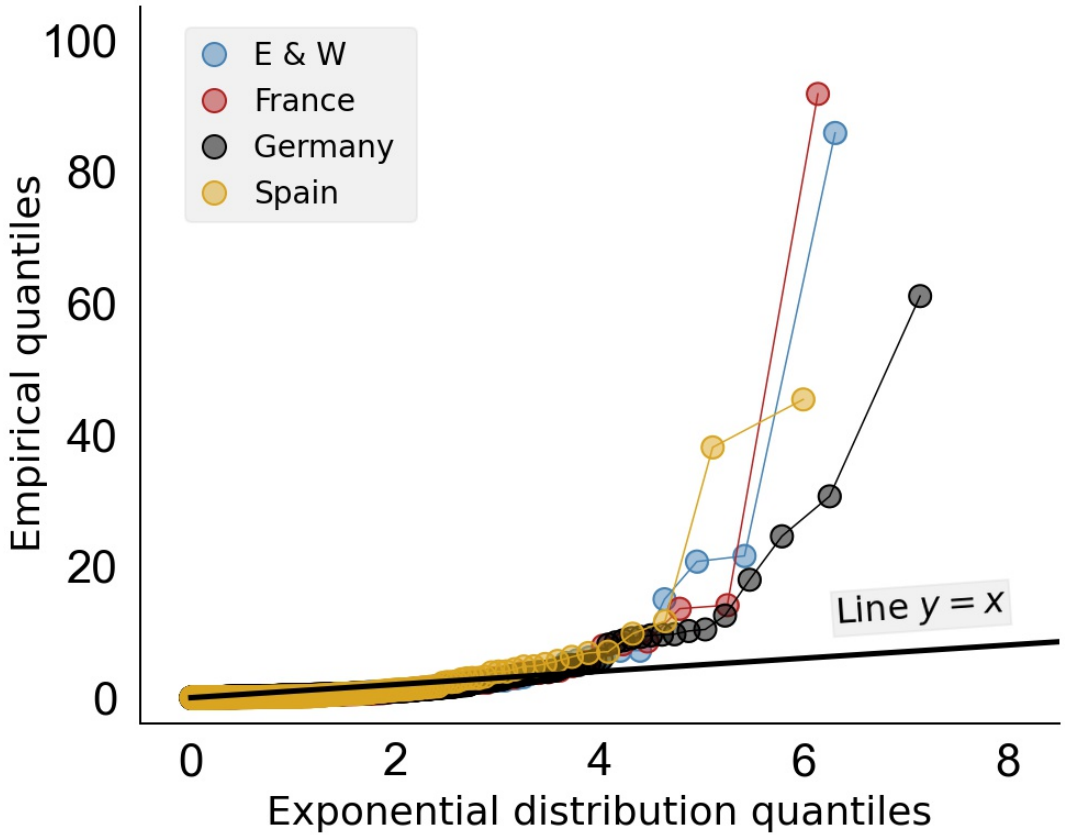


Figure 5.3: Q-Q plot for the distribution of population sizes corresponding to urban areas from England and Wales, France, Germany and Spain, with respect to an exponential distribution.

Figure 5.3 is a Q-Q plot, with theoretical quantiles on the horizontal axis, corresponding to an exponential distribution, and empirical quantiles on the vertical axis, corresponding to the distribution of population sizes for urban areas with more than 15,000 people from E&W, France, Germany and Spain. The parameter of the exponential distribution for the scatter plot corresponding to each country has been set to be the sample average of the data for that country. The solid black line represents the shape that the Q-Q plot would have if the empirical data was exponentially distributed. However, Figure 5.3 shows that the empirical data curves upwards on the right side of the plot. Q-Q plots that show this tendency indicate that the data has a heavier upper tail than the theoretical distribution, in this case, an exponential, i.e. the data exhibits more extreme values than what would be expected if it came from an exponential distribution. As explained in the previous paragraphs, finding the exact distribution of empirical data that follows a heavy-tailed distribution is challenging and sometimes not possible. For this reason, in this thesis, the urban population size data is described as heavy-tailed, but the specific distribution is not investigated. Furthermore, in this particular Chapter, synthetic data will be used for the analysis of the methodology, since this allows for the true distribution of the data to be known.

5.1.3 Probabilistic approach

Statistical tests can determine whether a hypothesis should or should not be rejected with a certain confidence level, but unfortunately, they cannot determine whether it should be accepted. For this reason, in this thesis, the use of synthetic data is considered more appropriate for a discussion in the methodology, given the difficulty in determining the distribution of empirical data regarding city population sizes.

The introduction of dragon kings also motivates the statistical approach taken here. This is manifested by the fact that X is modelled as a random variable which follows a heavy-tailed probability distribution. Variable Y, which could for instance represent the number of road accidents in an urban area, is also modelled as a random variable. The urban scaling model in equation (5.1) hypothesises that Y is related to variable X and consequently, Y's probability distribution must be conditional on X.

Therefore, as Leitão *et al.* [Leitão et al., 2016], the expression from equation (5.1) is interpreted as

$$\mathrm{E}[Y|X] = \alpha X^{\beta}. \tag{5.6}$$

The contribution of this Chapter is mainly methodological: it demonstrates that different estimation methods account for dragon kings in different ways and under certain circumstances, this can result in different estimated values for the scaling parameters. The contribution is also somewhat fundamental since the findings in this Chapter suggest that there might be inherent limitations in the precise estimation of scaling parameters from real world data.

The structure of the Chapter is as follows. First, the precise details of the algorithm to generate samples as well as the methods for the estimation of the scaling parameters are given. Then, the results of the analysis are explained. For these results, samples without dragon kings are used as a null model and are compared to samples with one dragon king. Additionally, samples where a city (the dragon king in particular) follows a different urban scaling model than the rest of the cities in the sample are also analysed. Finally the implications of the results on the precise estimation of scaling parameters from real world data are discussed.

5.2 Data and methods

5.2.1 Generation of random data samples

In essence, the approach taken here involves assessing the discrepancies between the values of the scaling exponents estimated from synthetic data and their known true values. First, samples of $n = 500$ simulated cities are produced, where each city has associated values of X and Y. The population sizes of these cities are randomly generated so that the underlying probability distribution followed by X is a power-law with exponent γ. Hence, the random variable X takes the value x with probability given by

$$\Pr(X = x) = \frac{\gamma - 1}{x_{min}} \Big(\frac{x}{x_{min}} \Big)^{-\gamma} \tag{5.7}$$

for $x \geq x_{min}$. Even though the power-law distribution is defined for continuous random variables, it can still give a good approximation of the distribution of city population sizes provided that x_{min} is sufficiently large [Clauset et al., 2009]. This particular choice of heavy-tailed distribution allows us to later apply the statistical tests proposed by Pisarenko and Sornette [Pisarenko and Sornette, 2012] for detection of dragon kings in the samples.

A variety of probabilistic models can be used to generate the value of Y associated to each city. In this thesis, two cases are explored that typically arise when analysing real data. For the first case, Y is assumed to have a Poisson probability distribution conditional on the value x of variable X. The probability that Y takes the value y is given by

$$\Pr(Y = y | X = x) = \frac{\mu^y e^{-\mu}}{y!}, \tag{5.8}$$

where $y!$ is the factorial of y, parameter μ is equal to the expected value

of the distribution $\mu = E[Y|X = x] = \alpha x^{\beta}$ and the variance is also equal to the expected value. A Poisson distribution is frequently used to model the probability of the number of occurrences of particular events in a fixed time or space interval, if these events occur with a known constant mean rate. However, as a consequence of heterogeneity in real data (e.g. data collected on different weekdays, times of the day, etc.) and the omission of relevant explanatory variables, real data sets tend to display overdispersion, that is, a variance larger than the mean [Hardin and Hilbe, 2007]. In order to account for this overdispersion, a second probability model for the generation of Y is explored here. In this case, Y is modelled through a negative binomial (NB) probability distribution, also conditional on X, which can be parametrised in terms of the mean μ and a parameter r that is related to the degree of overdispersion. Then, Y takes the value y with the following probability conditional on $X = x$

$$\Pr(Y = y|X = x) = \binom{y + r - 1}{y} \left(\frac{\mu}{r + \mu}\right)^{r} \left(\frac{r}{r + \mu}\right)^{k}. \tag{5.9}$$

The mean of the distribution is again $\mu = E[Y|X] = \alpha X^{\beta}$, however, the variance σ^2 now varies with μ according to $\sigma^2 = \mu + \frac{1}{r}\mu^2$.

Below, the specific algorithm to generate samples where the dragon kings obey equation (5.6) is given:

1. Start with $m = 0$ samples without dragon kings and $m = 0$ samples with one dragon king.
2. If $m = 1,000$ the algorithm is completed; otherwise, follow the steps below:
   (a) Draw a random sample of size $n = 500$ for variable X from a power-law probability distribution with parameters $\gamma = 2.2$ and $x_{min} = 15,000$. Denote the n values by x_i, with $i = 1, ..., n$ so that $x_1 \geq ... \geq x_n$.

(b) Transform the sample by taking x_1 and multiplying it by 100. Denote the new x_1 by x_{DK}.

(c) Check whether the original and the transformed random samples contain dragon kings. If there is one dragon king in the transformed sample and zero dragon kings in its original form, proceed to the next steps. Otherwise, discard the original and transformed samples and go back to step 2.

(d) For each value x_i in the original sample, draw an associated value y_i from a probability distribution that satisfies $E[Y = y_i|X = x_i] = \alpha x_i^{\beta}$. The true values of the scaling parameters are set to $\alpha = 0.01$ and $\beta = 1.15$. If y_i is drawn from a negative binomial distribution, set $r = 2$. For the transformed sample, the values of Y are kept the same for $i = 2, ..., n$ but y_{DK} is now conditional on x_{DK}.

(e) Store the values of X and Y corresponding to the original and transformed samples and increase m by one.

(f) Return to step 2.

The package 'powerlaw' [Alstott et al., 2014] is used to generate the power-law distributed values of X in step 2.(a). To check for the presence of dragon kings in step 2.(c), the DK-test and the U-test proposed by Pisarenko and Sornette [Pisarenko and Sornette, 2012] are applied. For the U-test, instead of finding the parameter values through the maximum likelihood method as prescribed by the authors, they are simply set to the true values, i.e. those used in step 2.(a) to generate the data related to X. It is considered that a sample does not contain any dragon kings if (i) the p-values given the U-test are above 0.1 for the top 100 cities, ranked by decreasing population size and (ii) the p-values given by the DK-test are all above 0.1 when the first spacing is compared to a number of spacings

ranging from 1 to 25. It is considered that a sample transformed according to step 2.(b) has one dragon king if (i) the p-values corresponding to the top 100 cities, ranked by population size, are above 0.1 except for the largest city, which must be below 0.1 and (ii) the p-values given by the DK-test are all below 0.1 when the first spacing is compared to a number of spacings ranging from 5 to 25.

As established by Sornette and Ouillon [Sornette and Ouillon, 2012], dragon kings are extreme events that arise from generating mechanisms which are not necessarily active for the rest of the entities under consideration. As a consequence, dragon kings tend to display unique behaviours and in the context of urban scaling models, this may have an impact on the estimated values for the parameters. Indeed, Gomez-Liévano *et al.* [Gómez-Liévano et al., 2020] show that the values of Y associated with extremely populous cities are prone to very large variance and this can result on biased estimators for the scaling exponents. To test these effects, samples where the value of Y associated with x_{DK}, denoted by y_{DK}, deviates from the scaling model satisfied by rest of cities are also generated. Instead, y_{DK} behaves according to a scaling model with a different exponent. This type of samples are generated by using the same algorithm as before but in step 2.(d), y_{DK} is generated ensuring that $E[Y|X] = \alpha X^{\beta^*}$, where $\beta^* = 0.5$. The effect of these deviations associated with dragon kings on the estimation of scaling exponents can be compared with the effect of deviations associated with other cities. In order to do this, this work also considers samples where the value of Y corresponding to a different city of smaller population size is generated according to $E[Y|X] = \alpha X^{\beta^*}$, while for all the other cities $E[Y|X] = \alpha X^{\beta}$ is satisfied.

5.2.2 Estimation of scaling parameters α and β

The values of Y associated to the cities in each sample are generated so that they follow a known distribution conditional on X. When Y follows a Poisson distribution, the Poisson regression must be applied to estimate the scaling parameters. Similarly, when Y has a NB distribution, the negative binomial regression must be applied. However, it can be difficult to ascertain the distribution of Y when real data is considered and this can lead to the wrong choice of regression model. Hence, in order to show how this can affect the estimated values for the scaling parameters, in section 5.3 the Poisson regression is also applied to samples where Y is NB-distributed and the negative binomial regression to samples where Y is Poisson-distributed.

Different sets of $m = 1,000$ samples are considered in section 5.3. For each of these sets, the scaling parameters are estimated via one of the above mentioned regression methods (a total of $1,000$ regressions for each set). This allows us to obtain a distribution for the estimated parameters and hence, provide 90% confidence intervals based on the 5th and 95th percentiles.

5.3 Results

5.3.1 Samples where all values of Y satisfy the same urban scaling model

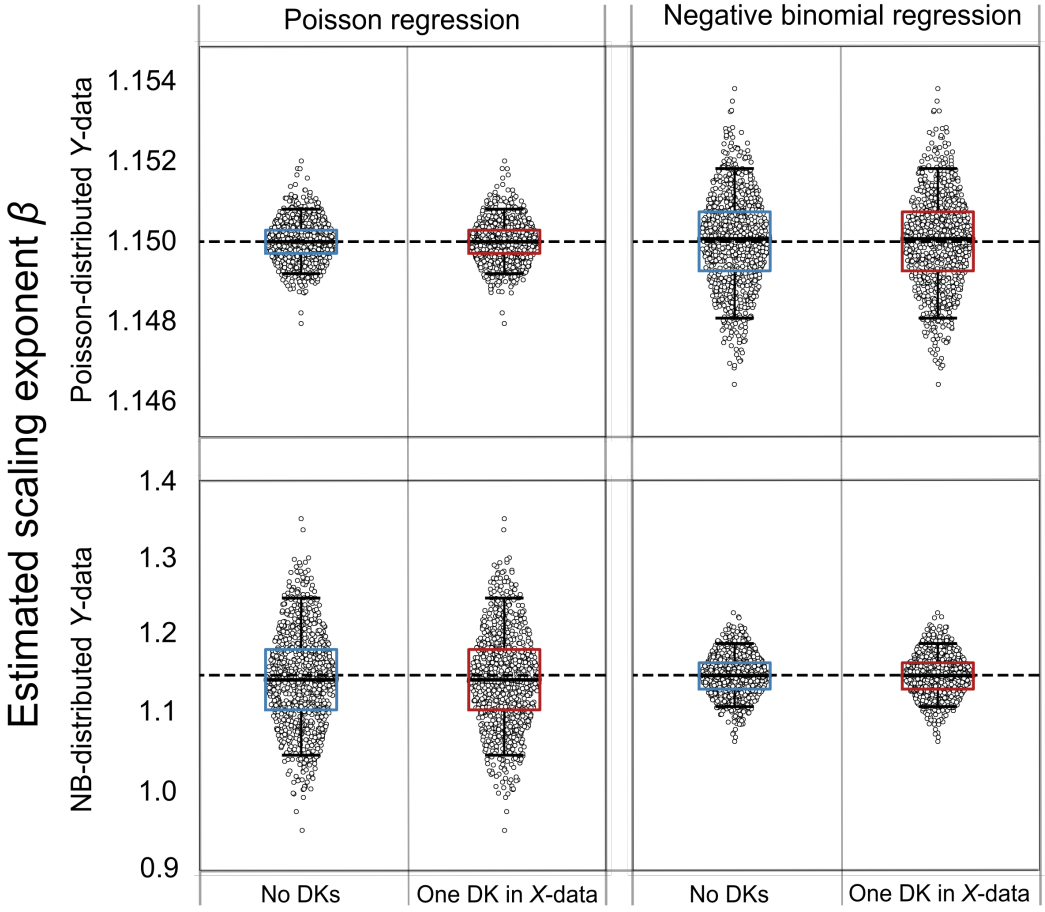


Figure 5.4: Scatter plots and boxplots showing the distribution of estimated values for parameter β from each of the $m = 1,000$ samples with $n = 500$. The bottom and top of the box indicate the first and third quartiles, the middle line in the box is the median and the upper and lower limits correspond to the 5th and 95th percentiles, giving a 90% confidence interval. The horizontal dashed line represents the true value of β, used to generate the synthetic data samples.

Figure 5.4 displays the estimated values of the scaling exponent obtained from performing simulations on different sets of randomly generated samples. Only the estimated exponent is shown here because it generally is the parameter of interest. The samples used to produce the left-most scatter plots and boxplots in each of the four panels do not contain any dragon kings, whereas those used to obtain the right-most scatter plots and boxplots have exactly one dragon king. For each sample, the Y-data was generated so that it satisfies equation (5.6) with $\beta = 1.15$. It is observed that for each probabilistic model and regression method, the distribution of estimated values for parameter β remains pretty much the same for samples with and without dragon kings. This suggests that having a dragon king does not change the results of the scaling analysis as long as $E[Y|X] = \alpha X^{\beta}$ holds for all cities.

The results corresponding to the top-right and bottom-left panels in Figure 5.4 show the effects of applying a regression method that assumes a probability model for Y other than the actual distribution followed by the Y-data. For example, if a negative binomial regression is applied to samples with Poisson-distributed Y, this results in a distribution for the estimated values of β which is a lot more spread than when the Poisson regression is used. Applying a Poisson instead of a negative binomial regression when the Y-data is NB-distributed has a similar effect.

From the bottom two panels in Figure 5.4 it can be learned that, not surprisingly, for both types of regression, the distribution of estimated values of β when the Y-data is NB-distributed is a lot wider than when the Y-data is Poisson-distributed (note the different scale of the vertical axis for the top and bottom rows). Then, it can be concluded that appropriate knowledge of the distribution of Y can help us obtain more precise estimates for the parameters of the urban scaling model that relates X and Y.

5.3.2 Samples where not all values of Y satisfy the same urban scaling model

Figure 5.4 shows that, as long as X and Y satisfy the same urban scaling model for all the cities in each sample, the presence of dragon kings does not have an impact on the estimated scaling exponent, regardless of the type of regression used for the estimation. Figure 5.5 shows the situation where the value of Y corresponding to a specific city in each sample satisfies an urban scaling model with exponent $\beta^* = 0.5$ instead of $\beta = 1.15$, as the rest of cities in the sample. The boxplots are obtained by estimating the scaling exponent from $m = 1,000$ samples. The first pair of boxplots on both the top and bottom panels, highlighted, correspond to samples with one dragon king where all the values of Y satisfy equation (5.6). The rest of boxplots correspond to samples where y_{DK} or y_i for $i > 2$ have been generated so that they satisfy $E[Y|X] = \alpha X^{\beta^*}$, with $\beta^* = 0.5$.

It is observed that the largest discrepancy with the true parameters corresponds to the case where the dragon king does not satisfy equation (5.6). The Poisson regression is particularly sensitive to the effect of the dragon king. This is especially the case when Y is Poisson-distributed. The negative binomial regression is not as sensitive to the effect of the dragon king, giving estimated values of the scaling exponent which are much closer to the true one. Generally, if a city with large population size deviates from the urban scaling model followed by the rest of cities, the effect on the estimated scaling exponents is more significant. The larger the city's population size, the more this tendency becomes apparent.

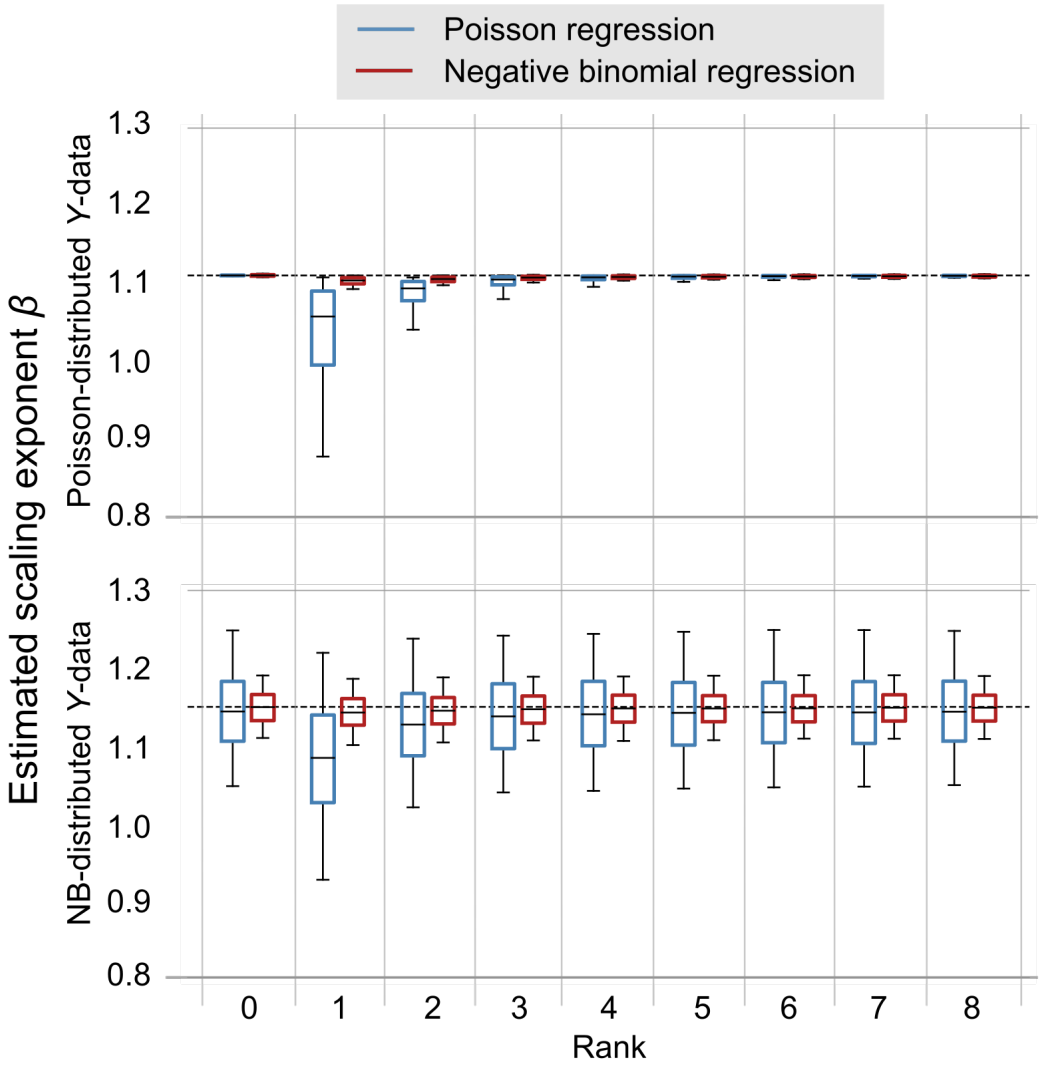


Figure 5.5: Boxplots showing the distribution of estimated values for parameter β from each of the $m = 1,000$ samples with $n = 500$. The X-data in the samples contains a dragon king. The horizontal axis represents the value of Y that has been generated according to either a Poisson or a NB distribution with $E[Y|X] = \alpha X^{\beta^*}$, where $\alpha = 0.01$ and $\beta^* = 0.5$. The rest of values of Y are generated according to a Poisson or a NB distribution with $E[Y|X] = \alpha X^{\beta}$, where $\beta = 1.15$. The vertical axis represents the values of the estimated scaling exponents. The bottom and top of the box indicate the first and third quartiles, the middle line in the box is the median and the upper and lower limits correspond to the 5th and 95th percentiles, giving a 90% confidence interval. The horizontal dashed line represents the true value of β, used to generate the synthetic data samples.

5.4 Discussion

From the results presented above, it can be concluded that the presence of dragon kings in the X variable does not affect the value of the estimated exponent β as long as Y satisfies the same urban scaling model for all the cities under consideration.

However, dragon kings tend to display a different behaviour from other cities so in real scenarios, Y might not necessarily follow the same urban scaling model as the rest of cities. If this happens, the estimated values of the scaling parameters given by the Poisson regression will be further from the true values than those given by the NB regression, since the former regression method gives more importance to the largest data points. The Poisson or negative binomial regressions involve maximising a likelihood function via the IWLS method. This, in turn, involves successive iterations of a weighted least squares problem [Ver Hoef and Boveng, 2007]. The weights assigned to different data points dictate their relative importance in determining the estimated parameters at each iteration of the IWLS method. Verhoef and Boveng [Ver Hoef and Boveng, 2007] show that the weight corresponding to the city with population x_i in the jth iteration is $w_i^j = \mu_i^j/(1 + \frac{1}{r}\mu_i^j)$. For the Poisson regression, this weight is $w_i^j = \mu_i^j$. The weights in the Poisson regression increase in proportion to μ_i^j, while the weights in the negative binomial regression tend to r for larger values of μ_i^j. Cities with larger values of X, such as the dragon kings, have corresponding larger expected values of Y and will therefore receive relatively more weight in the Poisson regression.

The findings of this Chapter, as published in [Cabrera-Arnau et al., 2020], raise the following dilemma. Regression methods that place relatively less weight on the larger city population sizes, such as the negative binomial regression, will produce estimates of the scaling parameters which

are closer to the true values. But, when it comes to cities, the interest is often in the really large fast-growing cities that host a significant percentage of a country's population. If the tendency is for large cities to evolve towards the dragon-king status, then urban scaling models are not a good model, as there is an inherent flaw in the parameter estimation procedure.

In conclusion, if estimation techniques that give more weight to the larger entities are used, then the estimated exponents may be invalid unless the same urban scaling model is satisfied by all the cities in the urban system. But, if an estimation technique is used that does not give as much weight to these 'outlier' cities which are prone to deviate from the mainstream behaviour, then the most interesting part of the urban scaling model risks being neglected.

Chapter 6

Relationship between road accidents and population size in functional urban areas from England and Wales, France, Germany and Spain

The work presented in this Chapter has been partially covered in the research paper entitled 'Urban population size and road traffic collisions in Europe' [Cabrera-Arnau and Bishop, 2021].

6.1 Introduction

As discussed in Chapter 4, road accidents are among the most common causes of death for certain age groups [Global Health Data Exchange, 2019]. At a worldwide level, approximately every 24 seconds someone dies as a consequence of a road accident [World Health Organization, 2018]. Besides the enormous emotional burden that each of these deaths leaves behind,

they also lead to significant financial losses. For example, in Great Britain it is estimated that the average cost of an accident in the year 2019 is above £100k ($140k), although for fatal accidents, this figure could be as high as £2.2M ($3M) [Department for Transport, 2019].

Much like wealth, road accidents are not uniformly distributed across regions. At a global scale, road accident death rates (measured as the number of deaths per 100,000 people that are caused by road accidents in one year) in low- and middle-income countries are about twice as large as in high-income countries (averages of 21.5 and 19.5 vs 10.9 per 100,000 population) [World Health Organization, 2018]. Data from 2019 [World Bank Group, 2019], depicted in Figure 6.1, shows that all the countries with road accident rates of more than 20 deaths per year per 100,000 people have a gross national income (GNI) per capita of less than 20,000 dollars, with the exception of Saudi Arabia. However, all the countries with a GNI per capita above 40,000 dollars have a road accident death rate of less than 10 per 100,000, except for the United States.

At a national scale, road accident fatalities are also unevenly spread. Rural areas have higher death rates, but most accidents actually take place in urban areas [Zwerling et al., 2005; Cabrera-Arnau et al., 2020].

Concurrently, the world is undergoing a rapid urbanisation process. As it can be learned from Figure 1.1 in Chapter 1, since 2005, more than 50% of the people live in urban areas, with this figure increasing year after year. It is estimated that by 2050, more than two thirds of the global population will then live in urban areas, with the percentage reaching 74% in Europe as of 2018 [Population Division of the UN Department of Economic and Social Affairs, 2018].

Given that most road accidents take place in urban areas and the population size of these urban areas is likely to increase due to urbanisation, the following question arises: does the number of road accidents increase with

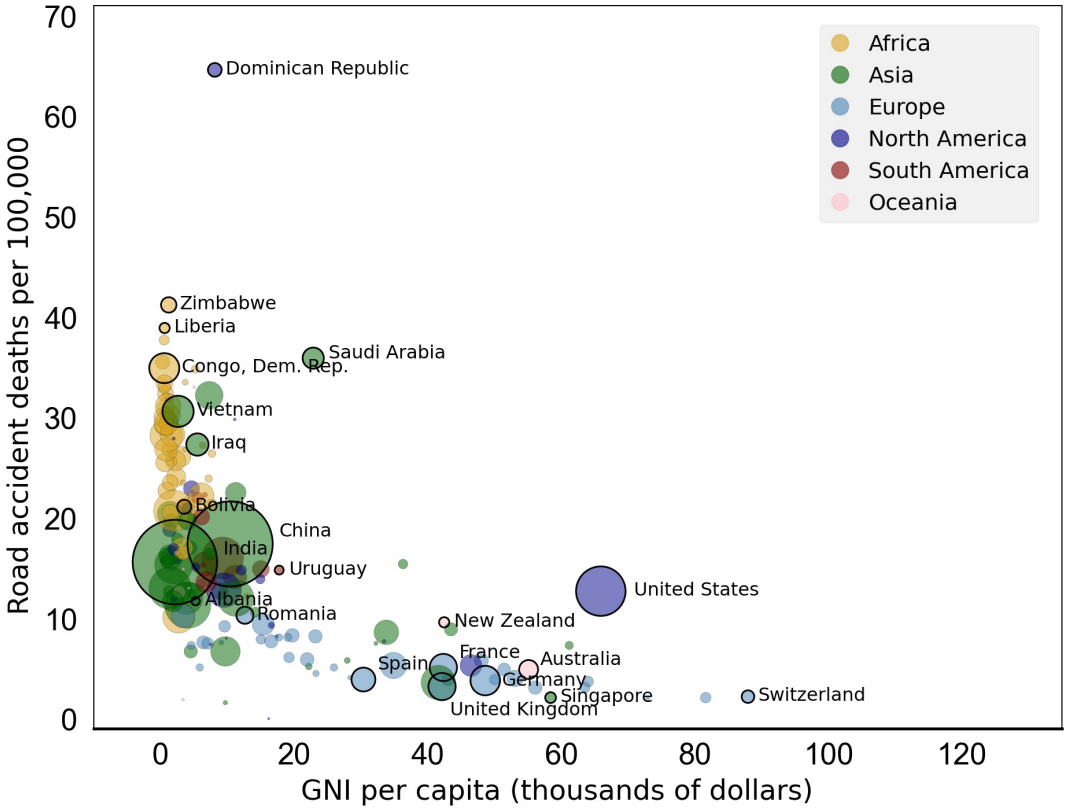


Figure 6.1: Number of road accident deaths per 100,000 vs. GNI per capita in all the countries of the world with available data. Data retrieved from [World Bank Group, 2019].

the population size of the urban area where they take place? As shown in [Louf and Barthelemy, 2014], traffic congestion increases in urban areas of larger population size and more traffic congestion leads to more opportunities for an accident. Additionally, traffic congestion can increase stress levels in drivers [Hennessy and Wiesenthal, 1997; Wener and Evans, 2011], and this can also lead to a greater risk of accident [Simon and Corbett, 1996; Taylor and Dorn, 2006]. Due to these factors, an affirmative answer to the question above could be postulated. However, previous research on the issue of whether traffic congestion has an impact on road accident rates has reached conclusions that might seem counterintuitive. For example, Shefer demonstrates, in a hypothetical situation, that a reduction in the

level of congestion could inadvertently cause an increase in road fatalities [Shefer, 1994]. However, Shefer only considers fatal accidents. Other works consider the total number of road accidents and reach different results. For instance, the authors in [Sun et al., 2016] analyse all the road accidents recorded by Shanghai Expressway Surveillance System in a three-year period and find that traffic exposure, congestion and merging behaviors all increase the risk of accidents on urban expressways. They also find that the risk factors are different in congested and non-congested flows. Despite the wealth of published research on the topic, our current understanding of how traffic congestion affects accident risks is still limited. As it has been reported in the review by Retallack and Ostendorf [Retallack and Ostendorf, 2019], the dominant result in the literature is a positive linear relation between the number of road accidents and levels of congestion/traffic volume. But Retallack and Ostendrof also mention some works that analyse finer temporal resolution traffic data and show a U-shaped relationship.

The aim here overlaps with the aim in Chapter 4: to analyse the direct impact of urban population size on the incidence of road accidents. The analysis is also based on data from England and Wales, France and Spain, but this time, with the addition of data from Germany. In order to achieve the aim, the population and number of road accidents corresponding to the urban areas under consideration have to be determined. However, there is no single way of establishing the boundaries of urban areas [Batty and Ferguson, 2011; Arcaute et al., 2015] and different criteria are often chosen according to the type of analysis to be performed. Similar to the approach taken in [Arcaute et al., 2015], a classification based on commuting flows is used in this Chapter. Based on these classification criteria, cities and towns that have traditionally been considered as different entities, may be classified as the same urban area.

6.2 Methods

The results in this Chapter are obtained from processing geographic and road accident microdata from England and Wales, mainland France, Germany and mainland Spain. England and Wales are two countries but, for ease of notation, they are referred to as only one entity denoted by E&W. Similarly, mainland France, Germany and mainland Spain will be simply referred to as the countries France, Germany and Spain.

6.2.1 Distribution of urban population sizes

In the Results section, there is a distinction between the behaviour of road accidents in two types of urban areas from each of the four countries of interest: the largest urban areas and the rest of smaller urban areas. Hence, it is considered necessary to give here a brief overview remarking the patterns displayed by the distribution of population sizes corresponding to the urban areas in E&W, France, Germany and Spain.

Urban population sizes have been found to follow heavy-tailed distributions, such as a power-law distribution [Gabaix, 1999; Sornette, 2009; Levy, 2009] or a lognormal [Eeckhout, 2004]. However, in practice, the population size of the largest urban area in a country is often larger than predicted by the underlying heavy-tailed distribution. These extremely large urban areas then become meaningful outliers and are sometimes referred to as dragon kings, a term coined by Lahèrre and Sornette in [Laherrère and Sornette, 1998]. Additionally, they have a special socioeconomic status forged by amplifying mechanisms for their own growth.

London and Paris would be examples of urban areas displaying dragon-king features. Their size is several times larger than the next largest urban area in their respective country and they are also primary nodes in the global socioeconomic network. Germany and Spain, however, are countries

which have experienced a higher degree of territorial divide throughout history and where different cities have been appointed as the capitals at different periods in time. As a result, these countries have more than one urban area with an unexpectedly large population size and with a central role in the socioeconomic landscape of the country. In Germany, there are actually many urban areas that fulfil these characteristics, in particular, the 'Big Five' metropolitan regions (Berlin, Hamburg, the Rhine-Ruhr metropolitan region, Frankfurt and Munich); in Spain, Madrid and Barcelona.

6.2.2 Population and number of road accidents in the urban areas

The urban areas used here are the functional urban areas (FUAs) established by Eurostat [Eurostat, 2021*a*], which are based on commuting flows [Eurostat, 2021*b*]. The data corresponding to E&W, France and Germany is from 2018 and, in the case of Spain, from 2015. This choice of years is due to the fact that, at the moment when this research was performed, those were the years with the most recent data in each of the countries under study.

The data corresponding to the small geographical hierarchies of each country is aggregated into urban areas and analysed further to produce the figures in the forthcoming sections. In the case of France, Germany and Spain, both population and road accident data is collected by the local administrative unit (LAU). LAUs have different names in different countries: *communes* in France, *gemeinden* in Germany and *municipios* in Spain. In E&W, data is available for lower level geographic hierarchies known as Lower Layer Super Output Areas (LSOAs), designed specifically to improve the reporting of small area statistics. However, urban areas may extend over several of these small geographic hierarchies. For example,

the urban area corresponding to Greater London would comprise 6,908 LSOAs while the urban area corresponding to Madrid would comprise 182 *municipios*.

Information related to population [Office for National Statistics, 2018*b*; Institut National de la Statistique et des Études Économiques, 2018*c*; Statistisches Bundesamt, 2018*b*; Instituto Nacional de Estadística, 2018] as well as the shapefiles for the LSOAs [Office for National Statistics, 2018*a*], the LAUs and the FUAs [Eurostat, 2021*a*] are publicly available for download. In the case of E&W, France and Spain, data bases can be downloaded where each entry is a recorded road accident [Department for Transport, 2018; Observatoire National Interministériel de la Sécurité Routière, 2018; Dirección General de Tráfico, 2015]. For each accident, the LSOA or LAU where it took place is specified. In the case of Germany, a data base is used where the accidents are already aggregated by the LAU [Statistisches Bundesamt, 2018*a*].

It is possible to make country-to-country comparisons of patterns that emerge as a result of considering a country's urban system as a whole. However, a word of caution needs to be said about the comparability of population data corresponding to urban areas from different countries. The population in an urban area is computed as the sum of populations corresponding to the small geographical hierarchies that lie within the boundary of the urban area. However, these geographical hierarchies are country-dependent and, except in the case of the LSOAs in E&W, are also subject to historical agreements.

Similarly, for all the countries under consideration, the number of accidents taking place in each urban area over a year is obtained as the sum of the number of accidents over that year in each small geographical hierarchy that lies within the urban area's boundary. But definitions as to what constitutes an accident may also vary. For example, the accidents recorded

in France are those that required some form of medical treatment [Observatoire National Interministériel de la Sécurité Routière, 2018], whereas in E&W, all reported accidents incurring personal injury, but not necessarily requiring medical care, are included in the national data base [Department for Transport, 2017*a*]. Furthermore, every country has different levels of under-reporting of data, especially when it comes to non-fatal accidents. Data related to hospitalisation as a result of an accident, surveys (e.g. National Transport Survey in Great Britain) and insurance compensation claims all indicate a higher number of casualties than are reported [Department for Transport, 2017*a*]. Hence, 300 accidents per 100,000 people in an urban area from E&W does not quite mean the same as in a French, German or Spanish urban area. It is for this reason that in Figures 6.5 and 6.6, the colour key is based on the percentage difference between the number of accidents per person in a given urban area and the corresponding country's average number of accidents per person in the urban areas under consideration.

6.2.3 Urban scaling models

Since Smeed's 1949 pioneering work regarding statistical aspects of road accidents [Smeed, 1949], the precision and availability of both geographic and road safety data have improved considerably, enabling many other authors to expand the field [Andreassen, 1985; Baker et al., 1987; Whitelegg, 1987; Andreassen, 1991; Erdogan et al., 2008; Anderson, 2009; Kumar and Toshniwal, 2016; Prieto Curiel, González Ramírez and Bishop, 2018]. Additionally, the more recent introduction of scaling models in the context of urban science [Bettencourt et al., 2007] offers a new avenue for modelling road accidents and understanding their behaviour.

Urban scaling models are based on the hypothesis that a quantifiable property Y varies with city population size X according to

$$Y(X) = \alpha X^{\beta} \tag{6.1}$$

with scaling parameters α and β. According to the value of the scaling exponent β, the scaling model can display three types of behaviour. If $0 < \beta < 1$, Y is said to grow sublinearly with X. Sublinear behaviour implies that the value of Y per person decreases with city population size. If $\beta = 1$, the scaling is linear and the values of Y per person are constant across city population sizes. If $\beta > 1$, Y scales superlinearly. When that is the case, the values of Y per person increase with city population size.

Scaling models have been applied widely (see e.g. [Bettencourt et al., 2010; Prieto Curiel et al., 2019]), in particular, urban scaling models were used in Chapter 4 to describe the relationship between the number of accidents of different degrees of severity and the population size corresponding to the set of 'built-up' areas (defined by a land-use classification criterion) from England and Wales, France and Spain.

In this Chapter, the analysis is extended to data from France, Germany and Spain as well as England and Wales and instead of the built-up areas, functional urban areas defined according to commuting flows (see Chapter 2) are considered.

Figures 6.2, 6.3 and 6.4 have been generated in order to, firstly, illustrate how the locations of road accidents are concentrated around built-up areas and secondly, show how the built-up areas considered in Chapter 4 and the functional urban areas considered in the current Chapter differ from each other. Figure 6.2 represents the locations of road accidents that took place in E&W during the year 2018. For visualisation purposes, only 20,000 randomly chosen accidents, which represent around 20% of the total figure,

are depicted in Figure 6.2. These accidents are concentrated in built-up areas, as it can be learned from comparing Figure 6.2 with Figure 6.3, where the latter shows the built-up urban areas in the UK. In Figure 6.4, the FUAs are depicted. The definitions for the FUAs are less restrictive than for BUAs, hence the larger patches.

An advantage of using urban scaling models is that they allow us to summarise the relative performance of cities across a vast range of population sizes under the same mathematical model. However, certain urban areas (frequently the largest ones in a region) are unique in that they play central roles in economic productivity of firms and workers [Puga, 2010], are especially prolific in certain industry sectors or have an extraordinary cultural output [Scott, 1997]. For this reason, it has been questioned [Arcaute et al., 2015] whether these urban areas, sometimes referred to as dragon kings, should be analysed alongside the rest or whether on the contrary, they should be considered as a separate category.

Here, urban areas displaying dragon-king features are included in the analysis. Following the results in [Cabrera-Arnau and Bishop, 2020], a negative binomial regression is used for parameter estimation since it places less weight on larger urban areas, hence making the parameter estimation procedure more robust with respect to observations associated with large urban areas. Further, the performance of the negative binomial regression is tested against a Poisson regression through the Akaike Infromation Criterion (AIC). For each country, the former method yields a lower value of AIC, indicating that the negative binomial regression is a model of higher quality than the Poisson regression for the data sets of interest.

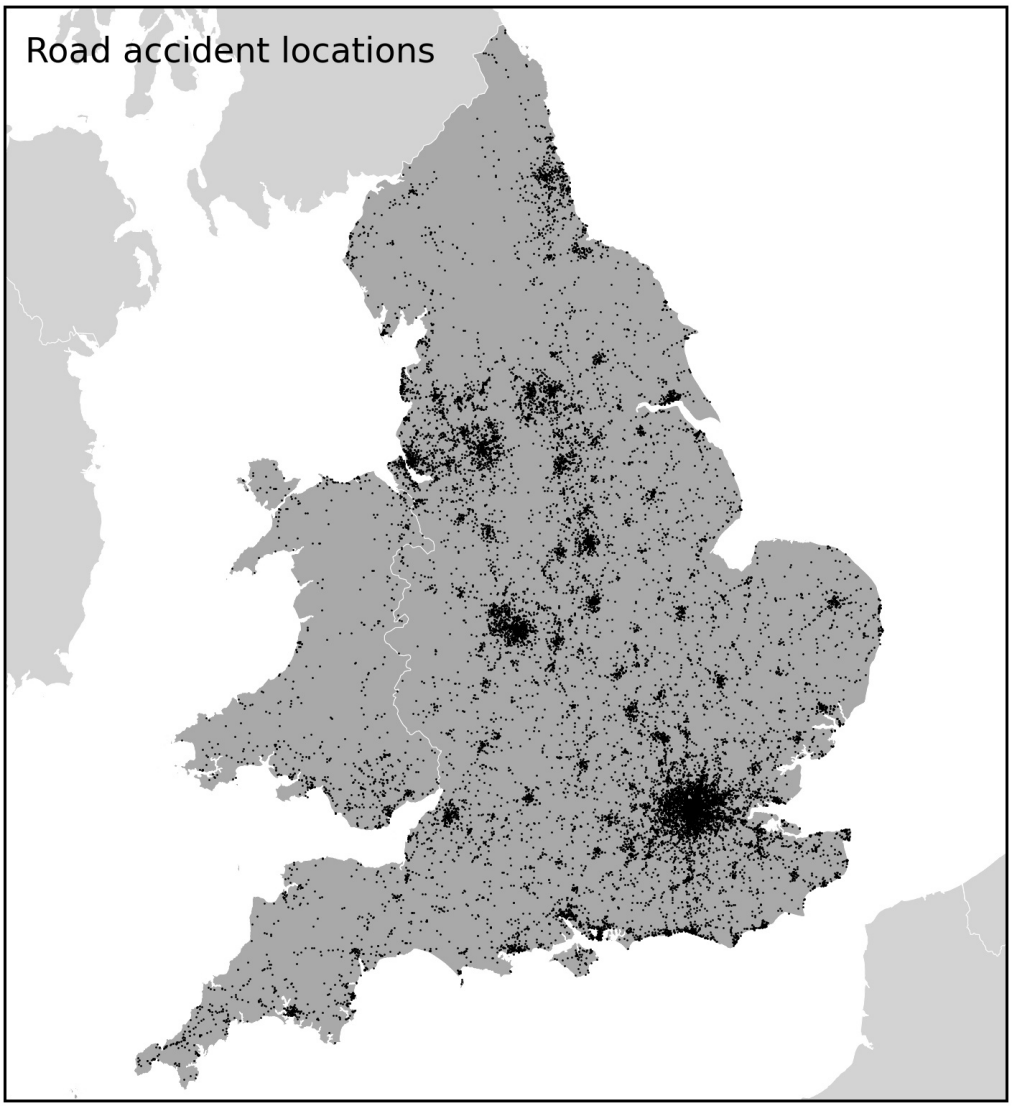


Figure 6.2: Locations of 20,000 randomly selected road accidents from England and Wales in 2019.

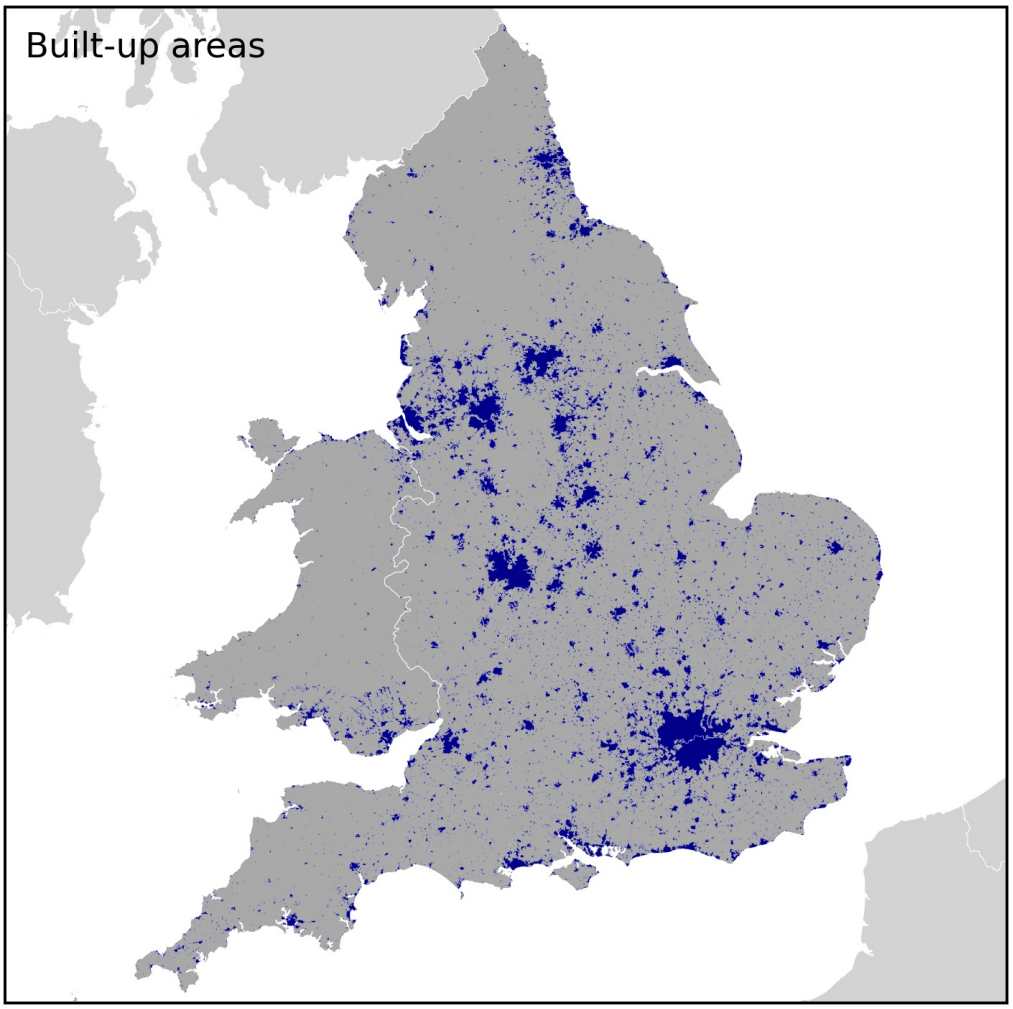


Figure 6.3: Built-up areas in England and Wales.

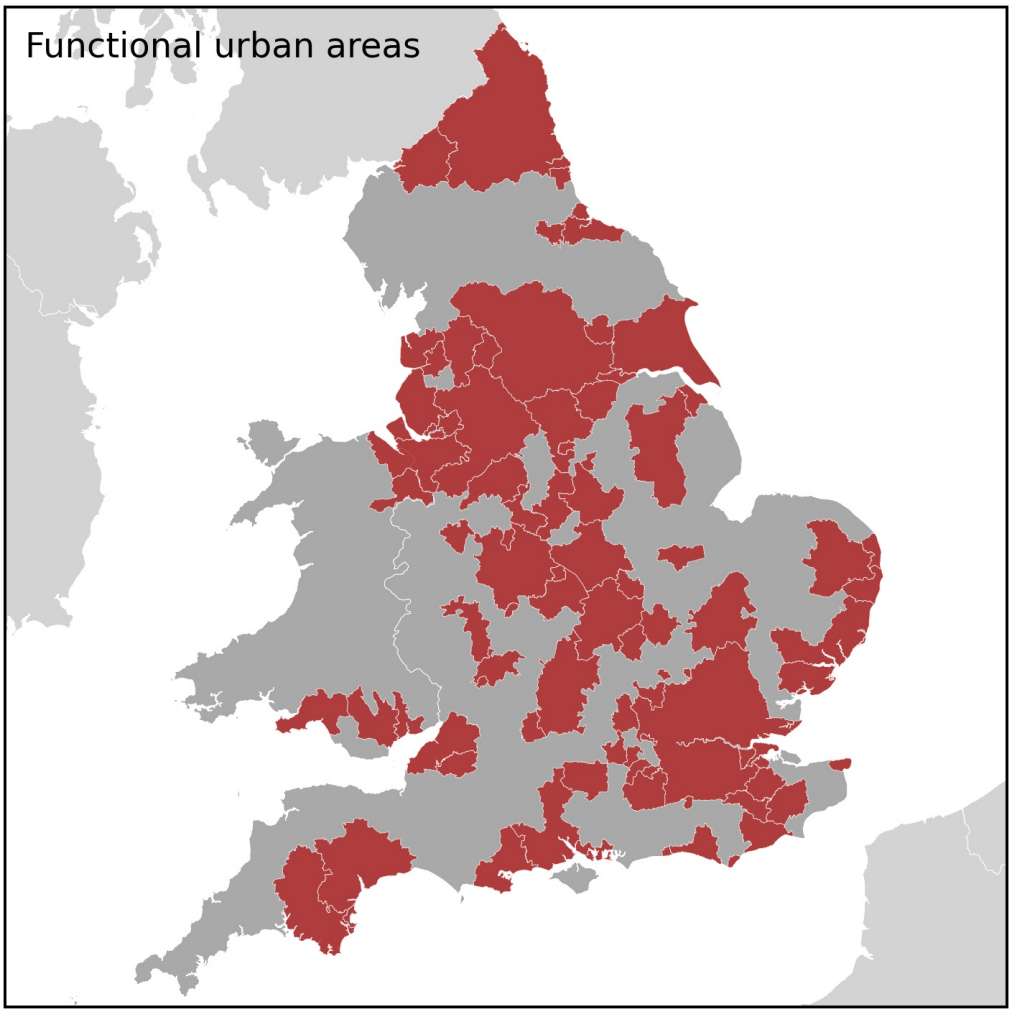


Figure 6.4: Functional urban areas in England and Wales.

6.2.4 Is the scaling behaviour significantly different from linear?

Answering the question of whether the scaling behaviour is significantly different from linear is equivalent to checking if the scaling exponents are significantly different from 1. To do this, a statistical test for significance based on a Monte Carlo simulation is performed. Let us assume the null hypothesis that the data corresponding to country A comes from a scaling model with parameter $\beta_0 = 1$. The parameters of the original sample, denoted as $\hat{\alpha}_A$ and $\hat{\beta}_A$, are also estimated. In each iteration i of the simulation, the steps below are followed:

- The populations of the urban areas from country A are kept the same as in the original sample.
- Then, random values for Y are generated distributed according to a negative binomial distribution with mean $\mu = \hat{\alpha}_A X^{\beta_0}$ and variance $\sigma^2 = \mu + \mu^2$.
- Finally, the value of the scaling exponent $\hat{\beta}_A^i$ is estimated corresponding to the sample generated in the ith iteration and store it.

Once the simulation process is completed, there will be an estimated value of the scaling exponent corresponding to each iteration. Then the p-value is computed as the proportion of stored values which satisfy $|\hat{\beta}_A^i - \beta_0| > |\hat{\beta}_A - \beta_0|$. If the p-value is smaller than a chosen significance threshold of 0.05, the null hypothesis can be rejected.

Applying 2,000 iterations of this method to the four countries under consideration, the following p-values are obtained: $p_{E\&W} = 0.93$, $p_{FR} = 0.55$, $p_{DE} = 0.61$ and also $p_{ES} = 0.81$ for E&W, France, Germany and Spain respectively, which are all above the chosen level of significance.

6.3 Results

6.3.1 Geographical distribution of road accidents in urban areas

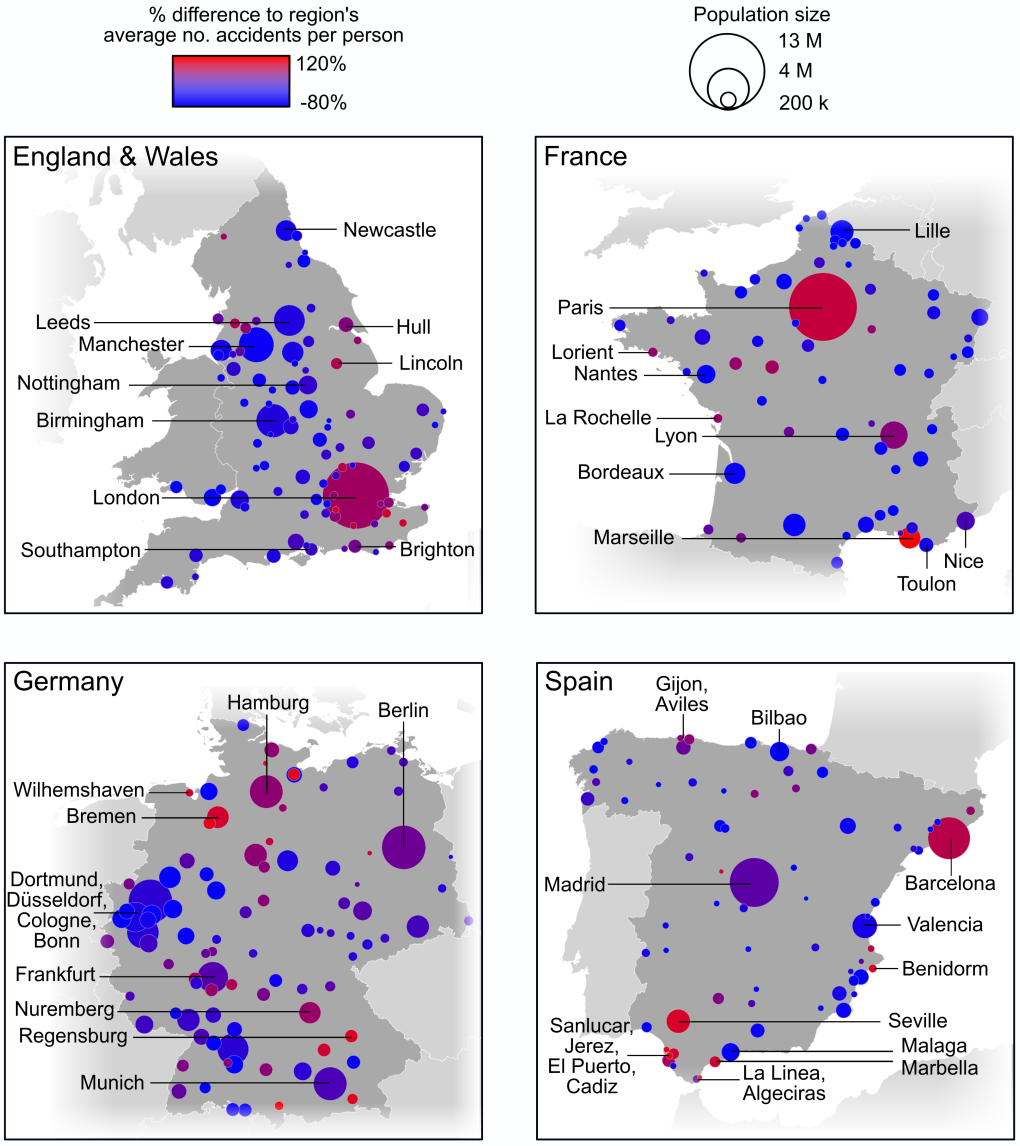


Figure 6.5: Map representation of the number of road accidents per person in urban areas from England and Wales, France, Germany and Spain.

In Figure 6.5, the population and number of accidents per person corresponding to the urban areas from E&W, France, Germany and Spain are plotted. A map layout is chosen for this Figure, as this helps with visualisation and understanding.

Largest urban areas

It can be seen in Figure 6.5 that the largest urban areas in E&W and France (London and Paris) stand out in terms of their large population size and high number of accidents per person. This is not the case for the German and Spanish counterparts, Berlin and Madrid. In Germany, the urban areas are more evenly spread across the whole range of population sizes, and so is the number of road accidents. In Spain, there are two urban areas, instead of just one, that stand out for their population size: Madrid and Barcelona. The number of road accidents is relatively high value in both population size and number of accidents. The fact that the four countries display different patterns is perhaps not so surprising, considering that their urban areas have been subject to unique historical developments. More details regarding this observation are provided in the Discussion and Conclusions section.

Other urban areas

The number of road accidents in smaller urban areas displays a high variability in all the countries. As a consequence, it cannot be discerned, *a priori*, whether urban population size plays a role in determining the number of road accidents for these urban areas. The fact that there is such degree of variability is an indication that there might be variables other than urban population size which affect the number of road accidents. For example, in E&W, Sheffield and Stevenage have similar road accident rates of approximately 175 per 100,000 people, however, the functional urban area

corresponding to Sheffield has a population size of 1.3 million, whereas the one corresponding to Stevenage has a population of just above 100,000 people. Similarly, Bremen and Willemshaven in Germany have a road accident rate of around 500 per 100,000 people per year, but their population sizes are also very different: Bremen's functional urban area has a population size of 1.4 million, whereas Willemshaven has a population size just below 175,000.

6.3.2 Scaling of accidents in urban areas

After inspecting individual cities in the previous section, a natural question arises as to whether the population size of a city has an effect on the number of accidents. To answer this question, we firstly propose the scaling hypothesis, which assumes that the number of road accidents Y in a given urban area is determined by its population size X according to an urban scaling model of the form $Y = \alpha X^{\beta}$. The two parameters associated with this model, α and β, can be estimated from the data. If the parameter β, known as scaling exponent, is found to be significantly larger than 1, then this gives an indication that the number of road accidents per person in an urban area increases with population size. In order to estimate the scaling exponent, the considerations from [Leitão et al., 2016] and [Cabrera-Arnau and Bishop, 2020], where the authors emphasise the need to account for the statistical properties of the data, are taken into account. Here, this is done by using a generalised linear model for regression. Details about this approach and more background about urban scaling models are provided in section 6.2 of this Chapter.

Figure 6.6 shows the data related to the urban areas in the four countries of interest as well as the scaling model that provides the best fit to the data, with 95% confidence intervals obtained by bootstrapping. In E&W and Germany, the estimated scaling exponent $\hat{\beta}$ has been calculated to be

slightly below 1, while in France and Spain, it is slightly above 1. However, in the methods section β is shown to be not significantly different from one ($p > 0.05$) in all four countries, hence indicating that, for the definitions of urban areas and road accidents used here, there are no significant effects of urban population size on the number of road accidents.

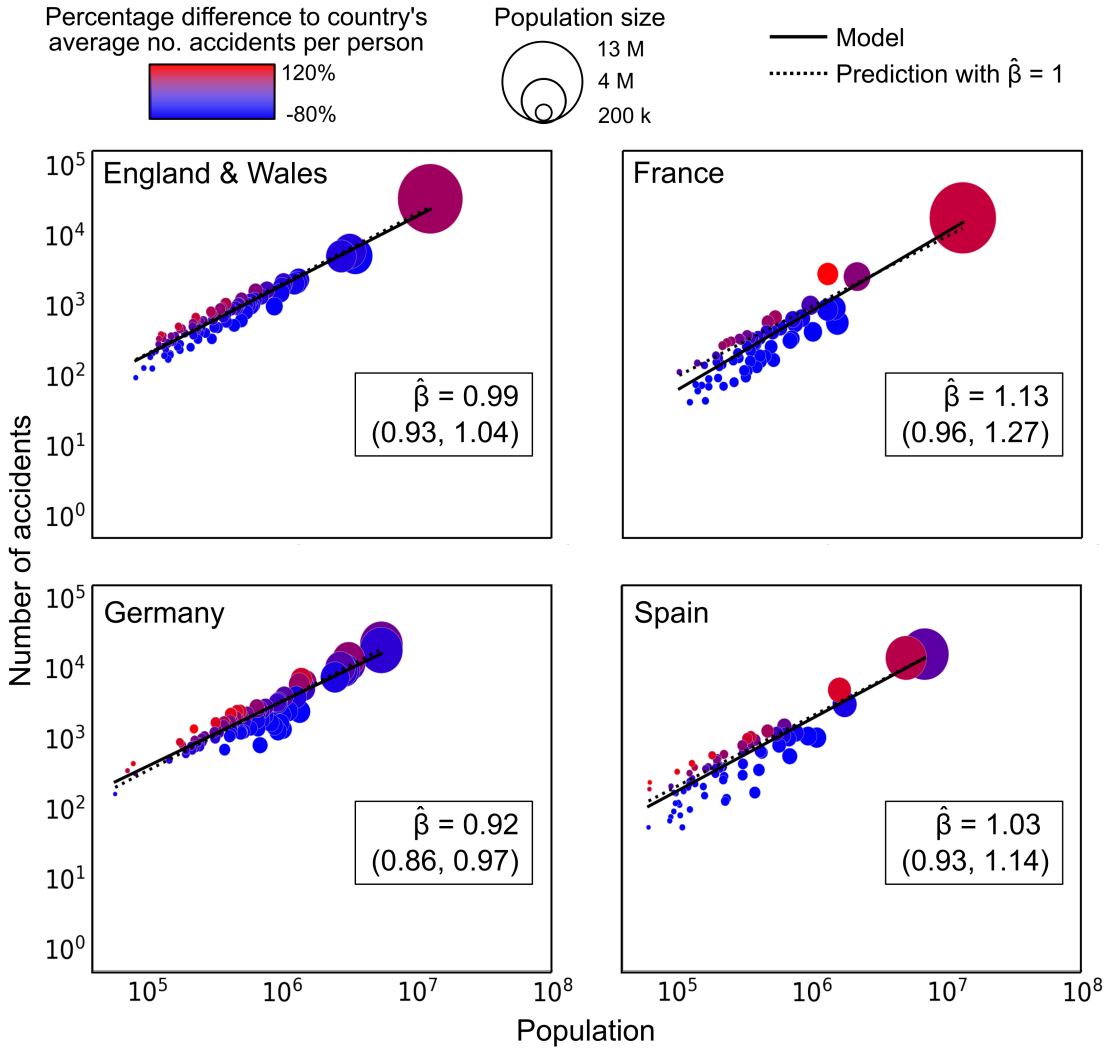


Figure 6.6: Urban scaling models corresponding to England and Wales, France, Germany and Spain.

In Figure 6.6, the high variability in the number of road accidents for urban areas of a given population size is perhaps even more evident. This is an indicator that, quite possibly, there are more variables influencing the number of road accidents apart from urban population size. If this is the case, urban scaling models should be replaced for other models that incorporate these additional variables.

6.4 Discussion and conclusions

It can be concluded that urban population size has no significant effect on the number of road accidents in urban areas from four European countries. This conclusion is based on the results obtained through the application of urban scaling models, which uncover patterns that emerge at a country-wide level.

These findings are in contrast with the results in Chapter 4, where urban scaling models were applied to describe the relation between the number of road accidents and the population of built-up areas from E&W (defined according to a land-use criterion). In Chapter 4, it was found that the number of road accidents scales superlinearly with urban population size. The discrepancies between the results are due to the fact that different urban areas are considered here and a different regression method for the estimation of parameters.

The behaviour of large cities, sometimes called dragon kings, is difficult to model due to their unique characteristics [Sornette, 2009; Arcaute et al., 2015]. Therefore, following the results from Chapter 5, a generalised linear model is chosen here for the estimation of the scaling model parameters that accounts for the dragon kings' unpredictability. This choice of generalised linear model assumes wider probability distributions for the number of accidents as the population size of the urban areas increases.

Turning the attention to individual urban areas, there are cases where the road accident rate is remarkably higher than the national average for a given population size. This tendency could be the result of the fact that other variables (volumes of traffic; traffic congestion; proximity to a port) may be playing a key role in determining the number of accidents. These variables are not analysed in this thesis but could form the basis for future work. In [Keuschnigg et al., 2019], the authors provide a possible approach to understanding the michromechanisims underlying the collective urban scaling behaviours. In particular, their analysis is concerned with the extent to which the increased social interconnectivity in cities is responsible for the urban wage premium. A similar methodology could be adapted to the context of road accidents in order to find more detailed information about the different drivers of urban scaling patterns.

In particular, two types of behaviour displayed by individual urban areas are highlighted here: that corresponding to the largest urban areas in a country and that corresponding to the rest of smaller urban areas. Firstly, it is found that the top largest urban areas in E&W and France, London and Paris respectively, display high road accident rates with respect to each country's average. Both E&W and France are countries that, despite their different levels of centralisation [Hooghe et al., 2010], have remained relatively unified in recent history. This has allowed their capital cities to forge their pivotal role, not only at a national level, but also as global cities [Kearney, 2020]. Here, it is found that both urban areas corresponding to the capital cities are also special when it comes to their number of road accidents per person. In contrast, Germany and Spain have either only been unified recently or have experienced surges of internal divide [Kaasa et al., 2014]. As a consequence, they display several urban areas that compete somewhat for the leadership with respect to socioeconomic power. In Germany, the 'Big Five' metropolitan regions (Berlin, Hamburg, the

Rhine-Ruhr metropolitan region, Frankfurt and Munich) are all prominent in terms of investment and market development. In Spain, there are two main urban areas with a central socioeconomic role: Madrid and Barcelona. Analogously, the incidence of road accidents is more spread across all of these urban areas instead of concentrating in just one as it was the case for E&W and France.

Secondly, the number of road accidents in smaller urban areas displays high variability in all the countries. This effect could be attributed to the different volumes of traffic found in different locations, since traffic flow [Wang et al., 2013] and, arguably, traffic congestion [Cabrera-Arnau et al., 2020], have been shown to be positively correlated with accident rate. For example, in E&W, the results recognise Hull and Lincoln as accident hotspots, which have also been ranked among the top ten cities with the highest levels of traffic congestion in the UK [INRIX, 2020]. A remarkable concentration of urban areas with an above-average incidence of road accidents per person is also detected on the South East of England, including London's satellite urban areas and other coastal urban areas. Like Hull, some of these locations are near to large ports. Ports are freight-generating points and hence, attract heavy goods vehicles from other places. Depending on how accommodating the surrounding infrastructure is, ports can limit the urban space, while at the same time, increasing traffic flow [Browne et al., 2017] and leading to more road accidents. Other urban areas that have a sea or river port are La Rochelle, Loirent and Marseille in France; Bremen, Hamburg and Regensburg (Danube port) in Germany and Aviles, Cadiz, Gijon and Seville (Guadalquivir port) in Spain. The number of road accidents in all of these locations is above the country's average.

It has been discussed before that ports have the potential to alter the urban hierarchy and, through history, have been key to shape the urban ag-

glomeration mechanism and the network organisation of cities [Lugo et al., 2020]. The authors in [Lugo et al., 2020], even suggest that while urban population size is still used as the most relevant differentiation factor in the study of urban systems, this measure should be associated with port sites in order to improve our understanding of urban hierarchies. The results of this Chapter hint at the possibility that indeed, ports may have an effect in the hierarchical organisation of an urban system such that urban areas with small to medium population size become very influential in the transport network and flow of urban resources. Appropriate traffic management in port cities could then help diminish the number of road accidents in those locations.

More generally, the results in this Chapter show how road accidents are spread across the different urban areas of four countries and therefore, can help determine the top-priority regions to be targeted by policies for the alleviation of disruption caused by road accidents. Particularly, the findings of this Chapter should be considered when countries apply any levelling-up strategies to improve aspects of certain regions that are yet to reach the overall national standard.

It remains as future work to improve our understanding of the causes that lead to the unusually elevated number of accidents in certain urban areas. This can be done by studying, for example, the traffic flow levels in these urban areas or their particular demographic composition, since it has been shown that certain demographic groups have an increased risk of being involved in road accidents [Petridou and Moustaki, 2000]. Other directions for future research include an analysis based on choices of urban areas other than the Eurostat functional urban areas or restricted to only a type of accident, e.g. fatal accidents.

At this point, it is worth mentioning some of the major limitations encountered when performing this research. While Eurostat attempts to

unify the urban areas for different countries, the definitions are still county-dependent, since they are based on the particular geographical hierarchies established by each country. Hence, the comparability between the results corresponding to the countries analysed here is compromised. Furthermore, the definitions provided by Eurostat are obviously limited to European countries. In this context, it is necessary to remark the need to standardise road safety data definitions and collection procedures so that more low-income countries, which tend to be the most affected by road accidents, can be included in the body of research, in line with the central, tranformative promise of the 2030 Agenda for Sustainable Development and its Sustainable Development Goals (SDGs) [United Nations Sustainable Development Group, 2021] to 'Leave No One behind'.

Chapter 7

Conclusions and discussion

In this thesis, the patterns created by media coverage and road accidents in urban areas have been analysed. Through the use of urban scaling models as the main mathematical framework in this analysis, new information has been obtained regarding the effect of urban population size on each of the phenomena of interest. The approach followed throughout the thesis corresponds to that of complex systems, where the collective behaviour of the people living in urban areas is considered, but the individual actions of each person are not investigated.

The study of urban scaling models in this thesis has risen questions related to the methodology in their application as well as some aspects of their validity. These questions arise especially in the presence of urban areas which display dragon-king features. Due to their extremely large populations and their central socioeconomic role, urban areas with dragon-king features can behave as a statistical outlier. When this is the case, the estimation of the parameters in the urban scaling models can be affected.

This Chapter summarises, firstly, the results and insights obtained from analysing data and applying mathematical models. The results give valuable information that can be used towards the accomplishment of the underlying goal of this work, which is to tackle serious social challenges in

light of urbanisation. However, solving social challenges usually comes with some practical difficulties, which cannot always be captured in the data analysis or the mathematical models. In this Chapter, these limitations are also discussed. Finally, there is a section dedicated to future work ideas and another one where final remarks are given.

7.1 Summary

7.1.1 Media coverage

Media coverage is determined by several factors which include advertisers, governments, other public institutions, interest groups and the community. The phenomenon of media coverage can be considered as a social challenge, due to its power to change social attitudes. For example, the way in which an event is portrayed by the media can influence the state of entire economies. Media coverage can also shape our attitudes to certain demographic groups, such as immigrants, and even determine our physical and mental health.

In this thesis, the spatial distribution of media coverage has been investigated by analysing the relationship between the media coverage and urban population size. The analysis considers different urban areas affected the Puebla earthquake, which hit Mexico on the 19th of September, 2017. Urban scaling is used as the framework to model this relationship. The data used for the analysis corresponds to the weeks following the occurrence of the earthquake, although data from a week before the earthquake has also been analysed in order to establish the baseline levels of media coverage. The results show that the baseline scaling behaviour is superlinear, meaning that larger urban areas receive more media coverage per person. In the week just after the earthquake, however, the scaling is still

superlinear, but the tendency is not as strong as it was before the earthquake, so the media coverage of earthquake-related content is more evenly spread across the urban areas. It takes about four weeks for the values of the scaling parameters to return to baseline values.

In addition to the spatial distribution, the work also looks at the temporal evolution of the proportion of coverage given by the media to the earthquake, showing that it takes about four weeks for this proportion to vanish.

The data used in the analysis regarding the media contents is sourced from a selection of Mexican newspapers, all of them with national coverage and written in Spanish. The actual contents are obtained from the newspapers Twitter accounts. The work is therefore based on the assumptions that the choice of newspapers are representative of the overall media activity and that the contents they decide to post on Twitter are representative of their individual activity.

7.1.2 Road accidents in built-up areas

Road accidents are still one of the leading causes of death among young age groups as well as one of the leading causes of disability. Like media coverage, road accidents are the result of many interacting factors which include legislators, manufacturers, drivers, infrastructure or environment among many others. In this thesis, the effect of urban population size on the number of road accident in urban areas has been investigated. The approach relies again on the application of urban scaling models. The findings obtained as a result of this analysis are of great importance considering that most road accidents take place in developing countries and these are precisely the countries where the urbanisation and motorisation processes are most accelerated.

In Chapter 4, the analysis was applied to built-up areas from England and Wales, France and Spain, which are urban areas defined according to a land-use criterion. In this case, the road accident data was from the years 2008-2018 for the case of England and Wales and France; and from 2008-2015 for Spain, since 2015 was the most recent year for which data was available at the time of the analysis. The urban scaling models were applied separately for accidents of different degrees of severity, using a Poisson regression for the estimation of the parameters in the models.

The data and the models reveal that road accidents scale superlinearly, so the more populous urban areas are also the ones with higher rates of road accidents per person. This trend could be attributed to the fact that larger urban areas tend to have higher levels of traffic congestion, hence increasing the chances of an accident. However, more traffic congestion also means lower travelling speeds on the road, which would make the accidents less severe. This observation, which was discussed in detail both in Chapter 4 and Chapter 6, raises a conflict between two desirable objectives: reducing the number of road accidents and reducing traffic congestion.

The analysis in Chapter 4 is also concerned with the temporal patterns followed by road accident data. It is found that fatal accidents are more frequent in rural areas than in urban areas. Their frequency is higher during the daytime hours than the nighttime. However, Friday and Saturday nights have a higher frequency than the other nights of the week. This rise in the number of fatal accidents is likely due to the fact that alcohol consumption is higher at these times.

In the case of serious and minor accidents, they follow similar patterns, with clear frequency peaks on weekdays corresponding to the rush hours and more spread frequency on weekends. The number of peaks is two or three depending on the number of rush hours in each country. In England and Wales and France, the evening rush hour has a higher frequency of

accidents, possibly attributable to the poorer visibility conditions in the dark and the increased stress levels of drivers after a whole day of work.

7.1.3 Dragon kings and urban scaling models

A more detailed analysis of the distribution of population sizes of urban areas serves as motivation for Chapter 5. It is observed that urban population sizes tend to follow heavy-tailed distributions, where the values corresponding to the upper tail of the distribution are more extreme than what would be expected if the data was exponentially distributed. The largest urban areas in the sample often have a population that goes even beyond what the underlying heavy-tailed distribution would predict. The unexpectedly large population size of these urban areas, combined with the fact that they usually have a central socioeconomic role, make them special with respect to the rest of cities in the country. For these reasons, these extraordinary urban areas are sometimes referred to as 'dragon kings'.

While dragon kings can be characterised statistically, this requires knowing the underlying heavy-tailed distribution of urban population sizes followed by the rest of cities. In Chapter 5, it is argued how inferring the underlying distribution is not easy and sometimes not even possible. Therefore, synthetic data for urban population size, represented by the random variable X, is generated in order to analyse how dragon-king urban areas may affect the estimation of parameters in urban scaling models.

The data samples are generated so that they follow a heavy-tailed distribution, but also contain one dragon king. The corresponding values of the Y variable for each randomly generated value of X can follow one of several probability distributions. The probability distribution of Y is conditional on X, according to a probabilistic interpretation of the a scaling model, given by $E[Y|X] = \alpha X^{\beta}$, where $E[Y|X]$ is the expected value of Y

given X and α and β are the scaling parameters, to be estimated according to an appropriate generalised linear model.

In Chapter 5, two distributions for Y are explored: the Poisson and the negative binomial distributions. It is concluded that the presence of dragon kings in the X variable does not affect the value of the estimated scaling exponent β, as long as Y satisfies the scaling model $E[Y|X] = \alpha X^{\beta}$ for all the cities under consideration. However, dragon kings tend to display behaviours that are different from other urban areas, so in real scenarios, Y might not follow the same scaling model as the rest of cities. When this is the case, the estimated values of the scaling parameters given by the Poisson regression will be further from the true values than those given by the negative binomial regression. This observation is due to the fact that the data points corresponding to the largest population sizes have more weight in the estimation of the scaling parameters in the Poisson regression, whereas the weights are more evenly spread across population sizes in the negative binomial regression.

This finding reveals a conundrum in the application of urban scaling models. If estimation methods that place relatively less weight on larger urban areas are used, such as the negative binomial regression, then the estimates of the scaling parameters will be closer to the true values. However, the largest urban areas are usually the most interesting ones, as they not only host a large proportion of a country's population, but are also central in the socioeconomic dynamics of the country. In the urbanising world, tens of megacities are likely to emerge in the near future and some of these will develop dragon-king features. Including them in the urban scaling models will compromise the precision of the estimated values of the parameters if methods such as the Poisson regression are used. If the negative binomial regression is used instead, then larger urban areas will loose some of their influence in determining the scaling parameters.

7.1.4 Road accidents in functional urban areas

The analysis in Chapter 6 consists, again, in the application of urban scaling models to understand the relationship between urban population size and road accident rates in urban areas. In Chapter 6, however, accidents of different degrees of severity are not analysed separately. While Chapter 6 can be regarded as a stand-alone chapter, the analysis is conceived as a way to explore some of the shortcomings of urban scaling models. The findings in Chapter 5, together with a deeper knowledge of the current literature, prompted the new approach followed in Chapter 6.

Firstly, the definition of urban area used in Chapter 6 is based on commuting flows and corresponds to the functional urban areas established by Eurostat. These functional urban areas have boundaries that encompass a much wider surface than the built-up areas considered in Chapter 4 and the land is not necessarily built up. The new definition has a homogenising effect among urban areas, so that the aggregated data for functional urban areas with different population sizes, on a per capita basis, will not be as different as if the urban areas were defined according to the more restrictive land-use criterion. Using functional urban areas instead of built-up areas is a way of testing how urban scaling models respond to different definitions in the data.

Secondly, instead of using a Poisson regression as in Chapter 4, a negative binomial regression is used in Chapter 6. This regression method is chosen in view of the results from Chapter 5, which show that the negative binomial regression is less sensitive to departures from the underlying scaling model, especially in the presence of outstandingly large urban areas. Furthermore, a more statistical perspective is taken on the estimated parameters and this time, tests to assess if the scaling exponent is significantly different from 1 are applied.

Thirdly, data from Germany is incorporated to the analysis in addition to the data from England and Wales, France and Spain. The inclusion of Germany facilitates, on the one hand, a more comprehensive study of road accidents across urban areas for different countries with different distributions of population sizes. Furthermore, countries like England and Wales or France have only one urban area with an extraordinarily large population size as well as a pivotal socioeconomic role, while other countries, like Germany or Spain, have more than one urban areas with these characteristics. Hence, Chapter 6 explores how the structure of the urban system from different countries might manifest in the distribution of accident rates across urban areas.

The findings of Chapter 6 are different from those in Chapter 4, but are not in contradiction. Through the application of urban scaling models, it is found that urban population size has no significant effect on the number of road accidents in urban areas from the four European countries. In Chapter 4, the scaling behaviour was found to be generally superlinear, although tests for the statistical significance of those results were not carried out. The difference between the results are most likely caused by the different definitions of urban areas and the different regression methods used in each Chapter.

The findings of Chapter 6 focus on two types of behaviour displayed by urban areas, either corresponding to the largest urban areas in a country with dragon-king features or to the rest of smaller urban areas. The largest urban areas in E&W and France, London and Paris respectively, which are also centres for socioeconomic power, display high road accident rates with respect to each country's average. In Germany and Spain, several urban areas compete somewhat for the leadership and analogously, the incidence of road accidents is more spread across all of these urban areas instead of concentrating in just one. In Germany, these are the metropolitan regions

corresponding to Berlin, Hamburg, the Rhine-Ruhr metropolitan region, Frankfurt and Munich, and in Spain, these are Madrid and Barcelona. The number of road accidents in smaller urban areas displays high variability in all the countries. This effect could be due to variable volumes of traffic found in different locations which in combination with traffic congestion, is believed to be positively correlated with the number of road accidents. For example, in E&W, Hull, Lincoln and some urban areas in the South East of England are found to have high road accident rates. Some of these locations are near to large ports, which due to their appeal to heavy goods vehicles from other places, could be the reason of the increased levels of traffic volume in the surrounding urban areas. Similarly, urban areas such as La Rochelle, Loirent and Marseille in France; Bremen, Hamburg and Regensburg (Danube port) in Germany or Aviles, Cadiz, Gijon and Seville (Guadalquivir port) in Spain are found to have relatively high numbers of road accidents, and all of them have a sea or river port.

The results in Chapter 6 show, once again, the way in which road accidents are spread across the different urban areas of four countries. These results, in combination with those in Chapter 4, can serve as a starting point to determine the regions that need to be targeted by policies for the reduction of road accidents.

7.2 Challenges and limitations

7.2.1 Interdisciplinary research

One of the biggest challenges of the research presented in this thesis is its interdisciplinary nature. The work is focused on the application of mathematical models to improve our understanding of some patterns found in real-world data related to social phenomena. The mathematical models

can be difficult from a technical perspective, although this challenges are somewhat expected and it is part of the job of a mathematician or more generally, a modeller to resolve them. However, dealing with real-world data about social phenomena is what makes this work uniquely challenging.

Firstly, processing and understanding the data requires knowledge of concepts and methods that go beyond the realm of mathematical modelling and rely on techniques from statistics, data analysis, geography and computer programming.

Secondly, interpreting the results requires contextual knowledge of the phenomena under study. For example, this work has investigated road accident rates in two types of urban areas and in order to interpret the results of the analysis, it is crucial to know how these urban areas are defined. This requires the researcher to access and read documentation provided by institutions from different countries about their methodology to establish the urban areas. Similarly, knowledge about different road accident definitions and the factors influencing road accident rates are also necessary to give an interpretation of the results.

Last but not least, the results of this thesis provide insightful information about social behaviour in urban areas and their relation to urban population size. However, translating these results from the research setting to the policy-making setting is not straightforward. To do this, even more interdisciplinary work would be needed involving economists, lawyers, demographers, urban planners, human rights experts and social scientists. This complicated task is not covered in this thesis, however, in one of the following sections entitled 'Practical implementation of results', some potential ideas about how the results of the thesis could be of practical use are discussed.

7.2.2 A note on the interpretation of results from urban scaling models

The findings obtained in the Chapters of this thesis are of particular relevance in the context of urbanisation, which involves a population shift from rural to urban areas. This population shift can take place through repurposing land for urban use, but it can also take place through the growth in population of existing urban areas. Urban scaling models are sometimes assumed to have predictive ambitions, and hence, to be able to answer questions about the effects of population increase in a single urban area as a result of urbanisation. If this assumption is accepted, statements could be made such as: 'given that the number of road accidents scales superlinearly, if London's population increases in the next year, then the number of road accidents there will also increase'.

However, when the urban scaling models are applied to cross-sectional data, as is the case in this thesis, this assumption is not necessarily correct. In order to make predictions about the trajectory in time of an urban area, a longitudinal analysis would be required. Since urban scaling analysis is performed cross-sectionally here, the results have a comparative interpretation as opposed to a dynamic interpretation. If the results from longitudinal and cross-sectional analysis were shown to be the same, it would imply that the urban characteristic under study is determined fully by the population size of the urban area. But generally, this is not the case, since urban areas are complex systems subject to the effects of urban networks as well as other factors such as agglomeration and clustering.

The scaling behaviours studied in this thesis should then be interpreted as systemic patterns that emerge as a result of a multitude of factors, but these factors can be related to the population size of an urban area. Then, for example, a superlinear scaling behaviour of road accident rates could

be attributed to factors such as traffic congestion or levels of stress as well as population size, but ultimately, these factors are more prevalent in larger urban areas. It would still of great value to understand what are the different components that make the scaling behaviour emerge and this can be done, for example, following the methodology of [Keuschnigg et al., 2019].

7.2.3 Practical implementation of results

The results obtained in this thesis give information about patterns followed by media coverage and road traffic accidents in several countries. These results require thorough interpretations in order to gain insights into the phenomena of interest. This thesis has attempted to provide these interpretations and thus, helps to elucidate the causes of the observed behaviours. The results have the potential to highlight locations and times where the incidence of a type of event, such as road accidents, is dangerously high. Hence, the interpretation of results can be followed by one further step, which involves, if needed, the creation of an action plan to change the current state of the urban system. For example, if the analysis detects a certain urban area as an accident hotspot, measures should be proposed in order to alleviate the incidence of road accidents in that location.

With regards to the media coverage, a superlinear behaviour is found based on data from urban areas affected by the Puebla earthquake from 2017. The superlinear tendency just after the earthquake is not as strong as it is on a baseline week or a four weeks after the earthquake. The media coverage of a natural disaster such as this earthquake can be crucial to determine aspects such as the amount of financial and humanitarian support received by different urban areas to recover from the catastrophe. It can also be sufficient to push investors away from companies based in the locations that have been most affected by the earthquake, hence adding to

the negative impact of the earthquake on the economy of the region. In this situation, some type of intervention to regulate the media contents could help protect the public interest. However, deciding whether this is ethical and if so, how it should be done are problems that escape the scope of this work.

In the case of road traffic accidents, Chapters 4 and 6 reach slightly different results, reflecting the differences in the methodology. Focusing on the temporal patterns found in Chapter 4, it is observed that Friday and Saturday nights see an unusually high number of road traffic accidents, possibly due to increased traffic and drink-driving. In order to establish an action plan, more information would be needed about the cause of this rise in the number of accidents. Supposing the reason is definitely drink-driving, more measures could be put in place to reduce the number of people driving under the influence, which could include campaigns to increase awareness, more traffic police control on the roads or more severe penalties for practicing this behaviour. However, whether these measures are enough of a priority and if so, when, where and how they should be applied are issues that involve a complicated decision-making process, where experts whose focus is the practical implementation of new regulations should participate.

Thus, the production of concrete action plans to address the points raised here is one of the major limitations of this work. Closing the gap between mathematical modelling and policy making, would require further collaboration with experts from different fields.

7.2.4 Lack of public data

Acquiring adequate public data sets was another of the challenges encountered in this thesis. There are countries, such as the United Kingdom, where the data sets can be easily accessed, are well curated and are updated regularly. However, other countries have hard-to-navigate websites

that can make the process of finding the data very time-consuming. Even though most online resources provide contact details for technical assistance, the support teams can be hard to reach. Additionally, the data sets are not always recent and their use might undermine the relevance of the research outputs.

In order to understand the data provided by institutions from the different countries analysed in this thesis, it is necessary to read specific documentation for each data set. Despite the efforts of many countries to make their data open, in some cases, the documentation is only available in their native language. For this reason, the number of countries studied here is limited.

Making public data open is not enough to facilitate scientific research or any other activities that might benefit from the data sets. Data should be FAIR as well as open, that is, it should meet the principles of findability, accessibility, interoperability, and reusability.

7.3 Future work

Some ideas for future work have already been discussed in the conclusions of Chapters 3, 4, 5 and 6. In this section, those that are of special interest are reiterated. Some additional ideas are also proposed in this section, but this time from a perspective that integrates the insights gained from the results of each individual Chapter.

7.3.1 Media coverage

The media coverage is analysed here both before and after the occurrence of the Puebla earthquake. A reason why this event was chosen is that it affected a wide geographical region encompassing several urban areas of different population sizes, hence facilitating the application of urban scaling

models. However, the earthquake did not impact all the urban areas equally and so, the strength of the impact could have been a factor determining the amount of coverage given by the media to each urban area. The strength of the impact was not among the variables taken into account in this thesis and therefore, more work in this direction would be useful.

The work on media coverage introduces the media coverage index as a general metric that could be used for other applications. Hence, it would also be interesting to find other events or topics that affect a whole region, preferably in a homogeneous way, but that also have a well-defined time of occurrence, since this is crucial to analyse the temporal evolution of the media coverage. Then, would the scaling behaviour associated with other events also be superlinear? Is there a correlation between the length of time that an event stays in the spotlight an the scaling behaviour? Can we identify a set of characteristics of each event that determine the scaling behaviour of the media coverage?

Finally, Twitter is used here as the platform of choice to extract data about the media contents. But, Twitter contents are not comprehensive, since there are many media channels that may not have a presence in this particular platform. Hence, are there any other easily accessible ways of obtaining data other than through Twitter?

7.3.2 Road accidents

The patterns followed by road accidents from England and Wales, France, Germany and Spain have been explored in this thesis. However, all of these countries have similar road accident death rates, which are well below the global average. It remains as future work to apply the methods and models introduced here to investigate road accidents in countries with higher road accident death rates, especially low and middle-income countries, which tend to be under-represented in the literature, but then, the lack of open

and FAIR data might be an issue.

Additionally, in this thesis accidents have only been classified according to their location and their degree of severity. However, there are many other classifications that would yield valuable results. For example, considering that more than a third of fatal road accidents in low and middle-income countries are among pedestrians and cyclists, future research could focus on applying the analysis by road user.

Many open questions remain regarding the true contributing factors to the patterns detected in this work. Scientific understanding of these factors can help close the gap between academic research and policy making, and so, of all the suggestions for future work mentioned in this section, this one is perhaps the most lucrative one. The analysis proposed in [Keuschnigg et al., 2019], could serve as a guidance to achieve these results.

7.4 Final remarks

This thesis has demonstrated that both media coverage and the number of road accidents follow clear spatial and temporal patterns. Understanding the patterns is key to unveil the underlying mechanisms of these phenomena and therefore, to design efficient plans that target socioeconomic disruption.

In particular, urban scaling models unlock one of the doors leading to a more sustainable urbanisation process. Knowledge of the relative social behaviours in urban areas of different population sizes can shape a country's political agenda by informing the process of resource allocation to different regions.

However, this thesis has also shown that urban scaling models are not infallible and the estimation of model parameters is a difficult task, which sometimes requires making assumptions about the data. Additionally, urban areas with dragon-king features may display behaviours which disagree

with those followed by other smaller urban areas. Hence, it is not always clear whether they should all be analysed simultaneously. If for some reason, it is deemed appropriate to analyse all the urban areas together, statistical models that account for the high degree of variability in the data, such as the negative binomial regression, must be chosen. When it comes to the estimation of the urban scaling model parameters, these methods usually give less weight to the larger urban areas and consequently, the estimated values are more reliant on the smaller urban areas. Hence, this thesis has raised the issue of the inability of urban scaling models to faithfully describe urban areas across all population sizes. It is important to remember, however, that mathematical modelling usually aims to provide simplified descriptions of a system that can help explain its behaviour. So, no matter how complicated a model is, reality is always more complex, and there will often be some aspects that the model is unable to reproduce.

Despite the possible issues raised when using urban scaling models, they are still a tool that firstly, motivates the collection of large amounts of data from urban areas and secondly, can help solve some serious social problems by uncovering non-obvious patterns of their behaviour.

References

Adams, J. G. U. (1987). Smeed's law: some further thoughts. *Traffic Engineering and Control*, **28**(2), 70–73.

Alirezaei, M., Onat, N. C., Tatari, O. and Abdel-Aty, M. (2017). The climate change-road safety-economy nexus: A system dynamics approach to understanding complex interdependencies. *Systems*, **5**(1).

Alonso, W. (1971). The economics of urban size. *Papers in Regional Science*, **26**(1), 67–83.

Alstott, J., Bullmore, E. and Plenz, D. (2014). powerlaw: A python package for analysis of heavy-tailed distributions. *PLOS ONE*, **9**(1), 1–11.

Amato, G., Bolettieri, P., Monteiro de Lira, V., Muntean, C. I., Perego, R. and Renso, C. (2017). Social media image recognition for food trend analysis. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1333–1336.

Anderson, T. K. (2009). Kernel density estimation and k-means clustering to profile road accident hotspots. *Accident Analysis & Prevention*, **41**(3), 359 – 364.

Andreassen, D. (1991). Population and registered vehicle data vs. road deaths. *Accident Analysis & Prevention*, **23**(5), 343 – 351.

Andreassen, D. C. (1985). Linking deaths with vehicles and population. *Traffic Engineering and Control*, **26**(11), 547–549.

Angel, S., Sheppard, C. S., Civco, D. L., Buckley, P., Chabaeva, A., Gitlin, L., Kraley, A. and Parent, J. (2005). *The Dynamics of Global Urban Expansion*. Transport and Urban Development Department, The World Bank.

Arcaute, E., Hatna, E., Ferguson, P., Youn, H., Johansson, A. and Batty, M. (2015). Constructing cities, deconstructing scaling laws. *Journal of The Royal Society Interface*, **12**(102), 20140745.

Auerbach, F. (1931). Das Gesetz der Bevölkerungskonzentration. *Petermanns Geographische Mitteilungen*, pp. 74–76.

Ausserhofer, J. and Maireder, A. (2013). National politics on Twitter: structures and topics of a networked public sphere. *Information, Communication & Society*, **16**(3), 291–314.

Bak, P., Tang, C. and Wiesenfeld, K. (1988). Self-organized criticality. *Physical Review A*, **38**(1), 364–374.

Baker, S. P., Whitfield, R. A. and O'Neill, B. (1987). Geographic variations in mortality from motor vehicle crashes. *The New England Journal of Medicine*, **316**(22), 1384–7.

Bar-Yam, Y. (1997). *Dynamics of Complex Systems*. Perseus Books. USA.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, **286**(5439), 509–512.

Barrat, A., Barthélemy, M. and Vespignani, A. (2004). Weighted evolving networks: Coupling topology and weight dynamics. *Physical Review Letters*, **92**, 228701.

Barthélemy, M. (2019). Modeling cities. *Comptes Rendus Physique*, **20**(4), 293–307.

Barthélemy, M., Bordin, P., Berestycki, H. and Gribaudi, M. (2013). Self-organization versus top-down planning in the evolution of a city. *Scientific Reports*, **3**(1).

Batty, M. (2009). *Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies*. Springer New York. New York, NY. pp. 1041–1071.

Batty, M. and Ferguson, P. (2011). Defining city size. *Environment and Planning B: Planning and Design*, **38**(5), 753–756.

Baudains, P., Fry, H. M., Davies, T. P., Wilson, A. G. and Bishop, S. R. (2016). A dynamic spatial model of conflict escalation. *European Journal of Applied Mathematics*, **27**(3), 530—553.

BBC (2010). How the UK's first fatal car accident unfolded. *Accessed June 2020*.

Beck, L. F., Downs, J., Stevens, M. R. and Sauber-Schatz, E. K. (2017). Rural and Urban Differences in Passenger-Vehicle–Occupant Deaths and Seat Belt Use Among Adults — United States, 2014. *MMWR Surveillance Summaries*, **66**(17), 1–13.

Berelson, B. (1952). Content analysis in communication research. *The ANNALS of the American Academy of Political and Social Science*, **283**(1), 197–198.

Bettencourt, L. M. A., Lobo, J., Helbing, D., Kühnert, C. and West, G. B. (2007). Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences*, **104**(17), 7301–7306.

Bettencourt, L. M. A., Lobo, J., Strumsky, D. and West, G. B. (2010). Urban scaling and its deviations: Revealing the structure of wealth, innovation and crime across cities. *PLOS ONE*, **5**(11), 1–9.

Bettencourt, L. M. A., Yang, V. C., Lobo, J., Kempes, C. P., Rybski, D. and Hamilton, M. J. (2020). The interpretation of urban scaling analysis in time. *Journal of The Royal Society Interface*, **17**(163), 20190846.

Bokányi, E., Kondor, D. and Vattay, G. (2019). Scaling in words on twitter. *Royal Society Open Science*, **6**(10).

Bornstein, M. H. and Bornstein, H. G. (2007). The pace of life. *Nature*, **259**.

Bouchaud, J.-P. and Mézard, M. (2000). Wealth condensation in a simple model of economy. *Physica A: Statistical Mechanics and its Applications*, **282**(3), 536–545.

Brelsford, C., Lobo, J., Hand, J. and Bettencourt, L. M. A. (2017). Heterogeneity and scale of sustainable development in cities. *Proceedings of the National Academy of Sciences*, **114**(34), 8963–8968.

Browne, M., Woexenius, J., Dablanc, L., Cherrett, T. and Morganti, E. (2017). Port cities and urban logistics. *The 22nd Annual Conference of The Chartered Institute of Logistics and Transport, Logistics Research Network*.

Brueckner, J. K. (1987). *Chapter 20. The structure of urban equilibria: A unified treatment of the muth-mills model. Urban Economics. Vol. 2 of Handbook of Regional and Urban Economics*. Elsevier.

Byrne, J. P., Mann, N. C., Dai, M., Mason, S. A., Karanicolas, P., Rizoli, S. and Nathens, A. B. (2019). Association Between Emergency Medical Service Response Time and Motor Vehicle Crash Mortality in the United States. *JAMA Surgery*, **154**(4), 286–293.

Cabrera-Arnau, C. and Bishop, S. R. (2020). The effect of dragon-kings on the estimation of scaling law parameters. *Scientific Reports*, **10**(1).

Cabrera-Arnau, C., Prieto Curiel, R. and Bishop, S.R. (2020). Uncovering the behaviour of road accidents in urban areas. *Royal Society Open Science*, **7**(4), 191739.

Cabrera-Arnau, Carmen and Bishop, Steven R. (2021). Urban population size and road traffic collisions in europe. *PLOS ONE*, **16**(8), 1–13.

Camagni, R., Capello, R. and Caragliu, A. (2013). One or infinite optimal city sizes? In search of an equilibrium size for cities. *The Annals of Regional Science*, **51**(2), 309–341.

Candia, C., Jara-Figueroa, C., Rodriguez-Sickert, C., Barabási, A.-L. and Hidalgo, C. A. (2018). The universal decay of collective memory and attention. *Nature Human Behaviour*, **3**, 82–91.

Carey, H. C. (1858). *Principles of Social Science*.

Chermak, S. M. and Gruenewald, J. (2006). The media's coverage of domestic terrorism. *Justice Quarterly*, **23**(4), 428–461.

Cilliers, P. (2002). *Complexity and Postmodernism: Understanding Complex Systems*. Taylor & Francis.

Clauset, A., Shalizi, C. R. and Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, **51**(4), 661–703.

Colizza, V., Barrat, A., Barthélemy, M. and Vespignani, A. (2006). The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences*, **103**(7), 2015–2020.

Coman, A. (2018). Predicting the decay of collective memory. *Nature Human Behaviour*, **3**, 18–19.

Combes, P. P., Duranton, G., Gobillon, L., Puga, D. and Roux, S. (2012). The productivity advantages of large cities: Distinguishing agglomeration from firm selection. *Econometrica*, **80**(6), 2543–2594.

Cottineau, C. (2017). Metazipf. a dynamic meta-analysis of city size distributions. *PLOS ONE*, **12**(8), 1–22.

Cottineau, C., Hatna, E., Arcaute, E. and Batty, M. (2017). Diverse cities or the systematic paradox of urban scaling laws. *Computers, Environment and Urban Systems*, **63**, 80–94.

Cvetojevic, S. and Hochmair, H. H. (2018). Analyzing the spread of tweets in response to Paris attacks. *Computers, Environment and Urban Systems*, **71**, 14 – 26.

De Vries, J. (1990). *Problems in the measurement, description and analysis of historical urbanization. In 'Urbanization in History: A Process of Dynamic Interactions'; edited by A. M. van der Woude, A. Hayami and J. De Vries*. Clarendon Press.

Deffuant, G., Neau, D., Amblard, F. and Weisbuch, G. (2000). Mixing beliefs among interacting agents. *Advances in Complex Systems*, **03**(01n04), 87–98.

Department for Transport (2017*a*). Reported Road Casualties in Great Britain: notes, definitions, symbols and conventions. *Accessed July 2020*.

Department for Transport (2017*b*). Reported road casualties in Great Britain, provisional estimates involving illegal alcohol levels. *Accessed July 2019*.

Department for Transport (2018). Road Safety Data. *Accessed March 2021*.

Department for Transport (2019). Accident and casualty costs (RAS60). *Accessed March 2021*.

Depersin, J. and Barthelemy, M. (2018). From global scaling to the dynamics of individual cities. *Proceedings of the National Academy of Sciences*, **115**(10), 2317–2322.

Dirección General de Tráfico (2015). Fichero de Microdatos. *Accessed March 2021*.

Ditton, J., Chadee, D., Farrall, S., Gilchrist, E. and Bannister, J. (2004). From imitation to intimidation a note on the curious and changing relationship between the media, crime and fear of crime. *British Journal of Criminology*, **44**(4), 595–610.

Ditton, J. and Duffy, J. (1983). Bias in the newspaper reporting of crime news. *British Journal of Criminology*, **23**, 159.

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A. and Danforth, C. M. (2011). Temporal patterns of happiness and information in a global social network: hedonometrics and Twitter. *PLOS ONE*, **6**(12), e26752.

Downs, A. (1972). Up and down with ecology: The "issue-attention cycle". *The Public*, pp. 38–50.

Duranton, G. and Puga, D. (2003). Micro-foundations of urban agglomeration economies. Working Paper 9931. *National Bureau of Economic Research*.

Editorial (2009). No man is an island. *Nature Physics*, **5**(1).

Eeckhout, J. (2004). Gibrat's law for (all) cities. *American Economic Review*, **94**(5), 1429–1451.

Eeckhout, J. (2009). Gibrat's law for (all) cities: Reply. *American Economic Review*, **99**(4), 1676–83.

Erdogan, S., Yilmaz, I., Baybura, T. and Gullu, M. (2008). Geographical information systems aided traffic accident analysis system case study: city of Afyonkarahisar. *Accident Analysis & Prevention*, **40**(1), 174 – 181.

Eurostat (2021*a*). The Geographic Information System of the Comission. *Accessed March 2021.*

Eurostat (2021*b*). Methodological manual on city statistics. *Accessed March 2021.*

Fischer, M. M. (2006). Chapter 2. spatial analysis and gis. *in* Spatial Analysis and GeoComputation: Selected Essays. Spatial Interaction Models and the Role of Geographic Information Systems. Springer Berlin Heidelberg. pp. 29–42.

Fragkias, M., Lobo, J., Strumsky, D. and Seto, K. C. (2013). Does size matter? Scaling of CO2 emissions and U.S. urban areas. *PLOS ONE*, **8**(6), 1–8.

Frette, V., Christensen, K., Malthe-Sørenssen, A., Feder, J., Jøssang, T. and Meakin, P. (1996). Avalanche dynamics in a pile of rice. *Nature*, **379**, 49–52.

Gabaix, X. (1999). Zipf's law for cities: An explanation. *The Quarterly Journal of Economics*, **114**(3), 739–767.

Galtung, J. and Ruge, M. H. (1965). The structure of foreign news: The presentation of the Congo, Cuba and Cyprus crises in four Norwegian newspapers. *Journal of Peace Research*, **2**(1), 64–90.

Gibrat, R. (1931). *Les inégalités économiques*. Librairie du Recueil Sirey.

Glaeser, E. L. and Maré, D. C. (2001). Cities and skills. *Journal of Labor Economics*, **19**(2), 316–342.

Glaeser, E. L. and Sacerdote, B. (1999). Why is there more crime in cities?. *Journal of Political Economy*, **107**(S6), S225–S258.

Global Health Data Exchange (2019). GBD Results Tool. *Accessed March 2021*.

Gómez-Liévano, A., Vysotsky, V. and Lobo, J. (2020). Artificial increasing returns to scale and the problem of sampling from lognormals. *Environment and Planning B: Urban Analytics and City Science*, **48**(6), 1574–1590.

Gong, P., Liang, S., Carlton, E. J., Jiang, Q., Wu, J., Wang, L. and Remais, J. V. (2012). Urbanisation and health in China. *The Lancet*, **379**(9818), 843 – 852.

Gonzalez, R., Cummings, G., Mulekar, M. and Rodning, C. B. (2006). Increased Mortality in Rural Vehicular Trauma: Identifying Contributing Factors Through Data Linkage. *The Journal of Trauma: Injury, Infection, and Critical Care*, **61**(2), 404–409.

Haldane, J. B. S. (1985). *On being the right size and other essays /J.B.S. Haldane; edited by John Maynard Smith*. Oxford Paperbacks.

Hall, C. M. (2002). Travel safety, terrorism and the media: The significance of the issue-attention cycle. *Current Issues in Tourism*, **5**(5), 458–466.

Harcup, T. and O'Neill, D. (2001). What is news? Galtung and Ruge revisited. *Journalism Studies*, **2**(2), 261–280.

Hardin, J. W. and Hilbe, J. (2007). *Generalized linear models and extensions*. Stata Press.

Haynes, K. E. and Fotheringham, S. (1984). *Gravity and Spatial Interaction Models*. Vol. 2 of *SAGE series in Scientific Geography*. Sage. United States.

Hennessy, D. A. and Wiesenthal, D. L. (1997). The relationship between traffic congestion, driver stress and direct versus indirect coping behaviours. *Ergonomics*, **40**(3), 348–361.

Hennessy, D. A., Wiesenthal, D. L. and Kohn, P. M. (2000). The influence of traffic congestion, daily hassles, and trait stress susceptibility on state driver stress: An interactive perspective. *Journal of Applied Biobehavioral Research*, **5**(2), 162–179.

Herbert, A. S. (1962). The architecture of complexity. *Proceedings of the American Philosophical Society*, **106**(6), 467–482.

Higgins, C. D., Sweet, M. N. and Kanaroglou, P. S. (2018). All minutes are not equal: travel time and the effects of congestion on commute satisfaction in Canadian cities. *Transportation*, **45**(5), 1249–1268.

Himelboim, I., McCreery, S. and Smith, M. (2013). Birds of a feather tweet together: integrating network and content analyses to examine cross-ideology exposure on Twitter. *Journal of Computer-Mediated Communication*, **18**(2), 40–60.

Hooghe, L., Marks, G. and Schakel, A. H. (2010). *The Rise of Regional Authority*. 1st edn. Taylor & Francis.

INRIX (2020). Global Traffic Scorecard 2020. *Accessed March 2021.*

Institut National de la Statistique et des Études Économiques (2018*a*). Base des unités urbaines. *Accessed March 2021.*

Institut National de la Statistique et des Études Économiques (2018*b*). Chapitre 1. Les unités urbaines. *Accessed March 2021.*

Institut National de la Statistique et des Études Économiques (2018*c*). Populations légales. *Accessed March 2021.*

Instituto Nacional de Estadística (2018). Cifras oficiales de población de los municipios españoles. *Accessed March 2021.*

Instituto Nacional de Estadística y Geografía (2015). Encuesta Intercensal (México). *Accessed May 2021.*

Janczura, J. and Weron, R. (2012). Black swans or dragon-kings? A simple test for deviations from the power law. *The European Physical Journal Special Topics*, **205**.

Jefferson, M. (1939). The law of the primate city. *Geographical Review*, **29**(2), 226–232.

Joppke, C. (1991). Social movements during cycles of issue attention: The decline of the anti-nuclear energy movements in West Germany and the USA. *British Journal of Sociology*, **42**(1), 43–60.

Kaasa, A., Vadi, M. and Varblane, U. (2014). Regional cultural differences within european countries: Evidence from multi-country surveys. *Management International Review*, **54**(6).

Kapitaniak, T. and Bishop, S. R. (1999). *A Dictionary of Nonlinear Dynamics.*

Kearney (2020). Global Cities Report. *Accessed March 2021.*

Keuschnigg, M., Mutgan, S. and Hedström, P. (2019). Urban scaling and the regional divide. *Science Advances*, **5**(1), eaav0042.

Kleiber, M. (1932). Body size and metabolism. *Hilgardia*, **6**(11), 315–353.

Kounadi, O., Lampoltshammer, T. J., Groff, E., Sitko, I. and Leitner, M. (2015). Exploring Twitter to analyze the public's reaction patterns to recently reported homicides in London. *PLOS ONE*, **10**(3), e0121848.

Krugman, P.R. (1996). *The Self Organizing Economy.* Blackwell Publishers.

Kumar, S. and Toshniwal, D. (2016). A data mining approach to characterize road accident locations. *Journal of Modern Transportation*, **24**(1).

Kuntsche, E. and Cooper, M. L. (2010). Drinking to have fun and to get drunk: Motives as predictors of weekend drinking over and above usual drinking habits. *Drug and Alcohol Dependence*, **110**(3), 259 – 262.

Lac, A., Handren, L. and Crano, W. D. (2016). Conceptualizing and measuring weekend versus weekday alcohol use: Item response theory and confirmatory factor analysis. *Prevention Science*, **17**(7), 872–881.

Ladyman, J., Lambert, J. and Wiesner, K. (2013). What is a complex system?. *European Journal for Philosophy of Science*, **3**, 33–67.

Lagarde, E., Chastang, J. F., Gueguen, A., Coeuret-Pellicer, M., Chiron, M. and Lafont, S. (2004). Emotional stress and traffic accidents: The impact of separation and divorce. *Epidemiology*, **15**(6), 762–766.

Laherrère, J. and Sornette, D. (1998). Stretched exponential distributions in nature and economy: "fat tails" with characteristic scales. *The European Physical Journal B - Condensed Matter and Complex Systems*, **2**, 525–539.

Le Maître, A. (1682). *La métropolitée, ou De l'établissement des villes capitales, de leur utilité passive et active ....* Amsterdam, B. Boekholt.

Leetaru, K. and Schrodt, P. A. (2013), Gdelt: Global data on events, location, and tone, 1979–2012, *in* ISA annual convention. *Vol. 2*. Citeseer. pp. 1–49.

Leitão, J. C., Miotto, J. M., Gerlach, M. and Altmann, E. G. (2016). Is this scaling nonlinear?. *Royal Society Open Science*, **3**(7), 150649.

Leon, D. A. (2008). Cities, urbanization and health. *International Journal of Epidemiology*, **37**(1), 4–8.

Levy, M. (2009). Gibrat's law for (all) cities: Comment. *American Economic Review*, **99**(4), 1672–75.

Liska, A. E. and Baccaglini, W. (1990). Feeling safe by comparison: crime in the newspaper. *Social Problems*, **37**(3), 360–374.

Louf, R. and Barthelemy, M. (2014). How congestion shapes cities: from mobility patterns to scaling. *Scientific Reports*, **4**(5561).

Lugo, I., Alatriste-Contreras, M. G. and Pumain, D. (2020). The role of ports in the dynamics of urban hierarchies. *Maritime Policy & Management*, **0**(0), 1–18.

MacKay, R. S. (2008). Nonlinearity in complexity science. *Nonlinearity*, **21**(12), T273–T281.

Malevergne, Y., Pisarenko, V. and Sornette, D. (2011). Testing the pareto against the lognormal distributions with the uniformly most powerful unbiased test applied to the distribution of cities. *Physical Review E*, **83**(3), 036111.

Malleson, N. and Andresen, M. A. (2015). The impact of using social media data in crime rate calculations: shifting hot spots and changing spatial patterns. *Cartography and Geographic Information Science*, **42**(2), 112–121.

Marshall, A. (2013). *Principles of economics*. Palgrave classics in economics. Palgrave Macmillan.

May, R. M. (1973). Qualitative stability in model ecosystems. *Ecology*, **54**(3), 638–641.

McDonald, S. (2009). Changing climate, changing minds: Applying the literature on media effects, public opinion, and the issue-attention cycle to increase public understanding of climate change. *International Journal of Sustainability Communication*, **4**, 45–63.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. and Alon, U. (2002). Network motifs: Simple building blocks of complex networks. *Science*, **298**(5594), 824–827.

Ministerio de Fomento (2018*a*). Áreas Urbanas en España. *Accessed July 2020*.

Ministerio de Fomento (2018*b*). Atlas Digital de las Áreas Urbanas. *Accessed March 2021*.

Molinero, C. and Thurner, S. (2019). How the geometry of cities explains urban scaling laws and determines their exponents. *arXiv:1908.07470*.

Montgomery, M. R. (2008). The urban transformation of the developing world. *Science*, **319**(5864), 761–764.

Morley, D. (2003). *Television, audiences and cultural studies*. Routledge. London, UK.

Newman, M. (2010). *Networks: An introduction*. Oxford University Press.

Observatoire National Interministériel de la Sécurité Routière (2018). Base de données accidents corporels de la circulation. *Accessed March 2021*.

Odlyzko, Andrew (2015). The forgotten discovery of gravity models and the inefficiency of early railway networks. *Œconomia*, **5**(1), 157–192.

OECD (2020). Road accidents. *Accessed March 2021*.

Office for National Statistics (2018*a*). *Lower Layer Super Output Area boundaries*.

Office for National Statistics (2018*b*). *Lower Layer Super Output Area population estimates*.

Office for National Statistics, Open Geography Portal (2011). Output Area (2011) to Built-up Area Sub-division to Built-up Area to Local Authority District to Region (December 2011) Lookup in England and Wales. *Accessed March 2021*.

Olteanu, A., Castillo, C., Diakopoulos, N. and Aberer, K. (2015). Comparing events coverage in online news and social media: the case of climate change. *International Conference on Web and Social Media*, **15**, 288–297.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. *Proceedings of the Seventh International Conference on Language Resources and Evaluation*.

Parliamentary Advisory Council for Transport Safety (2019). *Road Safety Since 2010 – Update with 2017 data*.

Peters, B. G. and Hogwood, B. W. (1985). In search of the issue-attention cycle. *The Journal of Politics*, **47**(1), 238–253.

Petersen, K. K. (2009). Revisiting Downs' issue-attention cycle: International terrorism and US public opinion. *Journal of Strategic Security*, **2**(4), 1–16.

Petridou, E. and Moustaki, M. (2000). Human factors in the causation of road traffic crashes. *European Journal of Epidemiology*, **616**(9).

Pisarenko, V. F. and Sornette, D. (2012). Robust statistical tests of Dragon-Kings beyond power law distributions. *The European Physical Journal Special Topics*, **205**(1).

Population Division of the UN Department of Economic and Social Affairs (2018). *UN World Urbanization Prospects: The 2018 Revision.*

PressGazette (2017). *NRS national press readership data: Telegraph overtakes Guardian as most-read 'quality' title in print/online.*

Prieto Curiel, R. and Bishop, S. (2017). Modelling the fear of crime. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **473**(2203), 20170156.

Prieto Curiel, R., Cabrera-Arnau, C., Torres Pinedo, M., González Ramírez, H. and Bishop, S. R. (2019). Temporal and spatial analysis of the media spotlight. *Computers, Environment and Urban Systems*, **75**, 254 – 263.

Prieto Curiel, R., González Ramírez, H. and Bishop, S. R. (2018). A novel rare event approach to measure the randomness and concentration of road accidents. *PLOS ONE*, **13**(8), 1–18.

Prieto Curiel, R., Pappalardo, L., Gabrielli, L. and Bishop, S. R. (2018). Gravity and scaling laws of city to city migration. *PLOS ONE*, **13**(7), e0199892.

Puga, D. (2010). The magnitude and causes of agglomeration economies. *Journal of Regional Science*, **50**(1), 203–219.

Pumain, D. (2000). Settlement systems in the evolution. *Geografiska Annaler. Series B, Human Geography*, **82**(2), 73–87.

Pumain, D. (2004). Scaling laws and urban systems. Working Paper 2004-02-002. *Santa Fe Institute*.

Ravenstein, E. G. (1885). The laws of migration. *Journal of the Statistical Society of London*, **48**(2), 167–235.

Reiner, T. A. and Parr, J. B. (1980). A note on the dimensions of a national settlement pattern. *Urban Studies*, **17**(2), 223–230.

Retallack, A. E. and Ostendorf, B. (2019). Current understanding of the effects of congestion on traffic accidents. *International Journal of Environmental Research and Public Health*, **16**(18).

Ribeiro, P. and Queiros-Condé, D. (2017). A scale-entropy diffusion equation to explore scale-dependent fractality. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **473**(2200), 20170054.

Richardson, H. W. (1972). Optimality in city size, systems of cities and urban policy: a sceptic's view. *Urban Studies*, **9**(1), 29–48.

Rishi, P. and Khuntia, G. (2012). Urban Environmental Stress and Behavioral Adaptation in Bhopal City of India. *Urban Studies Research*, **2012**. Article ID: 635061.

Schläpfer, M., Bettencourt, L. M. A., Grauwin, S., Raschke, M., Claxton, R., Smoreda, Z., West, G. B. and Ratti, C. (2014). The scaling of human interactions with city size. *Journal of The Royal Society Interface*, **11**(98), 20130789.

Schmidt, A., Ivanova, A. and Schäfer, M. S. (2013). Media attention for climate change around the world: A comparative analysis of newspaper coverage in 27 countries. *Global Environmental Change*, **23**(5), 1233–1248.

Scott, A. J. (1997). The cultural economy of cities. *International Journal of Urban and Regional Research*, **21**(2), 323–339.

Secretaría de Desarrollo Agrario Territorial y Urbano and Consejo Nacional de Población and Instituto Nacional de Estadística y Geografía (2015). Delimitación de las zonas metropolitanas de México. *Accessed April 2020*.

Seto, K. C., Güneralp, B. and Hutyra, L. R. (2012). Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools. *Proceedings of the National Academy of Sciences*, **109**(40), 16083–16088.

Seto, K. C., Reenberg, A., Boone, C. G., Fragkias, M., Haase, D., Langanke, T., Marcotullio, P., Munroe, D. K., Olah, B. and Simon, D. (2012). Urban land teleconnections and sustainability. *Proceedings of the National Academy of Sciences*, **109**(20), 7687–7692.

Shalizi, C. R. (2011). Scaling and Hierarchy in Urban Economies. *arXiv:1102.4101*.

Shefer, D. (1994). Congestion, air pollution, and road fatalities in urban areas. *Accident Analysis & Prevention*, **26**(4), 501–509.

Shefer, D. and Rietveld, P. (1997). Congestion and safety on highways: Towards an analytical model. *Urban Studies*, **34**(4), 679–692.

Simon, F. and Corbett, C. (1996). Road traffic offending, stress, age, and accident history among male and female drivers. *Ergonomics*, **39**(5), 757–780.

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika*, **42**(3-4), 425–440.

Simon, H. A. (1971). Designing organizations for an information-rich world. *Computers, Communications and the Public Interest: Conference Proceedings. The Johns Hopkins Press*, pp. 37–72.

Smeed, R. J. (1949). Some statistical aspects of road safety research. *Journal of the Royal Statistical Society. Series A (General)*, **112**(1), 1–34.

Smith, F. T. and Curtis, J. P. (2008). The dynamics of persuasion. *International Journal of Mathematical Models and Methods in Applied Sciences*, **1**(2), 115—122.

Sornette, D. (2009). Dragon-kings, black swans and the prediction of crises. *International Journal of Terraspace Science and Engineering*, **2**(1), 1–18.

Sornette, D. and Ouillon, G. (2012). Dragon-kings: Mechanisms, statistical methods and empirical evidence. *The European Physical Journal Special Topics*, **205**(1).

Srivastava, K. (2009). Positive mental health and its relationship with resilience. *Industrial Psychiatry Journal*, **18**(2), 75–76.

Statistisches Bundesamt (2018*a*). *46241-01-04-5: Straßenverkehrsunfälle, verunglückte Personen - Jahressumme - regionale Tiefe: Gemeinden*.

Statistisches Bundesamt (2018*b*). *Regional statistics*.

Steindl, J. (1965). *Random Processes and the Growth of Firms. A Study of the Pareto Law*.

Strano, E., Giometto, A., Shai, S., Bertuzzo, E., Mucha, P. J. and Rinaldo, A. (2017). The scaling structure of the global road network. *Royal Society Open Science*, **4**(10), 170590.

Strogatz, S. H. (2000). *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering*. Westview Press.

Sun, J., Li, T., Li, F. and Chen, F. (2016). Analysis of safety factors for urban expressways considering the effect of congestion in Shanghai, China. *Accident Analysis & Prevention*, **95**, 503–511. Traffic Safety in China: Challenges and Countermeasures.

Sutton, J. (1997). Gibrat's legacy. *Journal of Economic Literature*, **35**(1), 40–59.

Taylor, A. H. and Dorn, L. (2006). Stress, fatigue, health, and risk of road traffic accidents among professional drivers: The contribution of physical inactivity. *Annual Review of Public Health*, **27**(1), 371–391.

Tefft, B. C. (2017). Rates of Motor Vehicle Crashes, Injuries and Deaths in Relation to Driver Age, United States, 2014-2015. *AAA Foundation for Traffic Safety*.

Theofilatos, A. (2017). Incorporating real-time traffic and weather data to explore road accident likelihood and severity in urban arterials. *Journal of Safety Research*, **61**, 9 – 21.

Thompson, D. A. W. (2014). *On Growth and Form; edited by Bonner, J. T.* Canto Classics. Cambridge University Press.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, **46**, 234–240.

Trivedi, J., Sareen, H. and Dhyani, M. (2008). Rapid urbanization - Its impact on mental health: A South Asian perspective. *Indian Journal of Psychiatry*, **50**(3), 161–165.

Twitter (2017). Twitter Developer Platform. *Accessed November 2018.*

United Nations, Department of Economic and Social Affairs, Population Division (2018). World Urbanization Prospects: The 2018 Revision, Online Edition. *Accessed August 2020.*

United Nations Sustainable Development Group (2021). Leave No One Behind. *Accessed March 2021.*

Ver Hoef, J. M. and Boveng, P. L. (2007). Quasi-poisson vs. negative binomial regression: how should we model overdispersed count data?. *Ecology*, **88**(11), 2766–2772.

Vision Zero (2020). Vision Zero. *Accessed March 2021.*

Vosoughi, S., Roy, D. and Aral, S. (2018). The spread of true and false news online. *Science*, **359**(6380), 1146–1151.

Wang, C., Quddus, M. A. and Ison, S. G (2009). Impact of traffic congestion on road accidents: A spatial analysis of the M25 motorway in England. *Accident Analysis & Prevention*, **41**(4), 798 – 808.

Wang, C., Quddus, M. A. and Ison, S. G. (2013). The effect of traffic and road characteristics on road safety: A review and future research direction. *Safety Science*, **57**, 264–275.

Weisbuch, G., Deffuant, G., Amblard, F. and Nadal, J.-P. (2002). Meet, discuss, and segregate!. *Complexity*, **7**(3), 55–63.

Wener, R.E. and Evans, G.W. (2011). Comparing stress of car and train commuters. *Transportation Research Part F: Traffic Psychology and Behaviour*, **14**(2), 111 – 116.

Weng, L., Flammini, A., Vespignani, A. and Menczer, F. (2012). Competition among memes in a world with limited attention. *Scientific Reports*, **2**, 335.

Whitelegg, J. (1987). A geography of road traffic accidents. *Transactions of the Institute of British Geographers*, **12**(2), 161–176.

Wilson, A. G. (1967). A statistical theory of spatial distribution models. *Transportation Research*, **1**(3), 253–269.

Winters, J. V. (2011). Why are smart cities growing? Who moves and who stays. *Journal of Regional Science*, **51**(2), 253–270.

World Bank Group (2019). *The World Bank*.

World Health Organization (2018). *Global status report on road safety 2018*.

Wu, F. and Huberman, B. A. (2007). Novelty and collective attention. *Proceedings of the National Academy of Sciences*, **104**(45), 17599–17601.

Xu, C., Tarko, A. P., Wang, W, and Liu, P. (2013). Predicting crash likelihood and severity on freeways with real-time loop detector data. *Accident Analysis & Prevention*, **57**, 30 – 39.

Xu, W. W., Sang, Y., Blasiola, S. and Park, H. W. (2014). Predicting opinion leaders in Twitter activism networks: the case of the Wisconsin recall election. *American Behavioral Scientist*, **58**(10), 1278–1293.

Yang, J. and Leskovec, J. (2011), Patterns of temporal variation in online media, in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. ACM. pp. 177–186.

Youn, H., Bettencourt, L. M. A, Lobo, J., Strumsky, D., Samaniego, H. and West, G. B. (2016). Scaling and universality in urban economic diversification. *Journal of The Royal Society Interface*, **13**(114), 20150937.

Yuan, Y. and Liu, Y. and Wei, G. (2017). Exploring inter-country connection in mass media: A case study of China. *Computers, Environment and Urban Systems*, **62**, 86–96.

Zelenkauskaite, A. and Balduccini, M. (2017). "Information warfare" and online news commenting: Analyzing forces of social influence through location-based commenting user typology. *Social Media + Society*, **3**(3), 2056305117718468.

Zhong, C., Arisona, S. M., Huang, X., Batty, M. and Schmitt, G. (2014). Detecting the dynamics of urban structure through spatial network analysis. *International Journal of Geographical Information Science*, **28**(11), 2178–2199.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley Press.

Zwerling, C., Peek-Asa, C., Whitten, P. S., Choi, S-W., Sprince, N. L. and Jones, M. P. (2005). Fatal motor vehicle crashes in rural and urban areas: decomposing rates into contributing factors. *Injury Prevention*, **11**(1), 24–28.