Mathematical modelling of social systems

Rafael Prieto Curiel

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London

UCL Department of Mathematics

University College London

September 20, 2018

I, Rafael Prieto Curiel, 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

Mobility and migration patterns, the concentration of crime and opinion dynamics observed on the fear of crime are all examples of social systems in which complex patterns emerge that subsequently feed back into the overall system. This thesis describes new methods established to analyse such patterns which appear in social systems.

The main application area is in the field of crime science, but the methods developed here have wider applications within other social systems, some of which are also explored in the thesis, such as migration or road accidents. Based on new assessments of data, by utilising novel techniques of analysis and visualisation, models are also developed to determine how the perception of security is affected by particular crimes.

The new metrics and models developed here consider different types of situation. Firstly, for events which have low frequency and yet a high degree of concentration; secondly, the distribution of such events which allows them to be simulated under different conditions; and then finally, understanding the impact of different degrees of concentration.

An individual's fear of crime is the result of a mixture of factors which go beyond merely the actual crime experienced by that person, such as fear shared by others, memory of past events and of previous perceptions, crime reported in the media and more. This thesis quantifies fear of crime in a way that may prove useful to identify factors which increase fear of crime besides crime itself, explain why fear of crime emerges in a population and suggests policies for controlling fear.

Impact Statement

The research presented in this thesis develops new mathematical techniques to consider various aspects of quantitative social science. The applications are in human migration, the concentration of road accidents, the distribution of crime and the fear of crime. In recent years, the use of mathematical modelling for decision making has become almost the norm, whether this is in the use of general weather predictions, the path of hurricanes or the spread of infectious diseases.

In the area of crime, the use of models for predicting crime hot spots is now widely used by policing bodies. This body of work will contribute to this domain, adding new dimensions that will allow policymakers to make better decisions. In the case of crime and road accidents, the thesis considers a new methodology for dealing with sparse data. It gives a simple distribution, which is a close approximation to reality, from which rare events can be simulated. This methodology allows us to analyse past events taking into account their potential random aspects and so it has applications for policy design and the methodology could be used to determine upper limits of police intervention.

Fear of crime is a difficult and sensitive topic to analyse but quantifying fear reveals significant aspects both for academic and policy design purposes. This thesis analyses fear of crime from a regional perspective and shows that regions with less crime than others might yet have a higher degree of fear. Much police and political attention is given over to managing this fear. The thesis gives a mathematical framework to analyse fear of crime and shows the emergence of a generalised fear of crime even under

low crime (or even no crime) circumstances. The thesis also shows that fear might drastically increase after a specific event which may then take some time to reduce back to 'normal' levels. All of this has the potential to provide impact in the policy-making domain. However, having a better understanding of fear of crime has applications beyond crime studies as it helps to explain why extreme opinions emerge as a result of social interactions. Hence this work will have an impact more generally in the social science arena.

Collecting tweets and analysing their contents concerning crime and its fear, indicates differences between a quantified phenomenon, such as crime, its fear and how comments are shared by millions of users of social media. This thesis goes from actual crime to the fear of crime and then expressions of that fear on social media and it gives quantitative tools to analyse the apparent relationships between them. Hence the work may have wider impacts on our society.

Acknowledgements

The work presented here is the result of multiple collaborations. Firstly, I acknowledge the support and encouragement of Prof Steven Bishop, for his valuable contributions, guidance and never-ending patience. Also, to Prof Shane Johnson for sharing valuable experience and his insights.

Collaboration and working on the interface between mathematics and other sciences is a fundamental aspect of this thesis. I acknowledge working with Humberto González Ramírez on road accidents; with Sofía Collignon Delmar on metrics of crime concentration; with Luca Pappalardo and Lorenzo Gabrielli on migration; with Cristina Muntean and Stefano Cresci on expressions of fear on social media; with and with Philipp Heinrigs and Inhoi Heo on cities and spatial interactions and with Mara Torres Pinedo and Carmen Cabrera Arnau on media attention. With all the people I collaborated with, it was a pleasurable experience and a valuable learning process.

I am deeply grateful for the support of the Mexican Government through a Conacyt Scholar and also the encouragement of the UCL Mathematics Department and the MAPS Faculty.

My time at UCL and in London was a lifetime experience. To my friends from the Statistics Department, the KLB, Chalkdust, the UCL Mexican Society, Punto Decimal and the SuperKLB, thank you for being part of this period and for helping me construct my home away from home.

A very special thank you for the support I received from my friends and family from Mexico. Love and encouragement is felt even at large distances.

And finally, to my grandmothers, my brother and my mother. I would not be able to understand life without the four pillars of my life.

Contents

Chapter 1

Introduction

Mathematical modelling is the procedure whereby we consider the attributes and conditions of a system and then produce an abstraction and generalisation of them so that we can use the new model either to understand the interactions of the variables better or predict future outcomes. Whether we are considering a physical setting, in which variables might be observed or measured, or we are considering a social framework, in which variables might be much more difficult to quantify, typically, the mathematical modelling process concentrates on a specific phenomenon, reduces it to its more basic causes and mechanisms, formulates a hypothesis (which means transforming the system into equations), and then tests them against reality (which requires us to collect data or observations of the phenomenon).

The key element to start this modelling process is to decide what is relevant, so what variables need to be measured and what can be neglected as irrelevant or what cannot be simply measured. A decision is made; a higher precision might be achieved by considering more variables or a more complex system but at the cost of having to deal with a larger number of parameters, a system of overly complicated equations or an extensive amount of measurements and so, a simplification of the reality might be much more desirable in the context of mathematical modelling especially as this helps our understanding. Simplification is an important issue. For example, when

drinking a cup of water, to know the temperature of the liquid, we would rarely be interested in describing the temperature of each molecule but probably instead, be content that the mean temperature, expressed as a single-valued number, as a good enough approximation. In this case, more sophisticated models, which consider convection, evaporation, the distribution of different temperatures contained in the cup or hysteresis, could provide a more detailed description of the temperature of the water contained in the cup, but then most likely the complexity of the model would far exceed the complexity of its use.

The difficulties faced when modelling any phenomenon in a physical framework are exceeded when any social setting is considered. Models which deal with human interaction tend to simplify the microscopic, individual level in the hope to resemble the macroscopic, social behaviour. Thus, a voter's opinion might be modelled as a binary variable, a crime might be regarded as a point located on a map or a friendship could be considered as a link in a network; however, these simplifications made within a social context have helped us to understand the emergent patterns exhibited by voters (Düring et al., 2009), the levels of concentration of crime and the formation of criminal hotspots (Short et al., 2010) and the small-world phenomena observed in many social networks (Watts and Strogatz, 1998). Individual aspects of large and complicated systems are not necessarily the most relevant (Galam et al., 1982). Usually, details at a microscopic level are overlooked in favour of the macro features exhibited by the system as a whole. The mathematical approach is usually to study the emergent collective patterns when thousands or millions of people are considered with individual behaviours rarely being studied.

1.1 Aims of the thesis

Mathematical modelling allows our current knowledge of a specific setting to be challenged. In the case of social systems, mathematical modelling

facilitates our understanding of why collective behaviours emerge and allows us to construct 'what-if' scenarios and compare two different settings, exporting the techniques and results from one setting, or discipline even, to the other.

Mathematical modelling of social systems is the core objective of this thesis, and therefore, it has an interdisciplinary approach although it is based on mathematical modelling. Three specific aims form the core of this thesis, which then open broader discussions on the use of models for quantitative social sciences. Firstly, to analyse in the context of cities, the migration patterns which emerge after observing a specific population (such as a country or a continent) for a period of time. Secondly, to analyse, also in the context of cities, what are the main reasons which cause fear of crime in its population. Finally, to use the same modelling techniques applied in the case of migration and fear of crime, to other social settings, such as road accidents.

1.1.1 Migration patterns

Migration is one of the main reason why cities within the same region grow at different rates and, according to the World Bank, migration is identified as the main driver of city changes (Lee et al., 2015). For instance, during the past 30 years, Mexico City achieved the lowest fertility rate of the country but, nevertheless, its population has maintained the same national growth rate. The fact that in Mexico City people have fewer children than anywhere else in the country but it is still growing at the same speed (meaning that during the past 30 years Mexico City has doubled its population) can only be explained by migration.

This observation of migration, enhancing the growth in the population of Mexico City, is perhaps also the reason as to why Paris, London or Lagos have become the primary cities within their countries. Migration is one of the main drivers which has shaped cities and countries and has defined many national borders. Therefore, understanding what causes a person to

migrate, what patterns emerge when thousands of migrants are considered, why one city is more attractive than others to migrate to or why people tend to migrate more from one city rather than others, is a crucial aspect which has impacts on urban studies, sustainable growth and predicting the future trends of our cities.

Modelling and understanding migration in the US, considering cities as the observation units and migration between cities, between the countryside and the cities and form other countries to the cities is covered in Chapter 2.

1.1.2 Crime and fear of crime

Quantifying the fear of crime in a city, and understanding the patterns of fear is one of the aims of this thesis. Fear of crime, and not necessarily crime itself, is becoming more and more a serious issue around the World, particularly in big cities. It is the primary cause why people sometimes take expensive safety precautions, why some people might suffer from anxiety, social segregation and the reason why millions of people have moved from their cities looking for a safer place to live.

Understanding and measuring the dynamics of the fear of crime firstly requires a model of crime itself, or at least a way to represent it. The distribution of crime in a population allows different patterns in which crime is suffered to be simulated and highlights the relevance of using adequate tools to deal with events such as crime. One of the complicating issues about the analysis of crime is that it is highly concentrated and has a low frequency. Thus, the construction of the distribution of crime in a population and a metric for the concentration of rare events is the first step towards the analysis of the fear of crime and therefore, Chapter 3 is devoted to the analysis of rare events and Chapter 4 is dedicated to the study of rare events in the context of the distribution of crime in a population.

Determining the actual relevance that suffering a specific crime has on a population, the impact of crime being more or less concentrated and the dynamics under which fear of crime is shared across a population, is then

covered in Chapter 5.

Finally, in terms of fear of crime, different expressions collected at a city level from social media and from victimisation surveys allows the mathematical models and hypothesis of fear of crime to be compared with the observed reality. Chapter 6 is then the modelling of fear of crime at a city level put into practice.

1.1.3 Other social settings

Constructing different metrics for the migration patterns between cities or either the distribution of crime suffered by its population or the dynamics of its fear, requires the detection of the relevant attributes of the event so that characteristics which are either irrelevant, not quantifiable, not well defined or not a direct consequence of the event can be neglected. For instance, in the case of crime, the number of crimes suffered by the victims is often considered, although identifying the victim might not be clear (who is the victim of money laundering or corruption cases?) and counting the number of events might not be so straightforward (should we count every interaction between a rapist or a bully and its victim as a separate crime?).

One of the most powerful aspects of mathematical modelling is precisely that two distinct events, as different as they may be, are often mapped, using a similar system of metrics and equations, and therefore, can be envisaged using the similar models. For instance, opinion dynamics has been modelled using a similar system to that used to mimic the behaviour of an inhomogeneous gas; a model for human migration has been considered using concepts based on Newton's gravitational laws; the retaliatory behaviour of some gang fights has been well described using techniques from seismology, among many other examples.

The final aim of this thesis is to transfer concepts developed here for migration, crime and its fear, to other social scenarios. Particularly, the analysis of the concentration of crime is applied to understanding the concentration of road accidents. Chapter 7 is devoted to two special cases of social

scenarios, where the techniques described in previous chapters are used for different systems.

1.2 Models in social systems

Mathematical models of social systems have motivated research and advancements in mathematics and probability theory. For instance, Adolphe Quetelet observed conviction rates in France in 1835 and used tools from astronomy to measure the "true level of criminality" (Maltz, 1996). Following this, Simen-Denis Poisson was looking at the number of wrongful convictions during a specific time-interval and used the same data as Quetelet to develop what we now know as the *Poisson distribution*. Social systems have often helped in the development and advancement of mathematics and science by either providing the specific challenges or simply the data to unveil particular patterns.

There is, however, a serious challenge in the case of the mathematical models of social systems. Measuring the heat transfer between regions of a wire, the viscosity of a flow or the position of a star, for instance, is perhaps more natural and has its measuring instruments and units as opposed to measuring human constructs such as love, friendship, power or fear. Repeating a physical experiment allows reproducibility, gives certainty to the results obtained and transforms them into laws, meaning that simple, absolute and universal descriptions are attained. The well known *Hooke's law*, for instance, says that the force required to compress or extend a spring scales linearly with the distance. It can be expressed in a single equation, tested under different conditions and applied in an extensive list of physical settings. However, modelling social systems is fundamentally different.

For instance, the *first law of Geography* (developed by Waldo Rudolph Tobler in 1960s) states that "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). Although it is considered as a *law*, how related things are, how fast 'relatedness' decays

as distance increases, how to compare spatial and temporal distances, or what does it even mean to be related, how is it measured and what are its units are significant questions which show that this law is far from being a simple, absolute and universal description.

A similar law of social systems is the recent description of the *law of crime concentration at a place* (published by David Weisburd in his works between 2011 and 2015) which states that "for a defined measure of crime at a specific microgeographic unit, the concentration of crime will fall within a narrow bandwidth of percentages for a defined cumulative proportion of crime" (Weisburd, 2015). Although highly relevant for describing the spatial concentration of crime as an expected phenomenon when observing crime at places, it also highlights the complexity of any social description. The law of crime concentration assumes a defined measure of crime, which could be the number of crimes on each microgeographic unit, but also, the number of victims, the value of the stolen property, the number of casualties, the impact on the victims, the fear caused by the crimes, the cost of security and others. In addition, a defined cumulative proportion of crime could be a different proportion of crimes (as usually is the case), so "N% of places have M% of crimes" is reported in different studies for a wide range of values of N and M and so comparing the concentration of crime between different places is, in many cases, impossible (Lee et al., 2017). The fact that the concentration can only be described in terms of a "narrow bandwidth" but without any description of how narrow the bandwidth is, shows that uncertainty is expected, that without many observations of the same concentration, perhaps for different time periods, outliers are almost impossible to detect. Although the concentration of crime is discussed later (Chapters 3 and 4) the law of crime concentration shows the complexity of any mathematical description of a social system.

Finally, the *law of migration* (Ravenstein, 1885), first published by Ernst Georg Ravenstein in 1885, was developed by looking at migration at a

county level from and to the UK, Ireland and Scotland and states, among many vital issues, that *the majority of migrants move a short distance; migrants who move longer distances tend to choose big-city destinations; urban residents are often less migratory than inhabitants of rural areas; and large towns grow by migration rather than natural population growth*. The law of migration offers a valuable insight into the patterns formed by migrants, although again, far from being a universal description of human migration, with reproducible results which are tested under different conditions, it has serious issues to consider: how long or short are the journeys of the migrants, how big is the city chosen by long-distance migrants or how less migratory are urban inhabitants as opposed to rural inhabitants. Furthermore, the law of migration was developed looking only at the migration process from a very narrow time-interval and only in a very restricted set of countries, and so, how applicable are the results in other regions of the world, how has the law changed over the past decades, how are migration patterns affected by conflicts or disasters and how can migration fluctuations be explained, are relevant challenges not fully answered by the law of migration.

Every human feeling, every interaction with others and every social construct is so unique that the idea of measuring and modelling human or social behaviour with the objective to obtain a generalisation or an abstraction which applies beyond the observed individuals is the first challenge of social modelling.

Social models are inevitably incomplete and inaccurate, because of scientific limitations and a lack of data (Hunt et al., 2012), because it is often impossible to conduct experiments on the large-scale (Johnson, 2010a) and because conventional scientific approaches cannot be applied to many of the problems faced by our society (Johnson, 2000), but models of human behaviour have gained interest as the need for them grows, their results get more and more applied in policy and decision-making and their implications

are spread throughout more widely.

1.2.1 Mathematical modelling

Modelling of social systems brings forth three challenges. The first has to do with the actual information and data of events in a social setting. Any theory within the social framework, whether it is in the context of crime and the way it is suffered, the fear of crime and the way it is shared by the individuals, school bullying and more, has the risk of being nothing more than a hypothesis if it is not compared with reality using observed data, contrasted against other theories or even simulated with agents (Johnson and Groff, 2014). Therefore, an important, and highly sensitive part of models of social systems is the data availability, as it often limits their validation (Pepys, 2016). Modelling goes hand in hand with the availability and use of the data (San Miguel et al., 2012). To understand the fear of crime, for instance, it would be ideal to trace the perception of security that a person has before and after suffering a crime, and before and after reading about a crime in the newspaper, and to compare the two effects, but most of the data available in this context comes from observational studies, with perhaps yearly updates and other issues regarding the samples, the types of questions and the way the answers are recorded and more. Therefore, using actual data in the social framework becomes sometimes the largest challenge of the study.

Secondly, models of social systems are often challenging from a technical perspective. For instance, a dynamic model used for the spatial distribution of crime uses nonlinear partial differential equations to analyse crime pattern formation (Short et al., 2010); the study of the evolution of the distribution of opinions has been modelled using the Fokker–Planck equation (Düring et al., 2009; Düring and Wolfram, 2015); and a stochastic model has been used to describe human migration (Simini et al., 2012; McGinnis, 1968; Myers et al., 1967) are just a few of the examples in which advanced mathematics is needed. Often, social models have a component which requires mathematical techniques to obtain results, but also, other resources

such as statistics, network science, computation, big data, machine learning and spatial statistics (San Miguel et al., 2012) and beyond the quantitative aspects of the models, social models require an interdisciplinary approach. For some social models, simulation is the natural way of investigating their future states (Johnson, 2000).

Constructing, for example, a metric of happiness expressed in social media required storing and processing 46 billion words on 4.6 billion expressions (Dodds et al., 2011); models of human mobility, require processing large amounts of data, usually with a reference to a point in a map, for which the computational power required to deal with such large amounts of data is a new field of study in its own right (Pappalardo et al., 2015b).

From the computational perspective, a few years ago it was inconceivable to model more than just a few aspects of the individuals, but today we are capable of simulating large human systems (Bonabeau, 2002). Computers have changed the paradigm of modelling human behaviour allowing us now to deal with more complex interactions between its members and its environment (Pan et al., 2007); to understand the emergence of crowd behaviour in different situations; to challenge and, in some cases, to measure, some of the theories which were frequently used for crime pattern formation (Johnson and Groff, 2014), for instance, and in the context of other human behaviour and the design of policies (Johnson, 2000).

Finally, models of social systems are challenging because their outcomes might not be socially acceptable, desirable or popular. For instance, a model for the 2011 London riots showed the relevance of delivering police to scenes of disorder before control is lost (Davies et al., 2013) and rioters often select locations to engage in disorder in which previous demonstrations took place (Baudains et al., 2013); increasing the concentration of crime, letting fewer victims suffer more crime, improves the perception of security of a place (Prieto Curiel and Bishop, 2017); individual choice can lead to racial, gender or income segregation (Schelling, 1971); false infor-

mation might gain acceptance and once adopted by an individual, are highly resistant to correction and lead to the proliferation of fake news (Del Vicario et al., 2016); an individual from a larger city tends to have a higher income but at the same time is more likely to suffer serious crime or get AIDS (Bettencourt et al., 2007). Often, the outcomes of a model of a social system seem to challenge principles such as equity, liberty, dignity or freedom of speech, freedom of protest, and others. Modelling our society and its problems might give light to unpopular solutions.

In models of social behaviour, there might be difficulties in measuring its variables, the models involve complicated interactions requiring detailed techniques to deal them and they might provide challenging insights into the society. However, even if the process of modelling social phenomena has many critical components, multidisciplinary approaches and the use of newly available large amounts of data to work alongside increased computational processing power to deal with them is drastically changing the traditional boundaries of science.

1.2.2 Modelling human mobility and migration

Our understanding of human mobility has drastically changed over the past few decades. Data from social media (Noulas et al., 2012), mobile phone users (Lambiotte et al., 2008), where banknotes move during subsequent transactions (Brockmann et al., 2006), census (Levy, 2010), landlines (Schläpfer et al., 2014) and others (Pappalardo et al., 2015b; Rinzivillo et al., 2014; Pappalardo et al., 2016) have recently provided us with information about where humans move, the paths they follow and the patterns which emerge when looking at millions of people moving. The use of different sources of information, which work as a proxy to understand human mobility, has been successfully applied to understand social relationships in Belgium (Lambiotte et al., 2008), county-to-county migration in Kenya (Wesolowski et al., 2013) and others (Gonzalez et al., 2008; Schläpfer et al., 2014).

By looking at the trajectory of mobile phone users, for instance, a bursty

pattern was detected (Gonzalez et al., 2008) showing that usually two types of movements are observed (Rhee et al., 2011): a set of more regular, daily movements which are small, which represent the day-to-day life, and a set of less frequent and usually long-distance movements (Levy, 2010) which represent travelling, movers and other special long-distance journeys.

It is important to be able to model and hence predict both types of movements. Understanding the relatively short daily movements provides insight into why traffic jams form, for instance, allowing us to plan a more efficient public transport system, the impact of massive events (Giulianotti et al., 2015), helps to explain why some parts of a city have prosperous shopping centres whilst others are lagging behind in the consumers' attention and many more studies of urban dynamics. For example, understanding human movement within metropolitan cities and by using data from the social network Foursquare, allowed a *law for human mobility* to be formulated (Noulas et al., 2012), which says that variations in the human movements are predominantly due to different distributions of places across different urban environments.

Understanding the movement of humans over larger distances, on the other hand, allows us to detect the dynamics of specific types of migration: international migration (Lewer and Van den Berg, 2008; Karemera et al., 2000), migration from rural to urban areas (Todaro, 1969; Harris and Todaro, 1970), individual human mobility (Pappalardo et al., 2015b; Rinzivillo et al., 2014; Pappalardo et al., 2016), disaster-, climate change- or conflict-induced migration (Paul, 2005; Myers et al., 2008; Reuveny, 2007; Laczko et al., 2009; Naude, 2008; Ibáñez and Vélez, 2008). In the US, for example, data (of Commerce, 2015) shows that a person from the US moves on average 11.7 times during their lifetime, and typically moves a large distance (Levy, 2010) and also, between 2010 and 2014, for every person moving from the countryside to any city, there was roughly one person moving from a city to the countryside and five people moving between different cities. The

movement of people between different cities is the main dynamic observed in the metropolitan areas in the US.

Except perhaps for specific communities (such as an online gaming group or an academic research group) long-distance social links are explained partly by migration (Lambiotte et al., 2008) and it has been shown, by using nearly one billion phone calls in Belgium and in a second experiment, using data from social media, that the probability of having a social contact at a distance k decreases with k^2 (Lambiotte et al., 2008; Levy and Goldenberg, 2014).

Migration is a key component of the population dynamics and it is the main driver of city changes and the observed variation in city growth around the World (Lee et al., 2015). Micro decisions collectively determine the macro behaviour (Mansury and Shin, 2015) and therefore, different models of human migration, which consider various aspects of the phenomenon, are also a primary tool for the design of policies (Gnisci, 2008; WFP, 2017; Naude, 2008). The analysis of migration has changed over the past century, from being mostly focused in migration from rural to urban regions, to the analysis of movers from country to country, to a city-to-city and to conflict and disaster-induced migration, understanding why individuals decide whether or not to move, their path and their destination has been, and still is, a crucial question that is asked.

It is important to understand human mobility patterns at every spatial dimension. On a microscopic level, patterns of human mobility lead to road traffic, the concentration of crime in certain areas and is one of the causes of social segregation (Schelling, 1971), whilst at a macroscopic level, human mobility is the main reason why diseases are transmitted from far away regions, it helps to explain why some cities grow faster than others and it gives valuable information in terms of migration and international migration.

1.2.3 Modelling crime

Crime itself is a complex phenomenon since we observe unexpected social behaviours which are difficult to understand, control and, sometimes, even to quantify (D'Orsogna and Perc, 2015; Helbing et al., 2015). For instance, it is natural to assume that by enforcing longer prison sentences, harsher punishments or by increasing fines, less crime would be observed, but this is not usually true (Becker, 1968). It is perhaps expected that allocating more police reduces crime via deterrence, but this also might not be true (Kleck and Barnes, 2014), and in the case of fear of crime, it is frequently assumed that a city with fewer crimes experiences less fear, but again, results are often contradictory. Mathematical models of crime, thus, become powerful tools which help explain why this counter-intuitive behaviour emerges. For instance, a mathematical model for the spatial concentration of crime was used to describe the emergence of criminal hotspot patterns (Short et al., 2010); another model showed that when there are significantly high levels of crime, the probability of being arrested goes down and so criminals create a safer environment for themselves to commit more crimes (Glaeser et al., 1995); while another model showed the importance of delivering police to scenes of riots before control is lost (Davies et al., 2013). A useful review which shows some of the weaknesses frequently encountered with traditional economic and statistical models can be found in the work of Mirta Gordon (Gordon, 2010) and a valuable review of some of the powerful mathematical models in crime science can be found in the work by Maria D'Orsogna and Matjaž Perc (D'Orsogna and Perc, 2015).

In the case of crime, for example, assuming that every criminal has the same motivation to commit a crime might lead to an oversimplification of reality, leading to perhaps useless results which might even cause a potential hazard if they were to be extended to reality and used in a political context. But then, as in the example of the cup of water, how complex should a social model be? How many variables should be introduced and how many

parameters considered?

Assuming that a person experiencing social deprivation has a greater propensity to commit a crime, for instance, might be a reasonable assumption when considering the number of robberies in a certain region, but might be completely unrelated if we are considering other types of crime, such as rapes, gun crimes, gang fights or street violence. Thus, details of the individuals considered and their interactions, even at a microscopic level, might have an impact on the outcomes of the model.

From a policy design perspective, the set of measurements and models which accurately reflect the crime conditions within a region should be considered before any decision is made: improvements in what is not measured can be conjectured but not easily be verified. However, are the traditional measurements of crime precise enough and do they reflect the phenomena that we wish to quantify accurately? It has been identified, for example, that a small portion of the population usually suffers a disproportionate amount of crime (Grove et al., 2012), however, how much of the total crime is committed against repeated victims? Could this observed pattern be the result of criminals targeting random people and therefore simply bad luck for the victims? If a policy prevents a victim from suffering any subsequent crimes, is the expected result a genuine reduction of the crime in that region or is the policy creating crime displacement (Johnson et al., 2014), which creates, as a result, possibly even more victims?

A common practice is to take into account the average number of crimes suffered by the victims, which in general, is greater than one, meaning that a person who has suffered a crime is potentially victimised more than once. It is considered to be a proxy for the level of concentration of crime and a number that reflects whether crime is suffered by more or fewer people. Clearly, summarising an issue as complex and multidimensional as crime and its degree of concentration into a single number is desirable since it is simple, but there might be some issues with its mathematical construction.

There are multiple (and perhaps extremely different) scenarios that provide the same numeric result, and so using the average number of crimes blindly used as a metric might not precisely quantify the concentration of crime and it might not reflect the crime conditions from a region and therefore it should be used with care.

1.2.4 Modelling opinions

Going beyond crime itself, a relevant issue to consider is the fear that it causes. Statistically speaking crime is a rare event: even in highly victimised areas, the chances of suffering a crime are relatively low; however, it is not uncommon to be afraid of becoming the victim of a crime. The perception of being insecure is usually much more frequent than crime itself (Grogger and Weatherford, 1995), it has been associated with negative effects on quality of life (Jackson and Gray, 2010) and health (Ruijsbroek et al., 2015) and it has a strong social and political impact and causes prejudice against certain population groups. Hence the relevance of asking what is the impact of crime on the perception of security? Is fear of crime the result of other factors, such as the age of the person (Kershaw and Tseloni, 2005), their financial position (Tseloni, 2007), media coverage of certain types of crime (Chadee and Ditton, 2005), falsely conceived ideas about crime (Gilchrist et al., 1998) and fear to the unknown? Is the fear of crime even related to crime?

Having a fear of crime might not necessarily be wrong. A certain level of the fear of crime might even cause healthy precautions in the population (Jackson and Gray, 2010), and so understanding its causes, its dynamics and the way it is shared by individuals is relevant from the perspective of crime analysis and from a policy design perspective.

The perception of security or insecurity that an individual has of a region can be considered as an opinion and, as such, it might be quantified and modelled in the context of opinion dynamics, where in this case, extreme opinions represent whether a place is considered secure or insecure. Opin-

ion formation has been studied from many angles and different techniques, from agent-based models, to mean field theory and to kinetic models of opinion formation (Düring et al., 2009), where the individual opinion is usually modelled as a single-valued number contained in some closed interval which represents the extreme opinions, for example, the level of production of an employee in a plant (Galam et al., 1982) or left-right leaning voters (Düring and Wolfram, 2015) and individuals are considered to update their own beliefs due to two main reasons: interaction with others and a process of self-thinking. In the context of perception of security, other factors might have an impact on a microscopic level, such as crime itself and its degree of concentration, and therefore play a relevant part of the global behaviour. The emergence of crowd behaviour, in the context of the perception of security, might be highly sensitive to specific parameters of the model, such as the crime rates, any past experiences with crime, the topology of the social network considered and many others.

1.2.5 Modelling other social systems

Advances in crime science have moved at a fast pace recently. Some of the techniques developed, including key findings, might be applicable to the study of other social phenomena, such as the study of migration, school bullying or road accidents. For example, the fact that crime is usually suffered by repeat victims encourages a policy designed for people who already suffered a crime, since that is a good indication of a high risk of suffering future crimes. It is possible then to export that knowledge to other branches of social studies, such as the study of road accidents and accidents in general, where the occurrence of a road accident might work as an indicator of a high risk of future accidents. Insurers certainly think so as premiums go up once a claim has been made.

Chapter 2

Human migration

With more than half of the world's population living in an urban environment, it is crucial to understand the complex relationships, loops and feedbacks which emerge between a person and the city in which they live. Less than 10% of the world live in a metropolitan area with more than 10 million inhabitants, whilst most of the population live in secondary towns and, according to the World Bank, there are more than one billion people living in irregular settlements and slums around the world. Thus, modern urban dynamics face the disproportionate growth of large metropolises, but at the same time, the emergence and expansion of irregular settlements, the growth of the stagnant secondary towns and more complex urban dynamics.

This chapter introduces a mathematical modelling of human mobility and migration patterns, analysed from the context of urban dynamics. It is based on published research (Prieto Curiel et al., 2018b).

2.1 Modelling migration patterns

Humans have been migrating for millennia. From the first crossing of the Bering Strait to the Spanish conquest, from British and French colonial expansion to the influx of students to London today or the flow of skilled workers to Silicon Valley, migration has always been a central feature of human life.

Human migration is a sensitive topic which is easily politicised. It is often thought about in the context of international or illegal migration, most frequently from developing countries to developed ones and particularly, as something that needs to be stopped. The debate around migration would surely benefit from more data and mathematical modelling, and from fewer sensationalist media reports that often presents a distorted reality.

According to the International Organization for Migration, a *migrant* is considered as an individual who is moving or has moved from one location to another (so not considering movements within the same urban area), regardless of their legal status, the causes and without positive or negative connotation.

Modelling any social behaviour is complicated for many reasons. Firstly, it is impossible to observe all of the people involved or consider all of the reasons why they behave in the way that they do, which means that it is nearly possible to recognise only emerging patterns that arise from collective behaviour. Secondly, there will always be outliers. For example, evidence shows that a person who smokes 20 cigarettes a day is 26 times more likely to develop lung cancer than a person who does not smoke, but clearly, there will always be heavy smokers who remain cancer-free. Observing these 'lucky' smokers does not mean that evidence against smoking should be dismissed, but when social patterns are analysed, the general case is considered, that will not always apply to each and every single person.

2.1.1 Mathematical models of human migration

Today, human migration is one of the most debated concerns of the general public, governments and international agencies, due to the importance of integration policies, socioeconomic development and well-being. On the one hand, migrants contribute to the prosperity of their destination, to which they provide new skills, norms and community activities, as well as easing the pressures of an ageing population (World Bank; International Monetary Fund, 2015) and can enhance conditions in their place of origin by either re-

ducing unemployment, improving conditions by sending remittances and increases the resilience in the case of disasters (Nathan, 2014; Benach et al., 2011). On the other hand, human migration creates the political challenge of designing integration policies to allow newcomers to settle in unfamiliar environments, as well as prompting the need for improvement of social support systems (Taylor et al., 2016; Dustmann et al., 2008). Modelling, predicting and therefore understanding human migration is thus of fundamental importance for the formulation, planning and implementation of balanced policy programmes.

It is therefore not surprising that the study of human migration has attracted the interest of scientists from many disciplines. Some studies investigate the dynamics of specific types of migration, such as international migration (Lewer and Van den Berg, 2008; Karemera et al., 2000), migration from rural to urban areas (Todaro, 1969; Harris and Todaro, 1970), mobility in urban areas (Pappalardo et al., 2015b; Rinzivillo et al., 2014; Pappalardo et al., 2016), or disaster-, climate change- and conflict-induced migration (Paul, 2005; Myers et al., 2008; Reuveny, 2007; Laczko et al., 2009; Naude, 2008; Ibáñez and Vélez, 2008). Other studies focus on different models of migration, such as models based on a Markov process (Henry et al., 1971; Kelley and Weiss, 1969; Constant and Zimmermann, 2012) or the cumulative inertia model, according to which an individual is less likely to migrate if they spend more time in the same place (McGinnis, 1968) which was later validated with actual data (Myers et al., 1967). Two prominent migration models are the gravity model, which considers both the population size of the places of origin and destination and the distance between them (Anderson, 2010; Lewer and Van den Berg, 2008), and the radiation model, which additionally takes into account job opportunities in the vicinity of the place of origin (Simini et al., 2012).

There are many models of migration, each of which aims to capture a different aspect of the phenomenon. It depends on the purpose of the model

(the question being considered) and the data that is available. Although mathematical models cannot give a perfect description of the complex pattern that is observed in reality, they can help understanding the reasons why more people might migrate to or from specific locations, help explaining the impact of the distance between the origin and the place where the migrant moves to or certain policies on migration statistics and can be used to forecast, for example, the number of people who will migrate following a natural disaster.

Migration between cities, as well as from the countryside to the cities, has attracted particular interest in recent years. Internal migration is the main reason for urban growth (World Bank; International Monetary Fund, 2015) and the reason why most of the world's population now live in urban areas. A city's population size strongly affects the well-being of its inhabitants (Bettencourt et al., 2010; Pappalardo et al., 2015a, 2016), as large cities provide more efficient resources for their inhabitants (Bettencourt et al., 2007), who tend to develop more social contacts (Schläpfer et al., 2014), move in a more diversified way (Pappalardo et al., 2016, 2015a) and create more patents and bank deposits (Bettencourt et al., 2007). On the negative side, however, large cities suffer more infectious diseases (Bettencourt et al., 2007) as well as more crime (Glaeser and Sacerdote, 1996; Cullen and Levitt, 1999). Migration is the main driver of city changes (Lee et al., 2015) and the reason why some cities grow faster than others, providing a positive feedback leading to even more changes and further population growth (Prieto Curiel et al., 2017b). An important challenge is to understand, quantify and predict the impact of city size on human migration: are individuals in large cities more likely to migrate than individuals who live in small towns? Are individuals more likely to migrate to a city larger than their current one, or does population size not matter? How to quantify the attractiveness of a city for internal and international migrants based on its population size?

Here, human migration is analysed in the context of cities and its population (Batty, 2007). Migration dynamics are examined considering individuals as the inhabitants of a city or potential movers to them (Batty, 2013) but ignoring other individual aspects, such as age, income, education or gender. Starting from official census data, the migration fluxes from and to US cities are analysed and how these fluxes are influenced by the cities' population size is investigated.

The main finding is that migrants preserve city size, i.e., they prefer to migrate to cities with a similar size to the city of their origin. Moreover, a phase transition is observed, where the exponent in the new model changes from sublinear to superlinear at a specific population size. Building upon these findings, a data-driven scaling model is developed, which describes human migration as a two-step decision process, demonstrating that it can partially explain migration fluxes only on the basis of city size. The impact of distance on a gravity-scaling model of human migration is then considered, showing that it performs better than both the scaling and gravity models of human migration.

2.1.2 Migration data

The data source is a census which stores the number of migrants from one metropolitan area to another in the US (of Commerce, 2015), where a metropolitan area or *city* is considered here as a high population density area with strong economic ties and with a population larger than 50,000 inhabitants (based on the metropolitan statistical areas MSAs defined by the U.S. Department of Commerce). This gives 385 cities for which the internal migration process is quantifiable. Note that the area of Los Angeles was merged from the original data source with other three metropolitan areas (Riverside-San Bernardino, Oxnard-Thousand Oaks-Ventura and Bakersfield). These cities are collectively formed by approximately 268 million inhabitants, so more than 80% of the US population. The population size of individual US cities varies broadly from just above 50,000 inhabitants (e.g.,

Carson City) to nearly 20 million inhabitants (New York City and Los Angeles) while individuals living in towns or rural areas with less than 50,000 inhabitants are considered to be part of the countryside.

Using the available data, the following aspects of migration are analysed:

1. the probability that an individual chooses to migrate;
2. the destination picked by migrants according to the size of their city of origin;
3. the probability that an individual moves to the countryside;
4. the destination picked by international migrants.

2.2 Scaling of migration

It is assumed that there are n cities and in the following, X_{ij} is defined as the number of individuals migrating from city i to city j; X_{i*} is defined as the (total) outflow migration from i and X_{*j} as the (total) inflow migration to j, such that $\sum_j X_{ij} = X_{i*}$; and $\sum_i X_{ij} = X_{*j}$ and P_i denotes the size of the population living in city i, with i and $j = 1, 2, \ldots, n$.

2.2.1 To migrate or not

The probability of an individual deciding to migrate from city i is estimated by X_{i*}/P_i, which is the frequency of a resident leaving city i. This probability might depend on city size and it is detected by fitting a power law equation:

$$X_{i*} = \alpha P_i^{\beta}, \tag{2.1}$$

where α and β are parameters to be determined from the data (and then expressed as $\hat{\alpha}$ and $\hat{\beta}$ respectively).

Equation 2.1 is a functional form that does not assume that the probability of migrating either increase or decrease with city size. Instead, this is

a data-driven model so that the data provide evidence supporting whether the probability of migrating increases with city size, if $\hat{\beta} > 1$, referred to as superlinear (Bettencourt et al., 2010), decreases, if $\hat{\beta} < 1$, referred to as sublinear, or if it is independent, if $\hat{\beta}$ is close to one.

The exponent $\hat{\beta}$ is adjusted from the entire dataset and a sublinear behaviour is detected of the probability of migrating, with $\hat{\beta} = 0.8829 \pm 0.0147$, i.e., the probability that a person moves away from their city decreases as the size of the city increases. Moreover, the coefficient of $\hat{\alpha} = 0.1676$, indicating that the probability of migrating from a city ranges between 0.023 (as for New York City or Los Angeles) and 0.047 (for instance, in Carson City in Nevada or Victoria in Texas).

The results indicate that individuals from the smallest cities (say, with less than 100,000 inhabitants) are twice as likely to migrate than individuals from cities with more than 10 million inhabitants.

Patterns of human migration are quite variable among cities of different sizes and so noise is a relevant issue. The scaling equation detects a generalised pattern but it does not mean that all individuals from smaller cities have a higher probability of migrating than individuals from large cities. When equation 2.1 is fitted, the adjusted R^2 is 0.9033, meaning that there are other aspects which determine the individual probability of migrating (for instance, age) which in turn determine the collective frequency of migrating from each city. However, a general pattern in which individuals from smaller cities are more likely to migrate is, nonetheless, detected.

2.2.2 Migration to and from other cities

Having decided whether or not to migrate, the decision to migrate to a particular city of a given size is also affected by the population size of the origin city. For instance, if only individuals who used to live in a small city are selected (say with $50,000 \leq P_i \leq 200,000$) and fit equation 2.1 looking at the size of the cities to which they migrated, a sublinear behaviour with $\hat{\beta} = 0.8060 \pm 0.0263$ and adjusted $R^2 = 0.7101$ is encountered. A similar

sublinear behaviour is found when using a different "small city" threshold, for instance, for individuals who used to live in a city with less than 500,000 inhabitants ($\hat{\beta} = 0.8224 \pm 0.0206$ with adjusted $R^2 = 0.8061$), less than one million inhabitants ($\hat{\beta} = 0.8363 \pm 0.0175$ with adjusted $R^2 = 0.8554$) or other thresholds within that range (see the Appendix, Table B.1).

In contrast, a superlinear behaviour is found ($\beta > 1$) when fitting equation 2.1 but considering only the destination of individuals who used to live in "large cities" and decided to move, that is, if selecting only people who used to live in cities with a population larger than a certain threshold. For instance, for $P_i \geq 5$ million, a slight superlinear behaviour is found, as $\hat{\beta} = 1.0499 \pm 0.0337$, with adjusted $R^2 = 0.7163$. This behaviour gets more pronounced with a larger threshold, so that $\hat{\beta} = 1.1688 \pm 0.0506$ (with adjusted $R^2 = 0.5814$) with $P_i \geq 8$ million and $\hat{\beta} = 1.2984 \pm 0.0619$ (with adjusted $R^2 = 0.5327$) with $P_i \geq 10$ million (see the Appendix, Table B.1). This means that an individual who used to live in a large city (i.e., $P_i > 5$ million) is more likely to move to an equally large city than to move to a small city. Thus, individuals tend to preserve city size when deciding to migrate: an individual from a city with several million people is almost twice more likely to move to a city with several million people as compared to an individual from a small city and similarly, individuals from the smaller cities are more likely to move to equally small cities.

Migration patterns can also be analysed in terms of the influx of population into a city, interpreted as the arrival of people per 1,000 inhabitants. Although we have found that individuals who live in large cities are more likely to move to a large city the next year, that does not necessarily mean that the influx of people who arrive into a large city come from equally large cities since it depends on the distribution of city size. Thus, by fitting the power law equation

$$X_{*j} = \alpha P_j^{\beta} \tag{2.2}$$

the inflow of people who move to city j is now considered.

Again, taking into account only the influx of people who move to a city with population in the range $50,000 \leq P_i \leq 200,000$ (i.e., a "small city"), a sublinear behaviour is observed, with $\hat{\beta} = 0.7997 \pm 0.0299$ (with adjusted $R^2 = 0.6492$) and a similar sublinear behaviour when using a different population range, for instance, for the influx of people who move to a city of less than 500,000 inhabitants ($\hat{\beta} = 0.8159 \pm 0.0245$ with adjusted $R^2 = 0.7429$). In general, the impact of the sublinear behaviour gets more pronounced (that is, $\hat{\beta}$ gets much smaller than one) for intervals with smaller cities (see the Appendix, Table B.2).

In contrast, a superlinear behaviour is found when looking at the influx of people who move to a "large city". For instance, the analysis of the influx of people who moved to a city with more than 8 million people, shows a superlinear behaviour with $\hat{\beta} = 1.1180 \pm 0.0460$ (with adjusted $R^2 = 0.6053$) and similarly looking at the influx of people who moved to a city with more than 10 million inhabitants with $\hat{\beta} = 1.2539 \pm 0.0574$ (with adjusted $R^2 = 0.5538$). Roughly speaking, 1.7 people in every $1,000$ inhabitants of a city with millions of people (such as Los Angeles) will have lived in a small city during the previous year, but nearly 20 people in every 1,000 in a small city will have lived in a different small city the previous year.

2.2.3 Migration to and from the countryside

A person who lives in a city might decide to migrate to the countryside, and this decision is affected by the size of the origin city. By fitting a power law equation (equation 2.1) it is found that a person who currently lives in a city might decide to move to the countryside and, according to the data, the probability of moving has a sublinear behaviour, with ($\hat{\beta} = 0.6846 \pm 0.0273$, $\hat{\alpha} = 0.7214 \pm 0.3464$ and with adjusted $R^2 = 0.6199$). Thus, results show that a person who lives in a city with less than 200,000 people is four times more likely to move to the countryside than a person who lives in a city with 20 million inhabitants, such as Los Angeles or New York City.

Also, a person who currently lives in the countryside might decide to

move to a city and a scaling pattern for the size of their destination follows a sublinear behaviour ($\hat{\beta} = 0.5971 \pm 0.0342$, $\hat{\alpha} = 0.4299 \pm 0.4331$ and with adjusted $R^2 = 0.4421$). Thus, a person who currently lives in the countryside is six times more likely to move to a city with 200,000 inhabitants or less than to a city with 20 million people.

2.2.4 Migration to cities from another country

The destination of international migrants is also affected by the city size of the destination. An individual who arrives in the US from another country is more likely to move to a large city, that is, international migration also exhibits a superlinear behaviour. Larger cities in the US increase their population diversity, measured simply as the proportion of people who previously lived outside the US (Page, 2010), three times faster than smaller cities, ($\hat{\beta} = 1.1884 \pm 0.0339$, with adjusted $R^2 = 0.7610$) with an even more pronounced pattern for people from Africa ($\hat{\beta} = 1.5794 \pm 0.0728$, with adjusted $R^2 = 0.5500$) and Americas outside the US ($\hat{\beta} = 1.2808 \pm 0.0424$, with adjusted $R^2 = 0.7036$).

The inflow of international migrants for every 1,000 inhabitants varies according to the size of their destination. Thus, comparing the whole range of city size, it is observed that large cities with millions of people are increasing their percentage of the population from Africa and from Americas outside the US, 32 times and 5 times faster respectively than the smallest cities (results available in the Appendix, Table B.1 for the results divided by continent of origin).

2.2.5 Migration patterns

Fitting a scaling equation and considering the destination of people who lived in small cities results in a sublinear scaling pattern in terms of their probability of moving and their destination, whether it is considered “small” to be cities with less than 200,000 inhabitants or even less than 1 million. Similarly, considering only people from the “large cities”, where the term

"large" can be cities with more than 6 million people or more, gives a superlinear result in terms of the destination picked by its migrants. Thus, there is a phase transition between a sublinear behaviour for small cities to a superlinear behaviour in the case of large cities and, in general, this pattern tends to get more pronounced at the extreme values of city size, that is, $\hat{\beta}$ gets smaller for the smallest cities and more substantial for the largest cities.

Additionally, by considering the inflow of migrants, we observe a sublinear pattern for small cities and a superlinear pattern for the large cities. Thus, there is also a transition for the influx of migrants into a city.

Migration patterns, either the decision to move to another city, move to a small city given that the person lives in an equally small city, the inflow of people who move from another city or from another country are all influenced by city size, either the size of the origin or the destination city and some of the patterns presented here are sublinear and some superlinear (Figure 2.1).

The observed phase transition occurs roughly for cities between 1 and 5 million inhabitants. Below that, cities follow a sublinear pattern and above that, cities follow a superlinear pattern in terms of the destination picked by migrants.

Detecting a sublinear pattern in the destination picked by migrants required grouping cities with a population smaller than a certain threshold and to analyse the observed pattern from the whole group. Thus, it is by grouping cities with a similar population size that we are able to detect an emergent pattern.

To analyse the migration data, not just for "small" or "large" cities, an algorithm was executed, which takes a ranked list of the cities according to their size, using a logarithmic scale, and creates non-overlapping partitions using a moving window of various ranges and with a random starting point. This gives groups of similar cities in terms of their population size but varying what it is considered to be similar. The cities were then partitioned

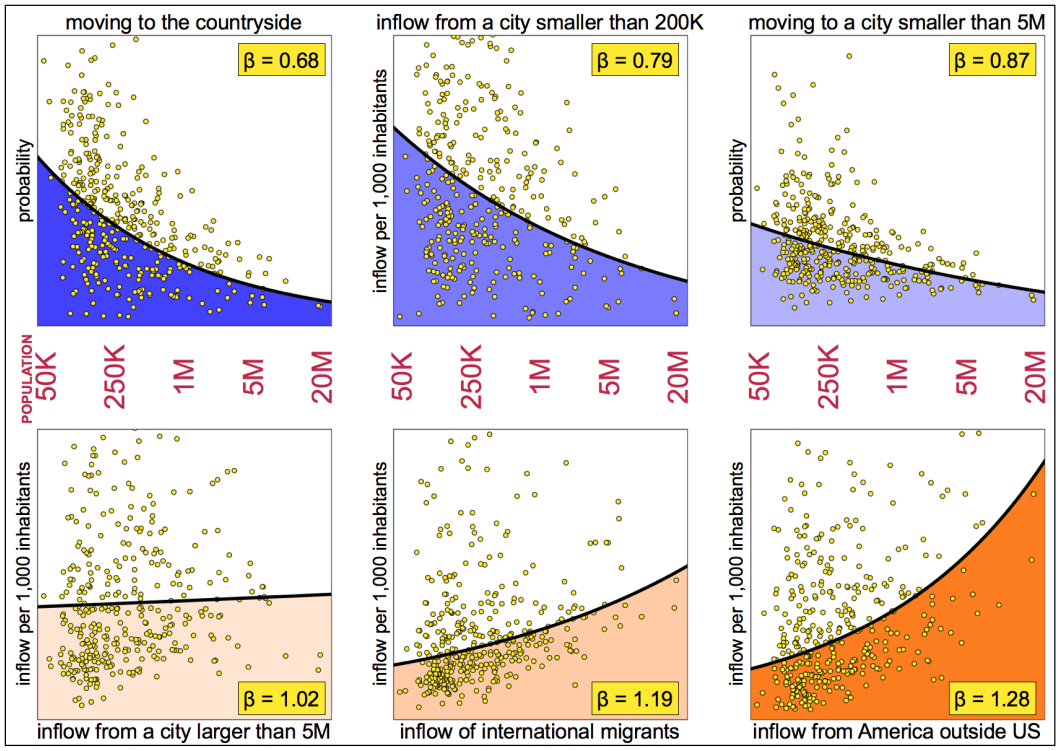

Figure 2.1: Selected scaling relationships fitted to the data. The dots represent data on each of the 385 cities in the US with its size on the horizontal axis and its corresponding figures for migration given on the vertical axis given in different units (as a probability or as the inflow of migrants per 1,000 inhabitants). Also plotted on the same diagrams are the results of the scaling relationship fitted to the data with the coefficient $\hat{\beta}$ given in each case. Top panel: three sublinear relationships. Bottom Panel: three superlinear relationships. A coefficient $\hat{\beta} \approx 1$, as established in the diagram on the left in the bottom panel (for the inflow from a city larger than 5 million), means that city size has a negligible impact on that flux. This establishes the approximate city size where a phase transition occurs.

1,000 times, each time considering a partition with a different starting point and a different width, such that on each partition, cities are grouped based on slightly different criteria. For instance, one partition might consider an interval $I_1 = 270,000 \leq P_i \leq 355,000$ while another time cities might be partitioned in such a way as to create an interval $I_2 = 290,000 \leq P_i \leq 390,000$. Thus, on each run of the partitioning process, particular cities might be grouped in different ways. Then, taking into account the destination picked by migrants from the different cities within each interval, we obtain the scal-

ing coefficient $\hat{\beta}$ by fitting equation 2.1 for each interval of cities. Intervals with no cities inside are ignored.

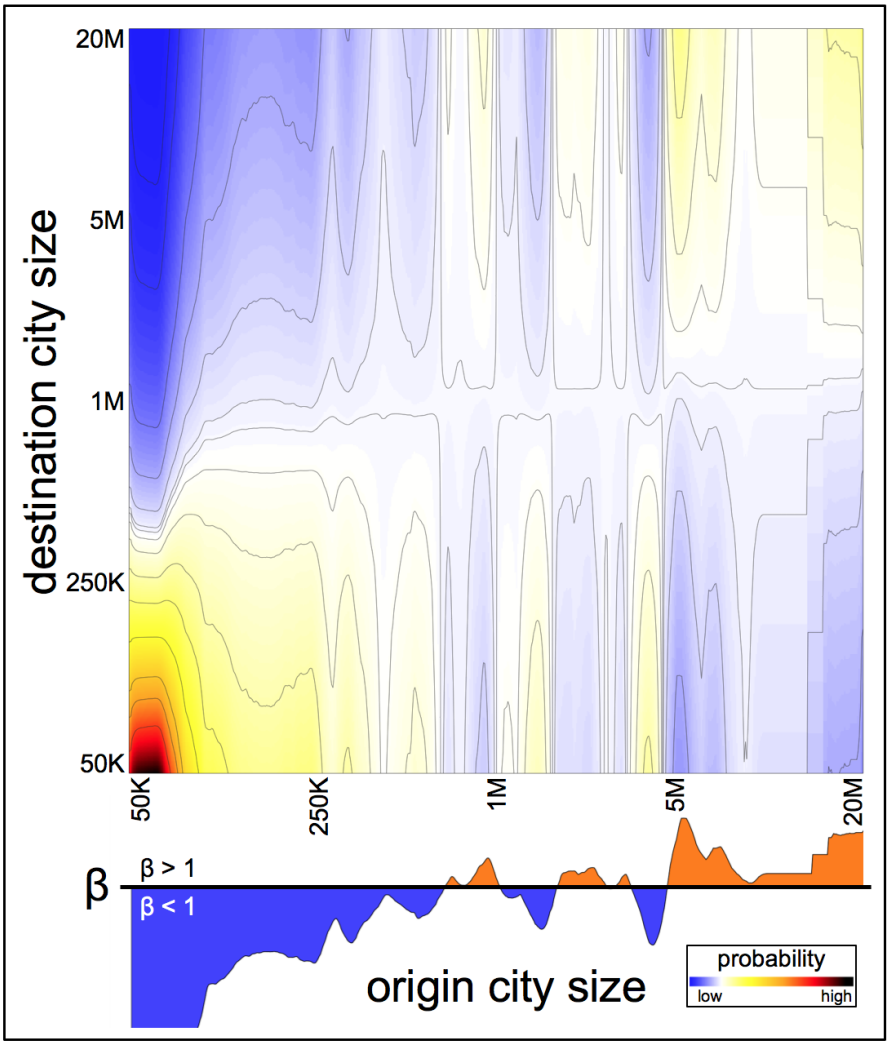

Figure 2.2: Probability of migrating conditional on city size, from a city of a given size (horizontal axis) to a city of a given size (vertical axis). The fitted values of $\hat{\beta}$ according to the city size (plotted in the lower panel) indicates if the probability of migrating to a destination with a given size follows a sublinear ($\hat{\beta} < 1$, in blue) or superlinear ($\hat{\beta} > 1$, in orange) behaviour.

The result after grouping 1,000 times the cities was roughly 33,000 intervals, and so the scaling equation was fitted this number of times; then, for each point in the population range, its corresponding values of $\hat{\beta}$ were averaged. Finally, for each point in the whole population range, an estimated value of the $\hat{\beta}$ is obtained, which smooths out any possible decision

of considering different ranges of city size. The results of the $\hat{\beta}$ for each population range gives a stable and consistent way of estimating the scaling pattern observed for the cities in the US (Figure 2.2).

For example, if the city of origin is larger than 4 million inhabitants (Figure 2.2), then the probability of migration follows a superlinear behaviour. In contrast, a strong sublinear behaviour is observed for small cities, particularly if the city has less than 100,000 inhabitants.

The size of the destination, picked by a person who chooses to migrate, depends on the size of the origin city. For example, the probability that an individual from a city with less than 100,000 inhabitants moves to a city with less than 100,000 inhabitants is 44 times larger than the probability that they will migrate to a city with 10 million inhabitants or more. The resulting relationship is thus given by a $\hat{\beta}$ coefficient which captures the probability of moving to a city with any size according to the size of the origin (Figure 2.2).

2.3 Models for the dynamics of human migration

2.3.1 Scaling model

City size plays a strong role in determining the patterns of migration: from the decision of whether to migrate or not (sublinear), whether to migrate to the countryside (sublinear), move to a small city (sublinear), move to a large city (superlinear) and in the destination for international migration (superlinear). Equation 2.1 determines the estimated probability that an individual living in a city with population P_i migrates from one year to the next (given by $\alpha P_i^{\beta-1}$). Figure 2.2 shows the relationship between the city size and the frequency of migration by considering the probability of each destination, given that an individual actually migrates. These two relations fully determine the dynamics of migration between different cities.

To take people from the countryside (51 million people) into account

within the model, it is considered that with a probability $\hat{p} = 0.0322$ an individual will migrate from the countryside to a city from one year to the next and their destination city follows a sublinear behaviour ($\hat{\beta} = 0.5971 \pm 0.0342$). Also, an individual who currently lives in a city might decide to move to the countryside with a sublinear probability ($\hat{\beta} = 0.6846 \pm 0.0273$). These relationships fully determine the dynamics of migration between the countryside and the different cities.

A two-step process is considered, by the simulation of the dynamics of internal migration observed in the US with people moving between different cities or between the countryside and the various cities. The Markov property is assumed, so that an individual's choice to migrate, as well as their destination, are based only on the current location (that is, the size of their city for people who live in a city or the fact that they live in the countryside). In the first step, it is simulated, for each individual, whether they migrate or not, while in the second step, the destination of the ones who have chosen to move is determined. Since both steps are affected by the observed scaling laws, migration is modelled as a decision problem (Schwartz, 1973). Assuming no deaths or births and ignoring international migration (both arriving and leaving the US) the population dynamics is fully determined.

The impact of city size in the migration pattern is summarised in Figure 2.3. A person from the countryside decides to migrate to a city (with probability $\hat{p} = 0.0322$) and its destination is picked following a sublinear pattern. A person from a city might migrate to the countryside (with a probability that decreases sublinearly with city size) or might decide to move to another city (with a probability that decreases sublinearly with city size too) although in that case, the destination is picked according to the city size of the origin and destination (Figure 2.2). Finally, a person who arrives from another country picks their destination following a superlinear pattern according to city size.

The observed scaling laws of human migration are used for modelling the migration process by considering the distribution of US population living

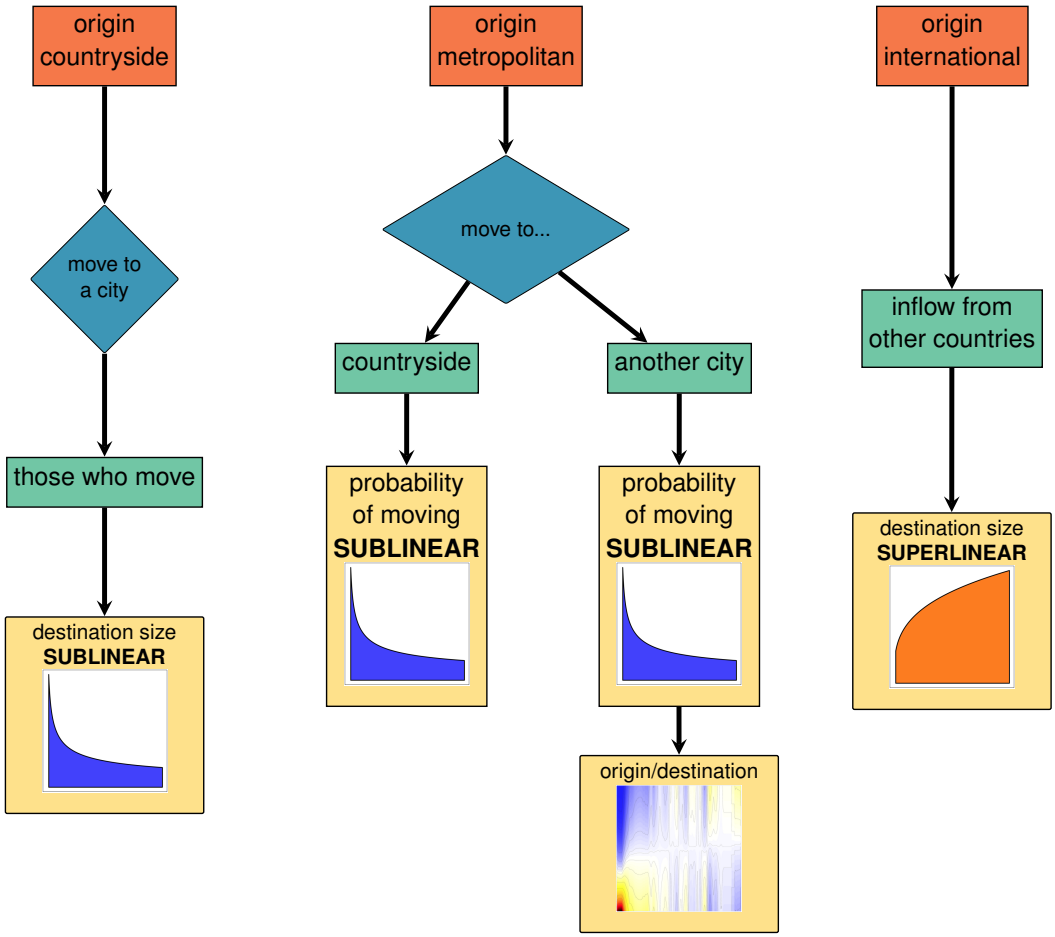

Figure 2.3: Model of the migration dynamics.

in different cities (83% of the US population) and the population living in the countryside (17% of the US population) and to consider the corresponding urban dynamics (Batty, 2007).

2.3.2 Impact of distance and the gravity-scaling model

Undoubtedly, physical distance has an impact on human migration (Schwartz, 1973) which is not considered in the scaling model, so far. Thus, using only city size as a variable to determine the probability of migrating and the destination picked by those who actually move, it is expected, for instance, roughly the same number of people moving to Los Angeles from Stockton-Lodi (a city in California with 684,000 inhabitants, located 500 kilometres away from Los Angeles) as those from Charleston (a city in South Carolina with 680,000 inhabitants, located 3,500 kilometres away from Los

Angeles) simply because both, Stockton-Lodi and Charleston have (almost) the same population. However, data shows that 7.5 more people moved from Stockton-Lodi to Los Angeles than from Charleston. Physical distance is indeed relevant.

The *law of migration*, developed looking at migration at county level from and to the UK, Ireland and Scotland states, among many key issues (Ravenstein, 1885), that the majority of migrants move a short distance. For more than a century there has been quantitative evidence that distance is one of the key aspects of migration and, in general, migration is inversely proportional to the distance between two locations.

As a consequence, one of the most commonly used models of human migration is called the *gravity model* (due to the similarity with the concept of physical gravity, in which objects are attracted to each other with a force directly proportional to their mass and inversely proportional to the distance between them) (Anderson, 2010; Lewer and Van den Berg, 2008). The gravity model predicts the flux of migrants F_{ij} between locations i and j as

$$F_{ij} = \frac{aP_iP_j}{d_{ij}^b}, \tag{2.3}$$

where a is a constant which needs to be estimated from the data, b is a constant which takes into account the impact of distance and d_{ij} is the geographic distance between the two locations, which in the current study, are cities in the US, although the model has been used to estimate the flux of migrants between countries, as for instance (Westerlund and Wilhelmsson, 2011; Karemera et al., 2000). Although the gravity model provides a valuable starting point for the analysis of migration, it has several drawbacks, for instance, it predicts the same flux from i to j as it predicts from j to i; it assumes a linear impact of the population of each location on the flux; in some cases it predicts more people leaving a location than the number of people in the location and other issues (Simini et al., 2012). There are

some modified versions of the gravity model which remove the linearity or the symmetry of the flux (Burger et al., 2009) but one of the main issues to consider when using the gravity model is that it ignores any scaling factor and so all modified versions of the gravity model assume that individuals from small and large cities behave the same and have the same probability of migrating, as opposed to the results demonstrated earlier based on data.

To rectify both models, the scaling model which considers the sublinear and superlinear properties, and the gravity model which considers the impact of distance, a modified scaling model, a *gravity-scaling model for human migration* is constructed by modifying the destination picked by migrants. It takes into account the impact of distance and it also considers the scaling factor observed in the probability of migrating and the preferential destination picked by those who actually move.

Consider a person from city i, with population P_i who has decided to migrate. According to the scaling model (Figure 2.3) the probability that the person moves to city j, say π_{ij}, follows a scaling pattern with some β (which could either be greater than one, if i is a large city, smaller than one if i is a small city or close to one if i is near the phase transition, according to Figure 2.2). The modified probability of moving from city i to city j, π'_{ij} is considered as

$$\pi'_{ij} = C\frac{\pi_{ij}}{d_{ij}}, \tag{2.4}$$

where d_{ij} is the geographic distance between cities i and j, and $C > 0$ is a number which makes the set of probabilities π'_{ij} sum to one. Although other expressions of the gravity model of migration consider the impact of the distance squared, or other functions of the distance, not necessarily linear (Burger et al., 2009; Westerlund and Wilhelmsson, 2011; Simini et al., 2012), here it is assumed that the probability that the person will migrate between two cities decreases as the distance between them increases. Note that the fact that it is a set of probabilities (i.e., they have to sum to one) means that the distance causes a non-linear impact.

The gravity-scaling model takes into account the observed scaling probability that a person will decide to migrate and takes into account the preferential migration observed between people from small or large cities and it also takes into consideration the impact that the physical distance has on the migration patterns. The gravity-scaling model gives the same results as the scaling model for the migration to and from the countryside and for the inflow of international migrants as the distance between a specific city and the countryside or a continent is not well defined.

2.4 Results of the new migration models

The fit of the power law equation (equation 2.1) is, in most cases, good (see the Appendix), as expressed by the high adjusted R^2 obtained from the data.

To determine the validity of the results of the scaling and the gravity-scaling model (as measured by how well it fits the observed data), the results are compared against the commonly used gravity model of human migration. The two parameters of the gravity model ($\hat{a} = 2.59 \times 10^{-6}$ and $\hat{b} = 0.753$) were estimated by minimising the mean square error of the model. The results of the scaling model and of the gravity-scaling model are obtained by simulating the model dynamics 100 times, considering 53.2 million people at each time (20% of the total urban population) who first decide whether or not to migrate and then choose the destination, both according to their city size. The median of the 100 simulations is reported.

Under the scaling model dynamics, 3.1% of the metropolitan population of the US migrates each year. Also, a random person from the cities in the US lives in a city with 4.92 million people, but after migration, it is expected that they will live in a city with 4.81 million people. Ignoring births and deaths and international migration, 80.5% of the movers went to a city with less population than their origin.

The observed migration between every pair of cities is compared and the predicted migration by the gravity, scaling and gravity-scaling models

and report the mean square error and the maximum error in absolute value in Table 2.1. Ignoring the impact of distance (as in the scaling model) shows large departures from the observed migration flows, but also, ignoring the scaling factor on migration (as in the gravity model) yields on large errors. The gravity-scaling model has the best results in terms of the fit of the migration flux (see Table 2.1).

model	mean square error	max error
scaling	102,112.4	15,547
gravity	82,278.8	25,929
gravity-scaling	58,288.8	9,592

Table 2.1: Results of the scaling, gravity, and gravity-scaling models. Mean square error and maximum error comparing the migration flow considering all pairs of cities as origin and destination. The smallest mean square error and the smallest maximum square error (in absolute value) are provided by the gravity-scaling model.

Also, the outflow and inflow of migrants from each city provided by the three migration models are compared. Results show (see Figure 2.4) that the gravity model, as opposed to the scaling model, has a systematic bias and underestimates the outflow of migrants for the smaller cities (those with the smallest outflow of migrants), as it ignores the fact that people from small cities are more likely to migrate, as described by equation 2.1. The gravity model also underestimates the inflow of migrants to small cities and this is mainly because the gravity model also underestimates the outflow of people from small cities which have preferential migration to equally small cities. In general, the gravity model has a systemic issue related to small cities, which is corrected by the scaling model. The gravity-scaling model, similar to the scaling model, takes into account the fact that people from small cities are more likely to migrate and so it does not present any systematic bias as the gravity model does.

When determining the validity of the scaling and the gravity-scaling model, note that internal migration from and to the countryside and international migration should be also taken into account. The scaling model

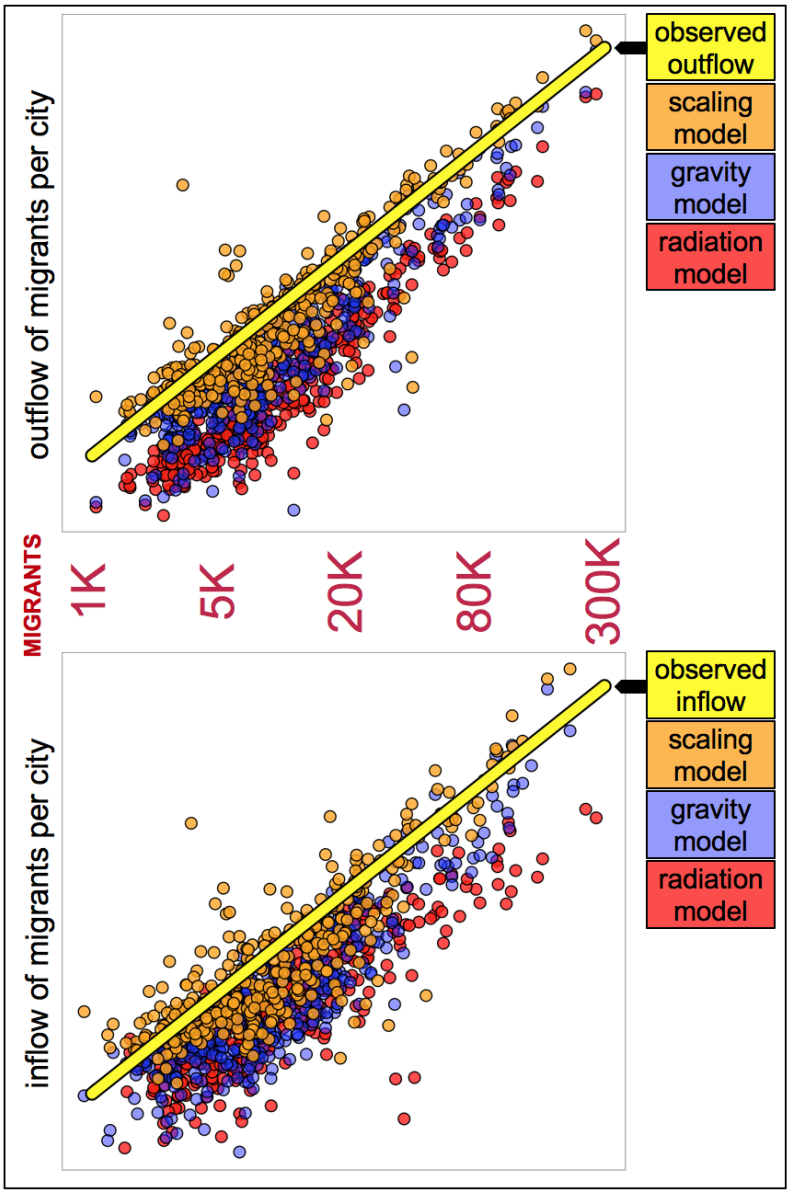

Figure 2.4: Observed outflow and inflow of migrants from each city against the predicted values of the models. The horizontal axis is the observed outflow or inflow of migrants from each city and the vertical axis is the prediction. The yellow line represents the identity (where the predicted value of the outflow or inflow of migrants from each city match the observed values, so there is a perfect match), so that observations closer to that line have a better fit.

predicts that roughly 1.51 million people will move from the cities to the countryside and they will more frequently be from the smaller cities, whilst 1.64 million people will move from the countryside to a city and they are more likely to move to smaller cities. The destination picked by people who

move from the countryside to a city has a sublinear behaviour: the scaling model predicts that less than 109,000 people from the countryside moving to the four largest cities of the US, simply because people from the countryside are more likely to move to small cities than large cities. Similarly, considering people who used to live in the countryside who then moved to the smallest 100 cities of the US, the gravity model predicts less than 67,000 movers, when in fact there were nearly 190,000 people moving. In this case, the scaling model predicts 156,000 movers, which is a much better fit.

The comparison of the three models reveals that ignoring the distance between cities (as it is done by the scaling model) does not provide a better fit in terms of the mean square error. However, a relevant issue is that the scaling model does take into account that people from small and large cities behave differently and therefore, does not have a bias, as opposed to the gravity model.

International migration is also affected by city size. Although it is not possible to determine the impact of the size of the origin city and it is not possible either to compare against the gravity model, the data does allow us to measure the scaling of international migration based on the destination. According to the scaling model, nearly 1% of the population of the largest cities lived in a different country the previous year, thus, increasing the diversity and multiculturalism of cities like New York, Los Angeles or Chicago. Large cities are increasing their diversity three times faster than small cities.

The scaling model is based on a set of observations in which noise is a relevant issue, so that a generalised pattern is detected (for instance, people from small cities have a higher probability of migrating) but it does not mean that all individuals from all small cities have a higher probability of migrating. The scaling model and the gravity-scaling model do not provide deterministic results. By simulating several times under the same dynamics, both models provide natural departures which could be observed under the same dynamics. For instance, between Houston and Dallas, there were

14,666 migrants and results from the 100 simulations of the gravity-scaling model show that a migration flux between 14,485 and 15,745 is expected under the same dynamics.

The power law correctly describes many aspects of the human activity, from the frequency of family names, the wealth of the richest people (Newman, 2005), the sizes of town and cities (Pumain and Guerois, 2004), the distribution of travelled distances (Wang et al., 2014; Brockmann et al., 2006) and now, the probability of migrating from a city, the probability of moving to the countryside, the probability that a person from a small city moves to a small city (and the other relationships indicated in Figure 2.2) and also the size of the city picked as the destination for international migrants are part of the list. Scaling laws play a fundamental role in the dynamics of migration.

2.5 Remarks

Internal migration is far more frequent than it is often assumed, and it is much more complicated than simply people from the countryside moving to the larger cities. Internal migration is highly influenced by economic activity and other environmental factors (Mansury and Shin, 2015; Garcia et al., 2015). People often leave the large cities to move to smaller ones or to the countryside. The destination picked by migrants is influenced by many factors, from the distance between the origin and the destination, to the economic activity, the employment rates and even unfavourable weather conditions (Henry et al., 2003).

Whilst migration is a topic which generates strong public opinions and significant media interest, it is frequently portrayed without much evidence and data to support the arguments. There is an urgent need for the debate about migration and, in particular, international migration, to be based on facts and for policies to be designed based on the observed phenomena rather than on the misguided opinions and news related to migration.

This points out the relevance of considering different levels of migration since they contribute towards strategic economic planning and development (Garcia et al., 2015). Indeed, the majority of the migrants move only short distances (Ravenstein, 1885) and 'gravity' is indeed a concept which roughly explains some parts of the migration patterns (Anderson, 2010).

2.5.1 An improved model of human migration

The initial scaling model examined human migration without considering the physical distance between cities, that is, only considering the city size. This stance is supported by the data which indicates that people from large cities are more likely to move to other large cities, despite the fact that large cities can be far away from each other and relatively scarce. The gravity-scaling model considers also the impact of distance and so it could be considered a modified version of the gravity model. The gravity-scaling model has a better fit to the observed data, is not symmetric, does not have a systematic bias (as can be observed in the gravity model) and takes into account scaling (from the probability of migrating to the preferential destination picked by migrants).

By considering scaling on migration patterns, the commonly used gravity model is considerably improved, highlighting the relevance of city size. A valuable aspect of both, the scaling and the gravity-scaling models is that rather than providing a deterministic number for the flux between two cities, they give a procedure to simulate migration providing intervals which could be observed under the same circumstances. Both models begin by taking into account the number of inhabitants of a city and simulate whether individuals move and, if so, where do they move to and therefore, there is a natural upper limit to the estimated number of people who leave that city, as opposed to the gravity model which might, under certain circumstances, estimate more people leaving actually live there.

2.5.2 A data-driven model of human migration

The scaling and the gravity-scaling models are based on current observations of migration but it does not mean that the same pattern has been observed previously, nor does it mean that the same pattern will be observed in the future. However, the methodology presented here allows scaling to be taken into account from and to the countryside, between cities and from international migrants and it goes beyond a static result observed only for a specific time interval and a particular region of the world. It highlights that in the studies of migration patterns, scaling might occur, it does so without assuming that scaling happens. In the case in which the exponent $\beta \sim 1$ the scaling might be negligible.

2.5.3 A large city versus a small town

Living in a large city may mean an improved access to education, job opportunities and income, among other "benefits", but on average and it does not mean a better education or income to all; however, the costs of living in a large city is experienced by all its inhabitants. The population living in Kibera, for instance (a slum in Nairobi, Kenya, with approx 1.2 million slum dwellers) or Rocinha (the largest favela of Rio de Janeiro) enjoy a limited number of the benefits of living in a large city but they pay the price for longer commuting distances, a higher price for the food and services, pollution, crime rates and more. Thus, although large cities provide certain benefits, more people moving into large cities does not necessarily translate to people enjoying a better standard of living, but might, unfortunately, translate into greater inequality and severe socio-economic problems within the cities.

The same applies to people from smaller cities. Take, for instance, the case of Carson City, one of the smallest cities in the US, where nearly twice the amount of people moved to Redding than to Sacramento (both in California), even though Sacramento is nearly 100 kilometres closer to Carson City than Redding is and Sacramento is 12 times larger in terms of population

size. According to the gravity model, 20 times more people moving to Sacramento than to Redding would be expected since it is larger and closer, but twice the amount of people moved to Redding than to Sacramento; scaling affects migration. Perhaps this is because Redding is a rural environment more similar to Carson City than Sacramento is, although this also warrants further explanation.

The observed patterns might change and migration to large cities might be consequently affected. For instance, fear of crime and fear of terrorism might discourage people from small cities to move to large cities, but at the same time, might encourage people from large cities to move to smaller cities, which may be perceived to be safer. The current main drivers of migration might be replaced by others, such as technology, an ageing population, jobs being lost due to automation, climate change or disasters, to name but a few. However, the methodology presented allows different scaling patterns to be traced through different time intervals, to be applied to international migration or migration from and to the countryside and the detection of quantitative and qualitative changes in human migration.

2.5.4 The role of distance in human migration

Although distance does play a crucial role in migration, either because of the mental cost of being far from the origin, the lack of information about distant places (Schwartz, 1973) or the actual monetary cost of moving, our results indicate that distance could also be expressed in terms of the lifestyle of the individual and not only in terms of physical distance. For example, the four most frequent destinations for an individual who used to live in New York City are Philadelphia, Miami, Washington and Los Angeles, which are 1,800 and 3,900 kilometres away from New York City in the second and fourth case, respectively. Modern communications and rapid transportation mean that the impact of physical distance is reduced so that in terms of migration, distance is becoming less relevant, while the differences in lifestyle between large cities and small cities or countryside are gaining prominence. There

are several reasons why the scaling laws affect migration patterns. Findings suggest that a relevant cause is that an individual chooses between the lifestyle of a large city or the lifestyle of a small one.

2.5.5 The scaling of international migration

There are still open questions regarding the scaling phenomenon in the case of international migration. Is a person from a large city more likely to move to another country, despite the fact that people from large cities are less likely to migrate? Is the relationship found here, where a person from a small city is more likely to move to equally small cities, also observed for international migration? Unfortunately, there is little information about the origin of international migrants who arrive in the US. However, in terms of their destination, a strong impact of the city size is found, which is even more prominent for people who previously lived in Africa or America but outside the US. An individual is less likely to move to a city in which they have less information (Schwartz, 1973), which might be the reason why people from other countries are more likely to move to a large city.

2.5.6 The scaling of migration in other parts of the world

Although the results obtained here are based on data for migration to and from cities in the US, a similar scaling pattern is expected in other countries, so that we predict that an individual from Paris is less likely to move to the countryside than a person from Tours, a smaller French city; a person from Guangzhou is more likely to move to Beijing or Shanghai since both cities have millions of inhabitants, even though they are at 1,200 and 1,900 kilometres away respectively; and Sidney, Melbourne and Brisbane are increasing their international population at a faster rate than the rest of Australia. Scaling should be relevant for other types of migration and in other regions of the world, although there might be other drivers, for instance, language, weather, conflicts or government-controlled migration policies and scaling might have a different magnitude.

So far, it has been noted that cities affect the decision that an individual takes in terms of migration. Whether it increases the probability that they will migrate, their destination, the inflow or international migrants or others, cities play a key role on migration patterns.

2.5.7 Challenges of migration

Migration might be very positive for both, the sending and the receiving location (Konseiga, 2006). Firstly, it can help adjusting to demographic changes, might ease the pressure of unemployment and lack of opportunities at the origin, but might also help by counteracting the decline of ageing in the receiving countries. Many countries are facing the effects of population ageing, most severely in the EU, where without migration, the number of people in the working age is expected to decrease by millions, but is also expected to increase slightly due to the inflow of migrants from other regions (World Bank; International Monetary Fund, 2015). Migrants construct and encourage a resilient network in case of disasters, providing relief and assistance to their affected communities of origin when needed (Laczko et al., 2009).

According to the World Bank, there are several challenges in terms of migration, which go from fostering and promoting legal migration flows (World Bank; International Monetary Fund, 2015), reducing the impact of the brain drain in the sending countries, protecting the rights and preventing the abuse of migrants, lowering the remittance costs and bureaucratic barriers among many. Thus, understanding the whole migration flow, from the small-distance movements of people near the large metropolis, to the internal migration and to the long distance international migration gives a whole perspective on the challenges, but also on the opportunities that migration provides.

Chapter 3

Rare events and their concentration

Many social (and non-social) aspects tend to be highly concentrated, such as wealth, the population of cities and even citations of scientific papers (Newman, 2005). The high concentration of certain social elements is a pattern which repeats and for which many tools have been developed to quantify its degree of concentration. However, when events are not only concentrated but also have a low frequency, such as the number of terrorist attacks on a city, the number of crimes suffered by individuals and others, the most common metrics tend to fail as measuring tools.

This chapter explains the complexity of rare events and develops a procedure to estimate the distribution and hence, the degree of concentration of rare events. It is based on published research (Prieto Curiel and Bishop, 2016a).

3.1 Events with low frequency

In many different practical contexts, being able to determine a measure of the degree of concentration of a variable is particularly useful. For instance, in the case of wealth distribution, the Gini coefficient has been used in many studies in Economic and Political Sciences. The Gini coefficient is a single-valued number which works as a summary for the whole distribution and it helps us to determine whether a country is moving into a more egalitarian

distribution of income or if its disparity is increasing, thus in a way it is a summary statistic which helps us compare different regions and over different time periods.

Similar comparisons are desirable in alternative contexts, for example, is crime more or less concentrated in specific regions after the introduction of surveillance systems in a city? Are road accidents more spatially disperse in Paris compared to Frankfurt? Are the number of claims that an insurance company receives from their customers being more concentrated this year? The data for these type of question has two characteristics which make it hard to deal with: it is both rare and also highly concentrated. In all these examples, the majority of the observations are equal to zero, but then, if a particular observation is not zero, then it is likely that the actual number is not small; so for instance, many accidents happen at the same place, leading to an accident black spot.

This phenomenon has been studied in different settings; in the case of crime, for example, although a high number of people or houses, in fact, suffer no crimes, those who suffer crime have an increased risk of suffering subsequent crimes (Grove et al., 2012) and as a result, the majority of the observations are equal to zero but then, some observations are far away from being zero (Johnson, 2010b). Another example comes from the study of human mobility patterns, where it has been studied by tracing the consecutive sightings of nearly half-a-million bank notes and also, by following 100,000 mobile users, that most of the individuals travel only over short distances (Gonzalez et al., 2008), which means that an individual is likely to be found only in a handful of different places and, if a threshold was placed on travel, most journeys would fall below the threshold.

Having a measure of the degree of concentration or dispersion of such events is useful since sometimes interventions (such as a policing strategy in the case of crime, or a road intervention in the case of road accidents) might result in the displacement of such events, rather than a genuine re-

duction, thus, resulting in a change of their level of concentration. However, traditional measures of the concentration of a variable (such as the Gini coefficient or the entropy), fail to work as a tool to compare different levels of the concentration or to track structural changes in the way that these events happen, due to their extremely small frequency and their high levels of concentration.

3.2 Data and distribution of a counting process

Here, the focus is placed on a counting process, that is, a variable that reflects the number of events with a certain property is observed, for instance, the number of emails that a person receives during a given day. Let X_i be the number of events that occurred over a fixed time interval, counted over some set $i = 1, 2, \ldots, N$, referred to as *individuals*. This could be, for example, the number of burglaries suffered by the i-th household during the period of one year, say, or the number of insurance claims from the i-th customer or postcode. Assuming that having one unit of these events does not affect future probabilities of having any additional units, the number X_i follows a Poisson distribution with rate $\lambda_i \geq 0$. Under this assumption, the number X_i becomes one observation from a Poisson distribution, which means that if X_i is small or even zero, it could be the result of a small rate, but it could also be (with small probability perhaps) the result of a large rate and it was just good luck, or vice-versa in the case that X_i is large. If a person suffered zero crimes last year, it does not mean that their rate is zero and they will never suffer crime.

It is also assumed that X_i is independent to X_j for $i \neq j$, which might be a strong assumption for the particular context under consideration and needs to be fully examined before moving to the following step. In the case of crime, for example, the assumption of independence is perhaps valid only for large populations, but it will be explored later in this thesis. Now, it is assumed that there is a way in which the N individuals may be col-

lected into $k \geq 1$ distinct groups, where group j say, has Q_j individuals (or equivalently, has a relative size $q_j = Q_j/N$), which have the same rate λ_j, with $j = 1, 2, \ldots, k$. Each one of the N individuals of the whole population belongs to one and only one group, so that $Q_1 + Q_2 + \cdots + Q_k = N$ (or written in terms of the relative size $q_1 + q_2 + \cdots + q_k = 1$). To avoid ambiguous definitions, the groups are ordered by their rate in increasing order, so $\lambda_1 < \lambda_2 < \cdots < \lambda_k$. This type of model is known as a mixture model (Böhning, 1998).

The distribution of a random individual, X_i, might be expressed as

$$q_1 \text{Pois}(\lambda_1) + q_2 \text{Pois}(\lambda_2) + \cdots + q_k \text{Pois}(\lambda_k), \tag{3.1}$$

which means that the individual is allocated into the j-th group (with probability q_j) and then has a Poisson distribution with the corresponding rate λ_j.

The number of groups k is crucial for the mixture model. An easy (but useless) solution is to assign each individual to a different group, however, solutions with larger numbers of groups are less useful since for each additional group, its size and its rate need to be estimated, so this increases the number of parameters of the model. The (non-parametric) maximum likelihood estimator (*mle* or *npmle*) helps to compare between models with a different number of groups, k, and to pick the best (in some sense) amongst them (Böhning et al., 1998) since in this case, no prior information on the number of groups is taken into account (McLachlan and Peel, 2004). Other techniques to estimate the number of groups, using bootstrapping, for example, are also available (Schlattmann, 2005). The model can be easily fitted using the statistical package CAMAN (Computer Assisted Analysis of Mixtures) (Schlattmann et al., 2015) in R (R Core Team, 2014) by considering the observed X_i, with $i = 1, 2, \ldots, N$ (Böhning et al., 1992).

The results obtained are: an estimate of the number of population groups $\hat{k}$, the corresponding rate for each group $\hat{\lambda}_j$, so that the collection

of the rate of each group can be viewed as a vector $\underline{\hat{\lambda}}$, and the relative size of each population group $\hat{q}_j$, also expressed as a vector as $\underline{\hat{q}}$. A goodness of fit test can help accepting or rejecting the distribution obtained (Böhning et al., 1992). A similar procedure using a mixture model has been used in different scenarios (Böhning, 1998), such as road accidents, mapping hepatitis B in Berlin (Schlattmann and Böhning, 1993) and many more examples in epidemiology (Böhning et al., 1998).

Two special cases are interesting from the mixture model. First, if $\hat{k} = 1$ then this means that the best way to explain the observations is simply as a Poisson process with rate $\hat{\lambda}_1$, which is a homogeneous distribution over the whole population. The second case is when $\hat{k} = 2$ and $\hat{\lambda}_1 = 0$, which means that the population can be divided into two groups, the first group has a rate equal to zero while the other group has a non-zero rate, which is a model known also as a Zero-Inflated Poisson Model (Böhning, 1998). Both scenarios, the homogeneous distribution and the Zero-Inflated Poisson Model, might be the result obtained from the mixture model.

The distribution of the rates $(\underline{\hat{q}}, \underline{\hat{\lambda}})$ is powerful by itself since it can be used to simulate different observations under that distribution so that the natural departures from the distribution can be understood. In general, the distribution of the rates is called the *profile*, so for example, for the number of crimes suffered by individuals it is the *victimisation profile*; for the number of crimes committed by every person it is the *criminality profile* and so on.

3.2.1 A concentration metric

The Rare Event Concentration Coefficient ($RECC$) works as a summary statistic and it is defined in terms of the distribution of the rates $(\underline{\hat{q}}, \underline{\hat{\lambda}})$ given by

$$RECC = \frac{1}{2\sum_{i=1}^{\hat{k}} \hat{\lambda}_i \hat{q}_i} \sum_{i=1}^{\hat{k}} \sum_{j=1}^{\hat{k}} \hat{q}_i \hat{q}_j |\hat{\lambda}_i - \hat{\lambda}_j|, \tag{3.2}$$

which is the Gini coefficient applied to the distribution of the rates. The Lorenz curve (Marsh and Elliott, 2008) and the Gini coefficient (Dorfman,

1979) of a distribution are often used as a measure of the concentration or dispersion of a variable, and so here they are applied to the mixture model. It is important to note that it is not the Gini coefficient computed directly from the observations X_i, but rather the Gini coefficient of the distribution of the rates $(\hat{\underline{q}}, \hat{\underline{\lambda}})$. A value of the Gini coefficient closer to zero is interpreted as the process being more homogeneously distributed across the population, and a value closer to one means that the process is more concentrated in some population groups.

The Lorenz curve and the corresponding Gini coefficient of the distribution of the individual rates are comparable between different time periods and over different regions, even in the case in which the number of individuals changes from one region to the other, or the total number of events of the process changes. With this simple tool, it is possible to compare the rates of processes in which there is randomness involved, and determine a useful metric for the concentration of events which are rare and tend to be highly concentrated.

3.2.2 Two scenarios from rare events

Two special cases might be obtained from the $RECC$. The first scenario, if the $RECC = 0$ then this means that the process is homogeneously distributed across the entire population so that every individual has the same rate $\hat{\lambda}_1$. This scenario might happen even when the individuals have different observations X_i since here, the distribution of the rates is considered and not the actual numbers X_i.

The second scenario is the case when from data obtained is a Zero-Inflated Poisson Model ($\hat{k} = 2$ and $\hat{\lambda}_1 = 0$). In such a case, the Rare Event Concentration Coefficient gives $RECC = \hat{q}_1$, the relative size of the group which has a zero rate.

3.3 Confidence intervals and estimates of uncertainty

This section summarises the steps followed to construct intervals for the distribution of the rates. The algorithm is easily executed using R (R Core Team, 2014) using two packages (Schlattmann et al., 2015; Zeileis, 2014) which are available online.

1. From the observed numbers X_i, run the CAMAN algorithm to obtain the estimated number of groups $\hat{k}$, the distribution of rates $(\hat{\underline{q}}, \hat{\underline{\lambda}})$ and its corresponding $\hat{RECC}$.

2. Assume that the estimated distribution is the *true* distribution, from which the N individuals can be simulated, with N being the population size and a mixture model

$$q_1 \text{Pois}(\lambda_1) + q_2 \text{Pois}(\lambda_2) + \cdots + q_k \text{Pois}(\lambda_k), \tag{3.3}$$

   so that first, the group of each individual is simulated and then its observed X_i, following a Poisson distribution with the corresponding individual rate λ_i.

3. Run the CAMAN algorithm on the simulated number of X_i to obtain

$$\left\{ k_{sim},\ \underline{q}_{sim},\ \underline{\lambda}_{sim},\ RECC_{sim} \right\}$$

The simulated $k_{sim}, \underline{q}_{sim}, \underline{\lambda}_{sim}$ and $RECC_{sim}$, provide departures which could be observed by having exactly the same distribution of the X_i (the assumed *true* distribution) but having different observations. By running the same procedure enough times (frequently 100 times) departures from what could be observed are obtained, so intervals for the $k_{sim}, \underline{q}_{sim}, \underline{\lambda}_{sim}$ and $RECC_{sim}$ are easily determined.

3.3.1 Rationale for a homogeneous mixture model

Assuming that there exists only a few homogeneous groups is quite a simplification and depending on the type of event, other assumptions made might be problematic. However, results are given in terms of only a few parameters and so results are easier to manipulate, but more importantly, the interest here is to obtain a global metric for the concentration of rare events. Indeed models which consider unit dependence might be more appropriate for modelling specific cases, such as the number of crimes which are committed in a region or a segment of a street or as a retaliatory process between gangs in Los Angeles (Mohler et al., 2012). Thus, the outcome of models which consider an inhomogeneous rate for each unit of observation is a rate λ_X which might depend on the time, place, individuals and/or gangs considered and indeed, if there is enough information to find a more detailed model instead of a constant rate for each group, the individual model might be much more precise.

The objective is to construct a global metric for the concentration of rare events and therefore, assuming a constant rate for each individual is maybe the best that can be done with the data available, and then, assuming a homogeneous rate for each group gives the best possible metric considering the restrictions mentioned above.

The distribution of the rates $(\underline{q}, \underline{\lambda})$ gives a simple description of the distribution of events so that a global metric for the concentration can be computed, but this should not be used at an individual level. The best way to model, for instance, the crime suffered by an individual during different times of the day, for example, is not by a Poisson distribution with a constant rate; the probability of suffering a robbery of a person whilst commuting back from work is higher than when sleeping or working and this would not be captured by a model with a constant rate.

3.4 Applications of the mixture model and the rare event approach

3.4.1 Volcanic Eruptions

An application of the $RECC$ has been completed via the study of volcanic eruptions. Information about the location the 1,532 different volcanoes in the world and their eruptions is available (Global Volcanism Program, 2013) and here, the number of confirmed eruptions for each volcano between 1966 and 2015 is considered (50 years of confirmed eruptions), giving a total of 1,746 eruptions.

Are volcanic eruptions a rare and concentrated event? In this context, out of the 1,532 different volcanoes, only 315 (around 21%) had an eruption in the last 50 years, yet, those volcanoes which had an eruption in the past 50 years, had on average 5.5 eruptions, meaning that volcanic eruptions are relatively rare and highly concentrated (see Table 3.1).

Eruptions	0	1 to 8	9 to 16	17 to 24	25+
Volcanoes	1,217	249	39	20	7
(%)	79.4	16.3	2.5	1.3	0.5

Table 3.1: Number of volcanic eruptions per volcano in the world between 1966 and 2015.

Results of the mixture model applied to the volcanic eruptions gives a total of $\hat{k} = 6$ groups, so that the 1,532 volcanoes are grouped in an optimal way into 6 groups; the first one has an eruption rate of $\hat{\lambda}_1 = 0$ and a relative size $\hat{q}_1 = 49.9\%$, so that nearly half of the volcanoes are not expected to have an eruption (Figure 3.1). The second group has an eruption rate of $\hat{\lambda}_2 = 0.17$ and a relative size $\hat{q}_2 = 34.2\%$, which means that nearly one-third of the volcanoes expect to have an eruption every 287 years. The group with the highest eruption rate has an eruption rate of $\hat{\lambda}_6 = 36.2$ with a relative size of $\hat{q}_6 = 0.3\%$, meaning that volcanoes within that group expect to have an eruption every 16.6 months. For volcanic eruptions, the $RECC = 0.883$.

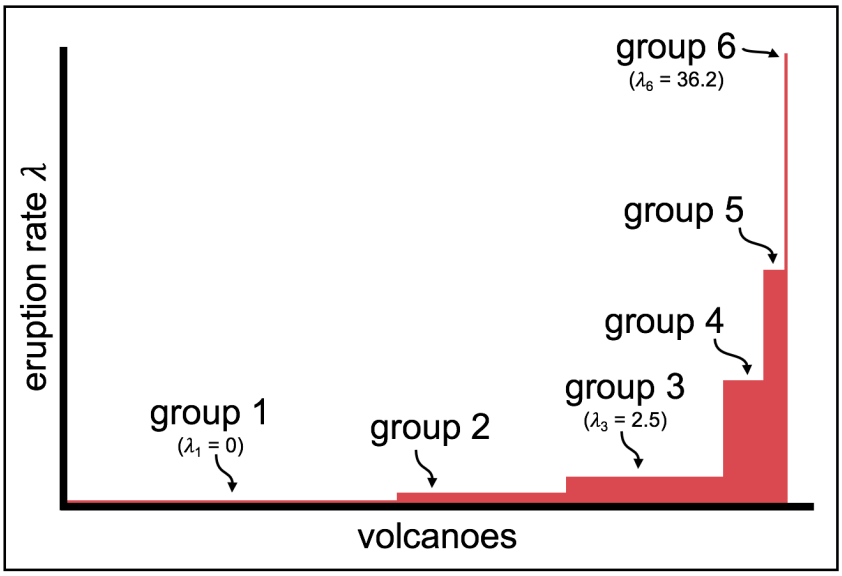

Figure 3.1: Volcanic eruptions profile between 1966 and 2015. The profile shows that a small part of the volcanoes concentrate the largest part of the eruptions between 1966 and 2015, however, more than half of the volcanoes are expected to have some eruption, with $\lambda_i > 0$.

The distribution of volcanoes throughout the world is highly similar to the positioning of the major tectonic belts and so many of the major volcanoes are clustered (Peterson, 1986). For example, the three most active volcanoes during the past 50 years were Etna (with 43 eruptions), in Sicily, Italy; Bezymianny (with 37 eruptions) and Klyuchevskoy (also with 37 eruptions), both in Kamchatka, Russia (see Figure 3.2). Additionally, volcanoes include a variety of cones and craters and some features are destroyed by continuing eruptions (Global Volcanism Program, 2013), which raises the question of how to deal with observations that might be highly correlated? For example, Bezymianny and Klyuchevskoy are 9.7 kilometres apart and so in that small region, there was a total of 74 volcanic eruptions in the past 50 years.

Clustering volcanoes which are at a distance smaller than 10 kilometres apart into *volcanic regions* allows the problem of correlated observations to be dealt with. By considering volcanic regions, so that Bezymianny and Klyuchevskoy in Kamchatka fall into a single region, instead of the 1,532

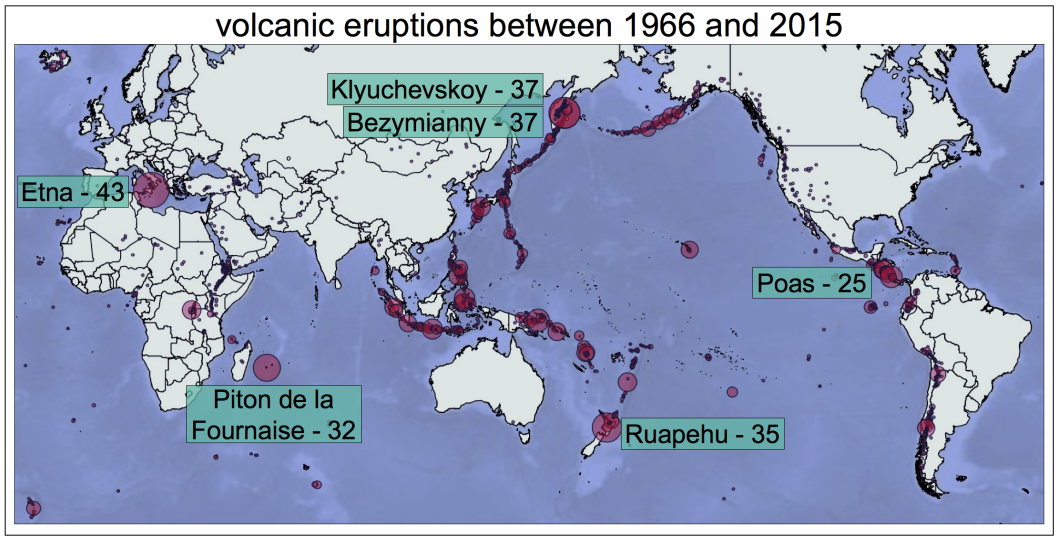

Figure 3.2: Distribution of volcanic eruptions between 1966 and 2015. The size of the disc represents the number of eruptions. Figure made with Natural Earth, free vector and raster map data `naturalearthdata.com` downloaded in August 2015, data from the Global Volcanism Program (Peterson, 1986) downloaded in June 2016 and R (R Core Team, 2014; Hijmans, 2016; Loecher and Ropkins, 2015).

volcanoes, 1,439 regions are obtained, and by taking into account the number of eruptions from each region, the mixture model and the corresponding $RECC$ can be computed. By following this procedure, the number of regions changes, the largest X_i changes (from 43 eruptions of Mount Etna to 74 eruptions in the Kamchatka region) and the mixture model also changes. However, when the 10-kilometre regions are considered, the $RECC$ changes from 0.883 to 0.879 and even to a value of 0.870 when clustering volcanoes into the considerably large regions with a radius of 20 kilometres.

By grouping observations which have a potential statistical dependence based on a physical attribute, such as nearby volcanoes or crimes separated in space within 200 metres (Mohler et al., 2012), groups/regions for which the assumption of independence is fairly reasonable is obtained. Thus, the $RECC$ is relatively stable when correlated observations are grouped based on a physical attribute.

3.4.2 Human Mobility Patterns

Another area in which the $RECC$ might be useful is in the study of human mobility patterns. It has been suggested that the way that individuals move might follow a Lévy flight (Gonzalez et al., 2008) which is a heavy-tailed distribution which might capture longer but less frequent journeys. Different research scenarios have been used, for example by following a large number of mobile users and by recording their position each time they interact with their mobile or by periodically recording their position. The phone towers divide the region into a Voronoi lattice (Okabe et al., 2009) and the data set provides the closest tower to a user so that the location is only recorded by the nearest tower which provides the communication service.

The number of times that a particular mobile user is recorded inside a tower vicinity gives an ideal setting for the study. It is reported (Gonzalez et al., 2008), for example, that from 186 measurements taken from a user, he or she was found to be only in the vicinity of 12 different tower vicinities. Moreover, the pattern of that person shows that nearly 90% of their time is spent in two locations and their neighbouring regions, most likely their house and their office. From the 186 measurements, 96 (51.6%) and 67 (36.0%) occasions happened in the two most preferred locations. In a similar study, some users were found to visit a much higher number of different vicinities (Song et al., 2010), and so the frequency in which users move through different vicinities allows the concentration of mobility patterns to be determined.

By counting the number of times that a user is recorded in different tower vicinities, produces different mobility patterns that users might have (Pappalardo et al., 2015b). The $RECC$ of the tower vicinities counts of different users gives a way to compare their levels of mobility and, for example, a smaller $RECC$ implies that a user has a higher degree of mobility than a person who has a larger $RECC$. A larger $RECC$ indicates that the person tends to move on a day-to-day basis only through a small number of neigh-

bours of their home city. In terms of human mobility, the $RECC$ takes into account the (potentially) highly concentrated nature of the regions in which a person moves, but also a random component which might motivate a person to visit places which they do not regularly attend.

3.5 Remarks

The Rare Event Concentration Coefficient $RECC$ based on the mixture model helps to compare the concentration rate of events which are not frequent and tend to be highly concentrated by taking into account the random nature of such events. Other measurements which are traditionally used for the concentration/dispersion are meaningless since they do not detect structural changes in the process, or they cannot be used to compare different regions or time intervals.

The Rare Event Concentration Coefficient $RECC$ is easy to compute and provides a summary statistic which is comparable and helps detecting structural changes in the dispersion of rare and highly-concentrated events, such as crime, road accidents or human mobility.

The $RECC$ is designed for rare events, so, in general, many zeros are observed, which do not ensure that the rate of the individuals is zero, so if a person, for example, suffered zero crimes last year it does not mean that their rate is equal to zero. The simplest possible model, which is a mixture model based on a Poisson distribution, might frequently have observations equal to zero with a rate being greater than zero.

3.5.1 A new tool for measuring the concentration of rare events

Considering events which have a low frequency, such as the number of crimes suffered by individuals or the number of terrorist events on cities, and constructing the profile of such events (the victimisation profile, in the case of crime) gives two valuable results. Firstly, a precise but simple description of the distribution of the events out of which it is possible to simulate and to

observe the expected departures from the distribution. Things might have a random element which is often ignored, but with the distribution and by simulating events, the probabilistic approach is a natural part of the event.

A typical approach to determine the concentration/dispersion of a variable (for example, using the Gini coefficient) fails to work as a measure of the concentration of road accidents due to their low frequency and their high level of spatial concentration. The methodology presented here, considering the distribution of the rates and the $RECC$, helps to overcome the low frequency of events, taking into consideration their random component and to obtain a distribution from which simulations can be easily computed. From the simulations, expected departures from the observed number of accidents can be detected, including outliers.

3.5.2 Extensions of the concentration coefficient

If events are not as rare, then it is possible to estimate the individual rates using a different technique than the mixture model and that tries to mimic the underlying pattern. For example, consider the rates at which underground stations serve their users, which is best modelled taking into account the hour and the day of the week; for instance, the number of mobile users within the nearest routeing tower vicinity, which might be modelled taking into account the time and space. More sophisticated models for a counting process can also be considered, for example, gang shootings may incite retaliation from rival gangs, and an earthquake increases the chances of a second earthquake, causing, in both cases, a self-exciting process (Mohler et al., 2012). In the latter case, estimating the individual rates, either as a function of time, space and/or past events, gives a much better approximation to reality. Thus, the Event Concentration Coefficient (ECC) can be constructed simply by computing the Gini coefficient of the individual rates, even in the case in which they were estimated using a different model. The resulting metric provides, as in the case of the $RECC$, a number between zero and one which reflects the level of concentration of such events.

Chapter 4

The distribution and concentration of road accidents

Road accidents are one of the main causes of death in the world but yet, road accidents have a low frequency and they tend to be highly concentrated when their spatial distribution is considered. Thus, road accident data is challenging to deal with and poses serious challenges for policy-making.

Here, the distribution of road accidents is modelled as a rare event and the accident profile of a city and of motorways is constructed. As a result, a distribution for simulating road accidents and a metric for the concentration is obtained which in turn, gives valuable insights for decision-making in terms of urban and motorways accidents. It is based on published research (Prieto Curiel et al., 2018a).

4.1 Road accidents

According to the World Health Organization, during 2013, more than 1.2 million people died around the world due to a road accident[1], one of the most frequent causes of death, 2.8 times the mortality due to Malaria and 3.3 times the mortality due to violence. Whilst the number of road accidents is now a global concern, it is, however, possible to either reduce their fre-

[1] Data from the World Health Organization, available at the Global Health Observatory data repository `https://bit.ly/2L89d9g`

quency or their impact: in the UK, for example, the number of road fatalities decreased from an average of more than 3,400 each year between the year 2000 and the year 2004 to an average just above 1,800 fatalities each year between 2010 and 2013[2]. This dramatic decrease in the number of fatalities in the UK indicates that accidents do not simply just occur and that through sensible policies, thousands of deaths around the world could be avoided.

Broadly speaking, road accidents have three potential causes: firstly, it could have something to do with the *driver*. It was shown that the chances of a driver having an accident are many times higher if he or she consumes high levels of alcohol (Horwood and Fergusson, 2000) or is fatigued (Sagberg, 1999) and accidents are considerably more likely to lead to a fatality if the driver exceeds the speed limit[3]. Secondly, accidents might have something to do with the *local environment*, for example, due to reduced visibility, the weather conditions, poorly designed junction, a poorly enforced speed limit, faulty traffic signals and more. Finally, an accident might occur simply due to *(bad) luck*, for example, a non-preventable failure in the car and so on. The first and second causes, attributed to the driver and to the environment, can and should be reduced to a minimum, both in terms of their frequency and their impact.

How to distinguish whether a certain region has an increased probability of accounting for an accident? Clearly, the road geometry, road obstacles and the level of traffic have an impact on the distribution of road accidents, but these tend to remain unchanged for long periods of time and are specific to a certain area so it makes any comparison between different cities, or even areas of a city, quite complicated.

If, for example, data shows a specific junction with several accidents, would that be enough to suggest that it is necessary to reduce the speed limit or put in a road intervention scheme? Is there a threshold as to the

[2]Data from the Department for Transport in the UK, available at `https://bit.ly/1JjD4iJ`

[3]Report from the Royal Society for the Prevention of Accidents `https://bit.ly/2aZioDQ`

acceptable number of accidents that a street or a road could experience and yet still be considered safe?

4.1.1 Heat maps and the random location of accidents

Numerous studies have been conducted to identify the spatial patterns of road traffic accidents and develop techniques to identify crash-prone locations. A frequently used tool to analyse the location of road accidents (as well as other spatially-distributed events, such as the location of crimes or gang fights) is a heat map (Anderson, 2009; Erdogan et al., 2008; Prasannakumar et al., 2011; Steenberghen et al., 2010; Anderson, 2007). This tool provides a graphical description of the location of a point process, which highlights areas or junctions more prone to accidents.

There are, however, two technical aspects with respect to heat maps which are often ignored: when a location is considered to be "hot", what is it compared with? and to what degree is the observed heat map the result of randomness? The relevance of randomness, in terms of its spatial distribution, is that every point process, no matter how it is generated and whatever the underlying distribution, will result in a set of observations being relatively close to each other, thus, even random points (where the term 'random' is used here for a uniform distribution) might be interpreted as having a "hot region" (Figure 4.1).

Although a heat map offers a visual tool for representing road accidents, it might actually result in misleading conclusions when the random element of the location of road accidents is not considered. The crucial difference between a point process that is generated by a uniform distribution and a point process with a different distribution, is frequently undetectable based on a simple visual inspection. A similar situation occurs when a single road is considered, an apparent concentration of accidents will appear, no matter how random or concentrated road accidents are. A formal statistical test against *Complete Spatial Randomness* can be constructed by considering the distance to the nearest neighbour of each point and compare this against

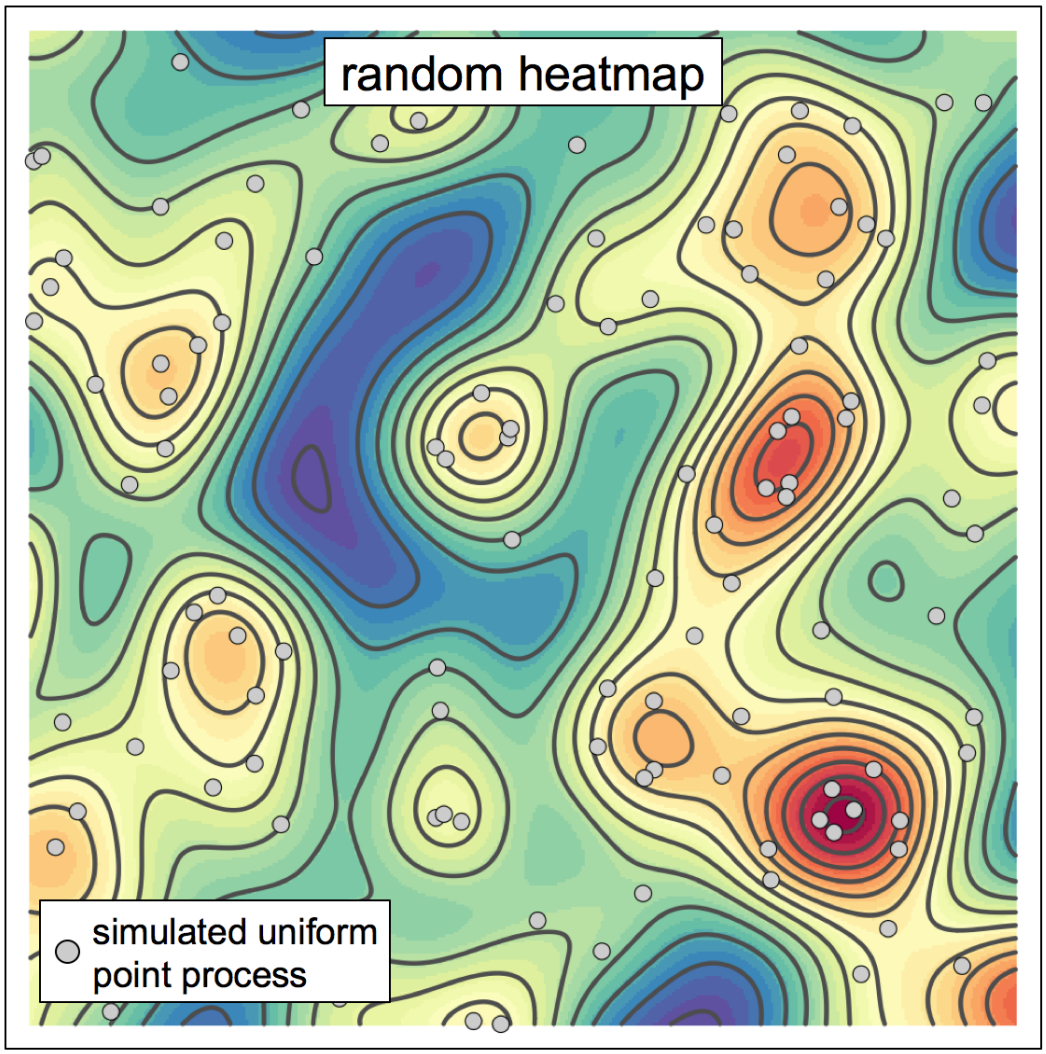

Figure 4.1: Heat map of a simulated point process that follows a uniform distribution. The underlying uniform distribution has the property that every region is expected to contain a number of points proportional to its area, thus, any apparent concentration observed in the map and any region with a higher, or fewer, points is only the result of randomness and not the result of a higher probability of observing a point in that region.

a uniform distribution (Diggle, 2014) and only orient efforts at a specific location when spatial randomness is rejected.

4.1.2 Concentration of road accidents

Road accidents might happen due to a mixture of environmental elements, for example, an obstructed visibility, excessive speed of road users, the curvature or quality of the roads, the street lighting and more. These conditions

perhaps repeat, almost under the exact same conditions, day after day and so it is expected to observe particular road junctions or segments with a much higher number of accidents than others if the environment is the main cause. However, accidents might also happen because of factors related to the driver or simply because of 'luck', and the chances are that interventions oriented to the road rather than the driver would not reduce this type of accident. A natural way to detect whether road accidents might be attributed to elements on the road rather than the driver is through its concentration. If there is an element which increases risk related to the environment, then more accidents would occur in this specific location than elsewhere and therefore a high concentration should be observed.

The degree of concentration of events has been shown to play a crucial role in other aspects, such as wealth (Lorenz, 1905; Yntema, 1933), the population of cities, the size of a forest fire (Newman, 2005) or crime. By considering the victims who suffer crime (O et al., 2017; Hope and Norris, 2013), the offenders who commit them (Wolfgang et al., 1987; Wolfgang, 1983; Martinez et al., 2017) and the places in which crime is executed (Weisburd, 2015; Lee et al., 2017), it has also been shown that crime is highly concentrated. In the specific case of the places in which crime is executed, a "law of crime concentration" has recently been developed (Weisburd, 2015) which provides a relevant reference in the study of crime at places.

Although crime and road accidents are fundamentally different events, they both share a low frequency, a high degree of concentration and the fact that both are, to a certain extent, unpredictable. Thus, both areas of research can utilise the tools developed to deal with their low-frequency but highly-concentrated type of events.

Statistically speaking, one of the things that make road accidents (as well as crime) hard to analyse is their low frequency. In London, for example, the road junction with the highest number of accidents has (just over) one accident every month, which makes them highly unpredictable and sta-

tistically hard to deal with. No relevant pattern concerning the day of the week or the time of the day of road accidents, can realistically be observed when the frequency of such events is so low. Moreover, since road accidents are low-frequency events, it is observed that the majority of road segments (or intersections) suffered no accidents within the period of the analysis. Hence, the Gini coefficient G, which is a popular measure of the degree of concentration (Dorfman, 1979), based on the count of accidents in each road segment, will reveal a high concentration of accidents, even when they are uniformly distributed amongst the segments in which accidents occurred. In other words, the Gini coefficient obtained directly from the count data does not take into consideration the fact that these events are rare, and will naturally regard the data as having a high degree of concentration. As a consequence, the Gini coefficient of low-frequency events might easily be misinterpreted and might make it difficult to compare the concentration of road accidents between cities or different motorways.

4.2 Spatial counts of the road accidents

Two sources of information and two types of analysis are used here to compare the concentration of road accidents. Firstly, data available from the Transport for London (TFL) website[4] allows the spatial concentration of road accidents within a city to be measured. Secondly, data available from the Ministry of Transportation from Mexico[5] allows the concentration of road accidents on motorways to be measured. The type of road accident and data from an urban environment is very different from that taken on motorways and therefore, two kinds of analysis are presented, based on a different discretisation of the observed road accidents.

Road accident data has, in general, two issues. A considerable number of non-fatal injury accidents are not reported to the police and are therefore not included in the available data, however, issues of under-reported

[4] Available at `https://bit.ly/295vkak`
[5] Available in Spanish at `http://www.sct.gob.mx/carreteras/`

accidents are considered minimal in the case of more severe accidents (Savolainen et al., 2011). Also, there might be a lack of precision related to the location of the road accidents, especially in the case of accidents on motorways, as there are fewer reference points. However, no systematic bias on the location of the road accidents should be observed and therefore, concentration metrics give reliable information about the underlying pattern.

4.2.1 Urban data - London

The data from the Transport for London contains information on road traffic collisions that involve personal injury occurring on public highways which have been reported to the police. Data is collected by the police at the scene of an accident or, in some cases, reported by a member of the public at a police station, then processed and passed on to Transport for London. The data, taken between 2005 and 2014, includes 242,782 unique collisions, with x, y space coordinates available. Accidents are subdivided into three categories: *fatal*, where death occurs in less than 30 days as a result of the collision, *serious*, if there are fractures or injuries requiring hospital treatment, and *slight injury*, where the accidents do not require medical treatment. Table 4.1 contains the reported frequencies between 2005 and 2014.

Category	fatal	serious	slight	total
frequency	1,670	27,788	213,324	242,782
%	0.7	11.4	87.9	100

Table 4.1: Observed frequencies of collisions in Greater London between 2005 - 2014

For the purpose of taking into account only the urbanest parts of the city, only the central area of London is considered here, which accounts for 70% of the road accidents registered by TFL occur.

4.2.2 Motorway data - Mexico

The motorway data considered here contains road traffic collisions registered on motorways in Mexico. The data is divided for each motorway and considers, for each accident registered by the police, the distance from the starting point of the highway. Unfortunately, the data does not include in which direction of the road the accident occurred.

The motorways analysed have Mexico City as their starting point, connecting the capital of Mexico with five large cities: Cuernavaca, Toluca, Pachuca, Puebla and Querétaro (Figure 4.2). There are two types of motorways, Federal Roads (free of charge) and Toll Roads and each city has both, a Federal Road and a Toll Road connecting them to Mexico City, except for the case of Querétaro for which the Federal Road first passes through another city (Toluca) and so it is not considered. In total, 9 motorways are considered for the study.

The length of the motorway and the vehicle flow rate is different for each of the 9 motorways considered. Both of these factors become relevant when it comes to studying road accidents. Longer roads or those with a higher number of vehicles are expected to have more accidents even if the risk for a driver is the same as compared to a shorter or less used road. Therefore, the flow, measured in *vehicle kilometre per year* units, makes the risk on each road comparable.

Taking into account the length of the road and the number of cars using it, allows a comparison of different roads to be made. For instance, in Table 4.2 it is possible to observe that the Toll Road between Mexico City and Querétaro has the highest number of accidents between 2015 and 2016, yet, is also the longest road among the 9 considered and has a considerably high vehicle flow. The Federal Road between Mexico City and Cuernavaca, on the other hand, has a higher accident risk and is more lethal (meaning that a driver is more likely to suffer an accident and it is more likely that the accident will result in a fatality) than in any other of the roads considered, but

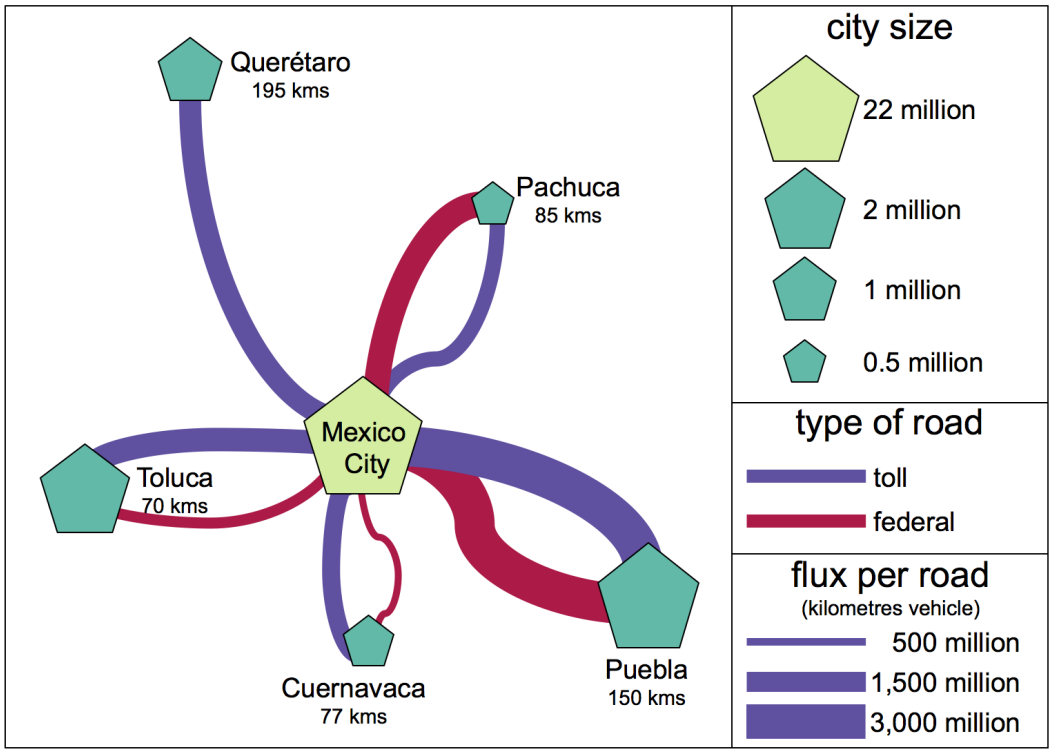

Figure 4.2: Main roads connecting Mexico City. Schematic representation of the nine roads which connect Mexico City and the five main cities in its peripheral region.

it is a short road with a reduced traffic flow and so it does not have as many accidents as the other roads. Thus, comparing the accident risk between different roads has to be based on the length of the road and the number of vehicles that use it or its flow.

The accident risk (number of accidents per vehicle kilometres of travel) and how lethal the accidents are, varies considerably between different roads. The road with the highest accident risk (the Federal Road between Mexico City and Cuernavaca) is actually 12.5 times more prone to accidents and 9.9 times more likely to have lethal accidents than the safest road (the Federal Road between Mexico City and Puebla).

4.3 Methodology for a spatial point process

It is important to determine when two accidents have occurred at the same location. Different levels of data aggregation have been used in previous

Destination	length	flow	accidents	victims	fatal
Cuernavaca Federal	60.5	467.9	105	159	22
Cuernavaca Toll	70.7	1204.7	106	128	23
Toluca Federal	55	764.5	117	90	20
Toluca Toll	55	1589.4	46	45	15
Pachuca Federal	62.5	1761.3	162	161	29
Pachuca Toll	62.5	1083.3	62	99	20
Puebla Federal	121	2736.3	49	63	13
Puebla Toll	121	2725.7	162	309	46
Querétaro Toll	164	1548.8	293	362	64

Table 4.2: Observed frequencies of collisions on the nine motorways which have Mexico City as origin between 2015 and 2016. The *length* of the road is measured in kilometres and *flow in of vehicles* is measured in millions of vehicle kilometres per year.

studies, from countries, provinces, counties, road segments, a point pattern process, road junctions and segments of a road with various lengths (Thomas, 1996).

The hypothesis that road accidents are homogeneously distributed (known as Complete Spatial Randomness or CSR) is easily rejected (Baddeley, 2010) by measuring the nearest neighbour distance for every road accident (Diggle, 2014). A map of where the accidents occurred during the past ten years, in the case of the London data (Figure 4.3), shows a very specific pattern, highlighting main roads and congested junctions.

4.3.1 Discretisation of the data

4.3.1.1 Urban environment

In the case of the urban space, a tessellation of the region of analysis is considered, that is, the city is divided into nearly 30,000 non-overlapping, regular hexagons, and the number of accidents within each hexagon is counted. A hexagonal tessellation is frequently used in cartography since it offers advantages in terms of the visualisation (Birch et al., 2007) and it offers equal-area units and minimal correlation with regularly spaced features, as opposed to a square grid (Carr et al., 1992). Hexagons of side length 40 metres provide a useful level of refinement for our analysis dividing the re-

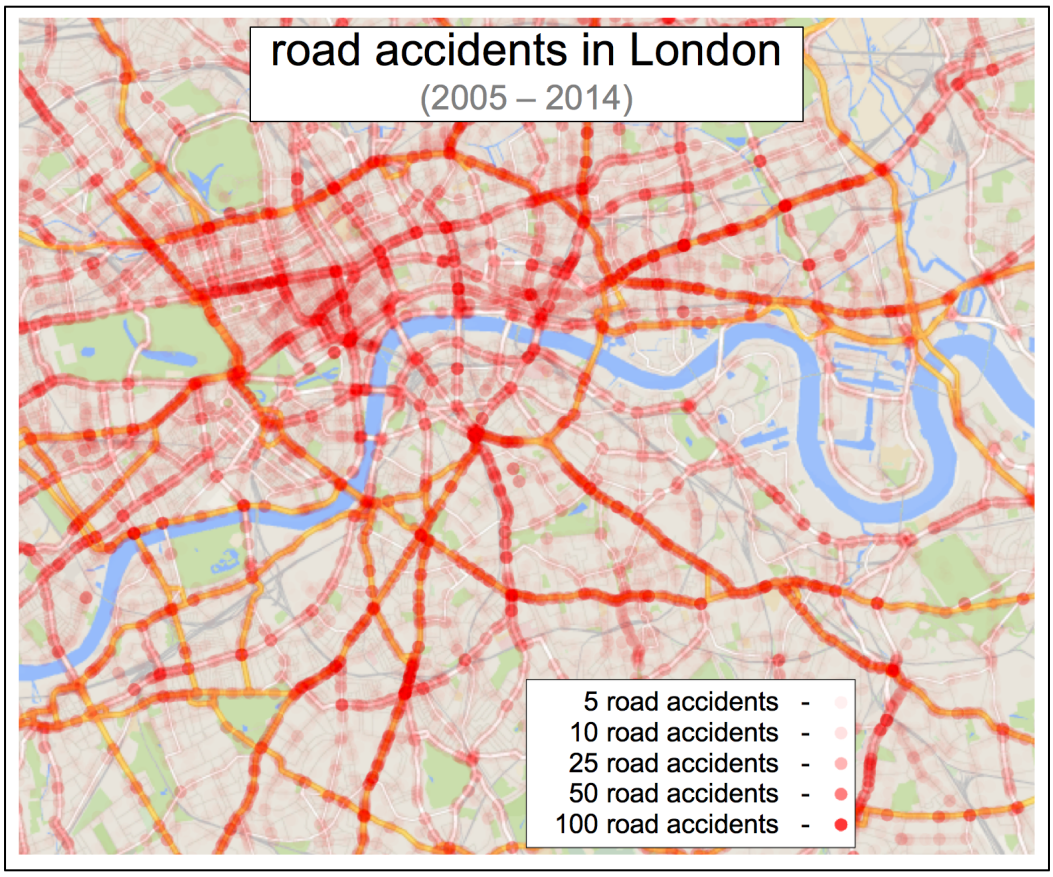

Figure 4.3: Registered Road Accidents in Central London between 2005 and 2014.

gion of central London into $N = 29,600$ tiles. Under this partition, Waterloo Bridge, for example, sits within four hexagons from its extremes on either side of the River Thames. Hexagon tiles are small enough that the region they represent are clearly identifiable and, although they do not match exactly with road junctions, they clearly represent parts of streets. Smaller tiles do not capture the patterns of road accidents and larger tiles tend to blend different regions into the same tile. Also, a similar measure of 40 metres is used for urban data in other studies (Steenberghen et al., 2010; Schuurman et al., 2009), and so this choice is likely to be close to optimal.

4.3.1.2 Motorway environment

In the case of the motorway data in Mexico, the highway is divided into non-overlapping segments of 500 metres and count the number of accidents within each segment. Due to the precision of the data, smaller segments do

not group accidents correctly and larger segments are not refined enough to identify a specific location of a highway. Also, 500 metres has been frequently used in other studies when a highway is partitioned (Erdogan et al., 2008; Bíl et al., 2013), so this level of partition is used for consistency. In addition, although there are some vehicular entrances and exits to the motorways between their origin in Mexico City and their outer destinations, these junctions have a reduced number of vehicles compared to the main roads and therefore, it is considered that through each segment of each motorway, the flow of cars is approximately the same.

Although using either a tessellation (in the case of the urban data) or a segmentation of the road (in the motorway data) has its disadvantages (such as a potential autocorrelation of the number of accidents) it does allow a region to be clearly identified, to cluster the accidents that are nearby and to consider different levels of refinement. Using this partition of the space transforms the data into a non-negative discrete variable, rather than a continuous measurement of the location of road accidents, which is easier to analyse.

Figure 4.4 shows the count of the number of road accidents recorded within each tile and the numbers show that there are many tiles with zero, or close to zero, accidents for the ten year period, but there are also a few tiles with more than 150 accidents. The tiling procedure gives comparable observations in terms of the number of accidents that occur, but not in terms of the risk that a driver experience by travelling across each tile since the number of drivers that go across each tile is significantly different. In fact, Figure 4.4 highlights roads in central London where most casualties occur. If the objective is to explain the reasons why a region has more accidents, a common technique is to divide the number of accidents by the traffic volume, so as to consider the *vehicle miles of travel vmt* (Jovanis and Chang, 1986). However, the objective here is to determine a measure of the concentration of such events.

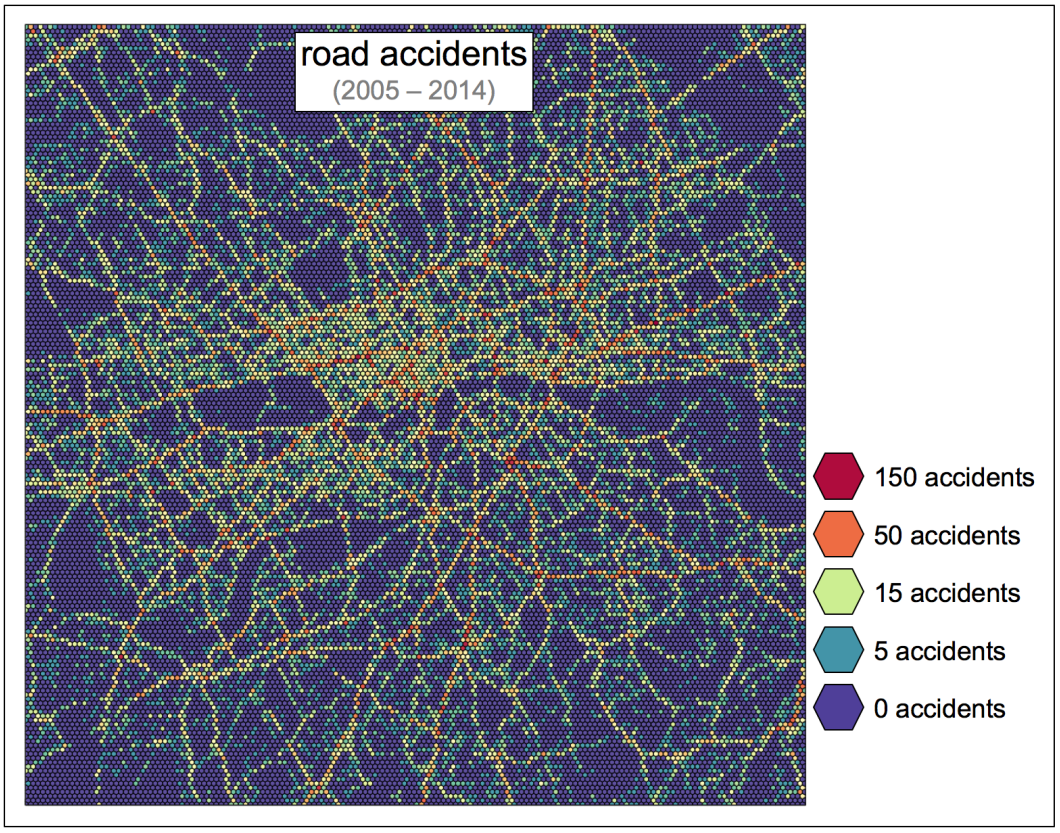

Figure 4.4: Number of accidents in Central London. Partitioning of Central London into 29,600 hexagonal tiles, with sides of 40 metres, and the count of accidents between 2005 and 2014.

4.3.2 Distribution of road accidents

The number of accidents within each motorway segment or within each hexagonal pixel, during a certain period of time (two years in Mexico and 10 years for the London data), might be equal to zero for obvious reasons (for example, for tiles which overlay a river or a park) or might be much higher in regions with a higher volume of traffic (Jovanis and Chang, 1986). If suffering an accident in one region does not affect future probabilities of suffering an accident (which might not be true if a road intervention takes place), then the number of accidents suffered in the i-th segment or region, H_i say, follows a Poisson distribution with rate $\lambda_i \leq 0$, where λ_i is referred to as the *accident rate*, representing the 'speed' at which the i-th segment or region suffers accidents hence, the number of accidents is simply an observation (or a realisation) from that Poisson distribution. An alternative

approach is to use a Negative Binomial distribution (Maher, 1990), by using Survival Theory (Jovanis and Chang, 1989), or other statistical models (Savolainen et al., 2011), but here, instead of trying to explain why a region has more accidents (perhaps through a regression technique) the objective is to measure their spatial degree of concentration, so it is assumed that regions have a different accident rate, without going any further.

Using a Poisson distribution for the number of road accidents observed on each segment has conceptual advantages. Firstly, the expected number of road accidents on a segment is given simply by its rate λ_i. Secondly, it allows us to sum the rates so that the number of road accidents in two segments, i and j, also follows a Poisson distribution, with rate $\lambda_i + \lambda_j$. Finally, the number of road accidents over k years also follows a Poisson distribution with rate $k\lambda_i$. Thus, it is easy to interpret the rate λ_i as the expected number of road accidents in the segment i.

In the case of the urban setting, two adjacent tiles might have similar rates, especially if the same road goes through both of them. In the case of the analysis of motorways, two neighbouring segments might also have similar rates if they experience accidents due to similar causes. Although there is a clear spatial structure that is highly relevant to the problem, it is assumed that each tile has a fixed accident rate.

With this approach, the analysis moves away from the observed count data for road accidents into the analysis of the rates, λ_i, of accidents. What is important is that it is a probabilistic metric, so it considers that a region might have been 'lucky' during one year and have only experienced a few accidents, or it might have been 'unlucky' and had many accidents. If a road segment has no accidents for a year, it does not mean it will never have them in the future, perhaps it is the result of a small rate λ_i. Transforming the observed data of road accidents into probabilities gives a different perspective on its distribution. For instance, consider a road segment i with rate $\lambda_i = 1$, so that exactly one accident is expected each year. There is a

high probability that the segment will have no accidents for one year, given by $\exp(-1) = 0.368$, and there is also a high probability that the segment will have more than one road accident, given by 0.264, which means that departures from its expected value of one accident per year are considered.

4.3.3 Inhomogeneous distribution of road accidents

To model the inhomogeneous distribution of accident rates, it is assumed that the N units (either tiles or segments) can be grouped into $k \geq 1$ distinct groups, where group j say, has a relative size of q_j (or, in other words, the group j has Nq_j units) and each group has an accident rate λ_j, with $j = 1, 2, \ldots, k$. Each one of the N units belongs to one and only one group, so that $q_1 + q_2 + \cdots + q_k = 1$. To avoid ambiguous definitions, the groups are ordered by their rate in increasing order, so that $0 \geq \lambda_1 < \lambda_2 < \cdots < \lambda_k$. This procedure is known as a mixture model (Böhning, 1998) and the (non-parametric) maximum likelihood estimator (*mle*) helps compute the optimal number of groups in which the units are grouped, denoted by $\hat{k}$ (Böhning et al., 1998), the corresponding accident rate for each group $\hat{\lambda}_j$ and the relative size of each of the groups, $\hat{q}_j$. The results of the mixture model (the number of groups, the accident rate and relative size) can be computed using the statistical package CAMAN (Computer Assisted Analysis of Mixtures) by considering the observed number of road accidents suffered in each of the tiles or segments and a test can help us accept or reject the distribution obtained (Böhning et al., 1992).

4.3.3.1 Rare Event Concentration Coefficient $RECC$

The distribution of the rates $(\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_k, \hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_k)$ obtained from the data is useful since it is possible, for example, to simulate accidents within each unit to understand the expected departures that simply a natural variability of the number of accidents would yield. In the case of highways, for instance, being aware of the rate of accidents from its origin to its destination gives a full description of the occurred accidents. However, to detect

a structural change in the accidents, the Rare Event Concentration Coefficient ($RECC$) is used (Prieto Curiel and Bishop, 2016a), where a value of the $RECC$ closer to zero is interpreted as road accidents being more homogeneously distributed across the city, and a value closer to one means that road accidents are more concentrated in some regions of the city. The $RECC$ is a coefficient comparable over different time periods, between different regions and even for different cities or type of accidents.

4.4 Road accidents profile and a metric for their concentration

4.4.1 Concentration of road accidents in urban environment

The Lorenz curve (Marsh and Elliott, 2008) and the $RECC$ for the road accidents in London between 2005 and 2014 are displayed in Figure 4.5 and results indicate that around 47% of the tiles considered have a rate equal to zero (not surprising, since London has lots of parks and a large river passing through it), but also, 33% of the tiles have an estimated rate of $\hat{\lambda}_j = 1.3$, meaning that within the period of ten years, these tiles expect to experience only 1.3 road accidents. These accidents are not considered to be related to the environment, due to the small rate, and so they could have happened anywhere. On the other hand, there are tiles with rates higher than 30 accidents over the ten year period, so they expect to have at least one accident every four months and so on. There is, however, a group of tiles with an estimated rate of $\hat{\lambda}_k = 86.6$, meaning that these tiles expected to have one accident every six weeks.

The level in which road accidents are spatially concentrated is surprisingly high. In Central London, 32% of the accidents happen in only 2.4% of the road junctions, and they get even more concentrated if focus is placed only on the serious and fatal categories. Table 4.3 shows the $RECC$ for the

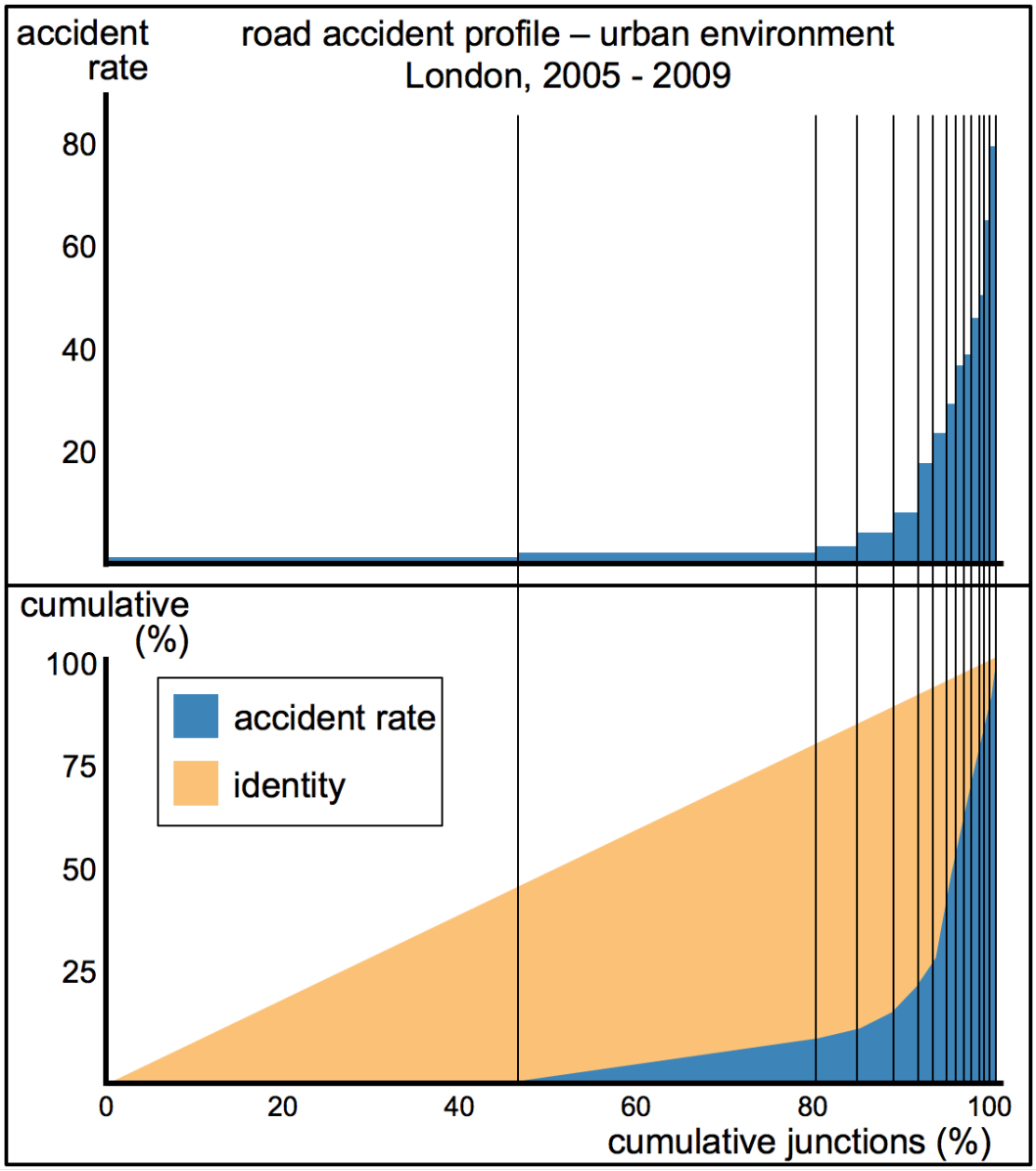

Figure 4.5: $RECC$ of accidents in Central London between 2005 and 2014. Accident rate $\hat{\lambda}_j$ and group sizes $\hat{q}_j$ (above). Cumulative accident rates in blue and the identity in yellow (below). The $RECC$ is represented by (twice) the area between the two curves.

road accidents and results are that fatal and serious accidents tend to be much more concentrated in only a few regions.

Considering only the serious and fatal road accidents, results of the mixture model are that around 64% of the tiles have a rate equal to zero. There are, on the other hand, a few tiles (roughly 0.4% of the surface or 109 road junctions) which have an accident rate of almost 11. This means

Category	Fatal	Serious	Slight	Total
$RECC$	0.8712	0.8198	0.8057	0.8055

Table 4.3: $RECC$ metric of the road accidents in London between 2005 - 2014. A value of the $RECC$ closer to zero means a more homogeneous distribution of road accidents, and a value closer to one means a higher degree of concentration.

that in the small region represented by the 109 tiles, someone is expected to suffer either a serious or a fatal accident every year. Table 4.4 shows the distribution of the accident rates.

Group	1	2	3	4	5	6
size $\hat{q}_j$	64.2	22.7	2.1	1.9	1.7	1.5
rate $\hat{\lambda}_j$	0.000	0.488	0.823	1.159	1.517	1.906
Group	7	8	9	10	11	12
size $\hat{q}_j$	1.3	1.2	1.0	1.0	1.0	0.4
rate $\hat{\lambda}_j$	2.337	2.839	3.466	4.359	5.860	10.950

Table 4.4: Estimated group size $\hat{q}_j$ and accident rate $\hat{\lambda}_j$ for the serious and fatal accidents in London between 2005 and 2014.

Serious and fatal road accidents have a surprisingly high degree of concentration. Results of the mixture model are that nearly half of that type of road accident happen in less than 5% of the tiles considered. However, another relevant component of road accidents is that nearly 25% of the serious and fatal road accidents occur in tiles in which only one accident every twenty years is expected. Perhaps accidents which occur at road junctions which have such a small rate cannot be attributed to the road itself and the chances are that they occurred due to causes related to the driver (such as alcohol consumption, driving when fatigued or more).

The $RECC$ between 2005 and 2014 for the road accidents in London does not show a drastic change in the way accidents are distributed across the city and so a certain stability is observed, despite the decrease in the number of accidents. Results are displayed in Table 4.5.

Tiles with the highest rates in London have specific environmental fac-

Year	2005	2006	2007	2008	2009
$RECC$	0.813	0.825	0.824	0.825	0.826
Year	2010	2011	2012	2013	2014
$RECC$	0.831	0.831	0.828	0.828	0.821

Table 4.5: $RECC$ for all road accidents between 2005 and 2014 in London.

tors which contribute to creating more dangerous roads. For instance, certain Underground stations which are transportation hubs, with a large number of pedestrians, are among the tiles with the highest rate in the city: such as Elephant and Castle, Hyde Park Corner and Camden Town. Also, some roads with a high flow have a consistent high accident rate, such as Euston Road and Kingsland Road (the A10 which is a main arterial road) and finally, relevant commercial streets are also among the locations with the highest accident rate, such as Oxford Street.

4.4.2 Concentration of road accidents on motorways

For the Mexican motorway data, comparing the distribution of the accident rates in the nine highways separately reveals that each road has a different pattern. In the case of the Federal Road between Mexico City and Puebla, the $RECC = 0.022$, meaning that the accidents are distributed almost following a uniform distribution along the whole road. This, however, does not mean that the road expects fewer accidents, but it means that from the origin to the destination, the accident rate remains practically the same at $\hat{\lambda} = 0.1978$. One way to interpret this, since the units of observation are segments of a road with 500 metres length and two years of data are considered, is that every 10 years a segment expects to observe one accident. Alternatively, for every 1,263 metres, one road accident is expected every year, irrespective of where on the road this measure is started from.

Road accidents are rare events and there is a need to use adequate tools to deal with them. The Federal Road between Mexico City and Puebla has the lowest possible degree of concentration, but it is only when the $RECC$ is computed that the uniform pattern is detected. A frequently used

metric to determine the concentration is the Gini coefficient. Unfortunately, computing the Gini coefficient directly from the number of accidents observed on each road segment is not adequate due to the abundance of observations with zero accidents. There is a correlation of -0.956 between the Gini coefficient computed in this manner and the average accident rate of the road (the number of accidents divided by the length of the road) meaning that both, the Gini coefficient and the average accident rate give the same information and do not provide any information in terms of the concentration. For instance, looking at the Gini coefficient, directly from the number of road accidents, in the Federal Road between Mexico City and Puebla gives a value of $G = 0.8211$, in which case, the wrong interpretation would be that accidents in that road are highly concentrated. Furthermore, the Gini coefficient evaluated for the Federal Road between Mexico City and Puebla turns out to be the highest among the nine roads considered here, and hence it can be wrongly concluded that on this road the accidents are more concentrated than on any other road (although looking at the rates, a uniform pattern is observed).

Accidents have a low frequency and so, in the case of the Federal Road between Mexico City and Puebla, only 49 accidents are distributed along 242 units of 500 metres (121 kilometres of road) meaning that, due to the low frequency of road accidents, at least 79.7% of the observations are equal to zero. In general, the low frequency of events (high count of observations with zero events) increases the Gini coefficient: the share of events for a great part of the population is zero, thus meaning more inequality in their distribution. However, by taking into account the distribution of the rates of road accidents in the Federal Road between Mexico City and Puebla and not just the number of road accidents, results show that almost every segment of that road has the same accident rate and there is practically no concentration of accidents along that road.

Another consequence of the low frequency of accidents is that the Gini

coefficient computed directly from the number of accidents tends to give similar results between different roads, with small or negligible differences between them and, in the worst case scenario, with the wrong results and interpretation (Prieto Curiel and Bishop, 2016a). The Gini coefficient of the roads with the lowest $RECC$ (the Federal Road between Mexico City and Puebla, with a Gini coefficient of $G = 0.8211$) and the road with the highest $RECC$ (the Toll Road between Mexico City and Cuernavaca, with a Gini coefficient of $G = 0.7401$) shows that using the traditional Gini coefficient, the wrong interpretation that *road accidents have a lower concentration on the Toll Road between Mexico City and Cuernavaca* would be obtained.

Other roads also have a certain degree of uniformity with regards to their accidents. The Federal Road between Mexico City and Pachuca, for instance, has $RECC = 0.274$ and three types of segments are identified: 17% of them have an accident rate equal to $\hat{\lambda}_1 = 0$, while 74% of them an accident rate of $\hat{\lambda}_2 = 1.36$ and a small segment of the road, 8.2% a rate of $\hat{\lambda}_3 = 3.4$, meaning that the majority of the road is to a certain extent risky and only 17% is risk free.

The nine roads in Mexico have a different rate distribution of their accidents (Figure 4.6). From a $RECC$ close to zero, observed in the Federal Road between Mexico City and Puebla, to a $RECC = 0.559$ observed for the Toll Road between Mexico City and Cuernavaca.

Computing the $RECC$ for the nine roads altogether (Figure 4.7) shows that nearly 30% of the segments have an accident rate of $\hat{\lambda}_1 = 0$ but also, that there are two groups with a high rate. The Federal Road between Mexico City and Pachuca has a set of road segments of five kilometres (not necessarily contiguous) with a rate of $\hat{\lambda}_q = 3.44$, meaning that there is a small number of segments which consists of 5 kilometres of the road in which we expect to observe 17 road accidents each year. Also, on the Toll Road between Mexico City and Cuernavaca, there are two segments (so, one kilometre in length) which have an accident rate of $\hat{\lambda}_r = 5.05$, much higher than

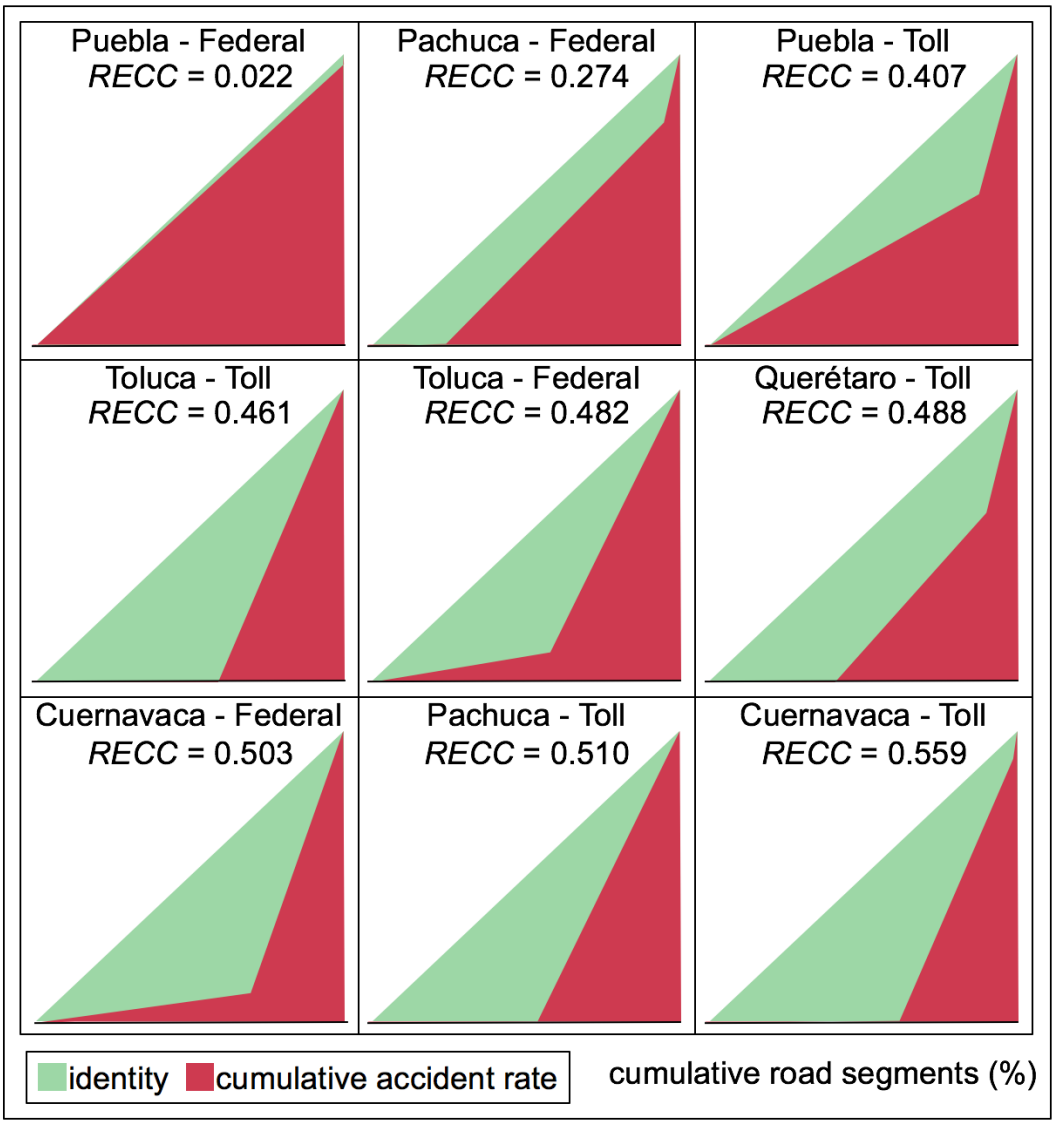

Figure 4.6: *RECC* of accidents in each of the nine main motorways with origin in Mexico City.

in the rest of the nine roads. On that specific kilometre (again, not made of contiguous 500 metres segments) the expected number of accidents each year is more than ten.

Environmental factors that contribute to the chance of having an accident can be identified in the road segments with high accident rates. For instance, among the highest rate segments ($\hat{\lambda}_q = 3.44$) of the Federal Road between Mexico City and Pachuca, segments situated at kilometres 35, 72 and 75 are found, where the first segment is located within an urban area,

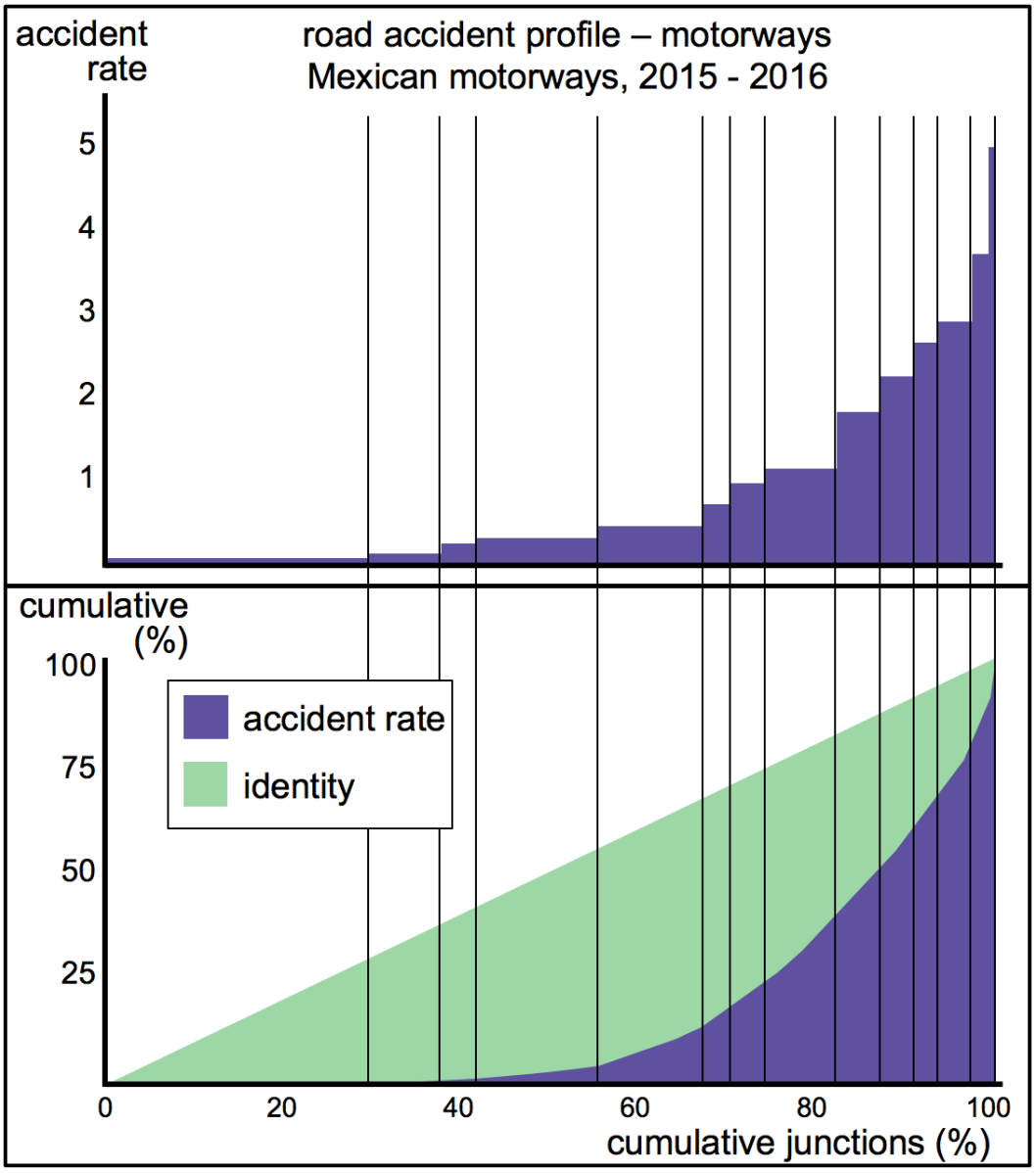

Figure 4.7: *RECC* of accidents in the nine main motorways with origin in Mexico City considered altogether. Accident rates $\hat{\lambda}_j$ and group sizes $\hat{q}_j$ (above). Cumulative accident rates in purple and the identity in green (below). The *RECC* is represented by (twice) the area between the two curves.

and the other two segments have junctions. Likewise, the two segments with the highest rate of accidents ($\hat{\lambda}_r = 5.05$) of the Toll Road between Mexico City and Cuernavaca are situated at kilometres 56 and 64, in a part of the road with a steep slope (7%), with the second segment having a turning place on the high-speed lane (left lane). These local factors: the presence

of junctions, the steep slope (7%), and the turning place on the fast lane are likely to be responsible for the high rate of these segments.

The accident rate along these two higher rate sections have been identified using the CAMAN procedure and the $RECC$ and this high concentration is attributed to the environment and an intervention to determine whether it is related to the road conditions, its visibility, its design or speed limit should take place. In total, in these two segments (one on the Federal Road to Pachuca and the other one, on the Toll Road to Cuernavaca) which are less than 0.7% of the 772.3 kilometres of roads considered, there are 4% of the road accidents.

In addition, the different values of the $RECC$ observed on the nine roads which originate in Mexico City are not the result of longer roads (so a higher number of observations) or a higher flow, nor as a result of a higher number of accidents, but due to other environmental reasons. For instance, the Toll Road to Toluca and the Federal Road to Puebla both have had less than 50 accidents in two years (in fact, they are the two roads with the lowest number of accidents) but the $RECC$ in the first case is 0.461 and in the second case is 0.022, meaning that even when they observe a similar number of road accidents, they follow a different concentration pattern.

4.5 Remarks

4.5.1 Concentration of road accidents

A typical approach to determine the concentration/dispersion of a variable (for example, using the Gini coefficient) fails to work as a measure of the concentration of road accidents due to their low frequency and their high level of spatial concentration. The methodology presented here, considering the distribution of the rates and the $RECC$, overcomes the low frequency of events, taking into consideration their random component and also it provides a distribution from which simulations can be easily computed. From the simulations, expected departures from the observed number of acci-

dents can be detected, including outliers.

Results for the urban environment show that road accidents are highly concentrated, especially those that fall into the serious and fatal category. This result could be useful for policymakers: by focusing their resources on less than 5% of the road junctions, they are considering the regions where nearly half of that type of accident occurs.

Results for the motorway environment show a much smaller concentration degree. In the case of the Federal Road between Mexico City and Puebla, road accidents are considered to be distributed almost uniformly along the road, meaning that statistically speaking, they have the smallest possible concentration. Also, the procedure introduced here, including the use of the $RECC$, allowed a comparison between different roads and a higher accident rate in two specific segments of the highways was observed (one on the Federal Road to Pachuca and the other on the Toll Road to Cuernavaca). On these specific sections, accidents might be closely related to environmental factors and so perhaps, some of these accidents could have been avoided by a road intervention scheme, such as a reduction in the speed limit.

4.5.2 Urban planning and road accidents

For a city planner, a quantitative tool such as the $RECC$, derived from the mixture model, provides the ability to compare between different severities or over different time periods to determine the effectiveness and impact of a safety program. For events, such as road accidents, which are rare and have a high degree of concentration, a tool which allows valid comparisons between different cities becomes a valuable asset enabling to learn from past experiences.

The ability to identify regions of a road or a city which have environmental factors that increase the risk of an accident enables infrastructures to be re-designed accordingly. For instance, in the case of London, these results might be used to justify the plans to transform Oxford Street, one of the Lon-

don's roads with the highest accident rate and with the highest number of road fatalities, into a pedestrian street[6]. Having identified a segment of a road which puts its users at a higher risk due to its environmental factors, means that something can (and should) be done to reduce that risk.

4.5.3 Other applications in the analysis of accidents

The methodology presented here could be easily applied to other types of accidents by adjusting the parameters. For example, the tiling procedure could help a risk manager to identify whether there are regions in some industrial complex with an increased rate of an accident, and the $RECC$ can be used for purposes other than the analysis of accidents, for example, by monitoring the number of people who required the assistance of the coastguard along different parts of the shoreline or it could be used by an insurance company to determine any changes observed in the distribution of accidents.

[6]More information about the transformation of Oxford Street into a pedestrian road is available at `https://bit.ly/2h8yOhA`

Chapter 5

Quantitative measurements of crime

A mathematical model for the number of crimes suffered by the individuals, committed by criminals or the spatial distribution of crimes, requires (as in the case of road accidents), to take into account their low frequency and their observed high degree of concentration. Crime is also a rare event, and that has severe implications in terms of security strategies and crime prevention. Only a few individuals tend to commit most of the offences, a few places (such as street segments) concentrate most of the crimes, and a few individuals tend to be the victims of crime more frequently.

In this chapter, a distribution of the number of crimes suffered by individuals is analysed considering its low frequency and high level of concentration. The methodology allows constructing a distribution of the rates in which individuals suffer crime which is similar to the observed phenomena, allows simulating the number of crimes suffered by individuals and constructing a metric for comparing the concentration of crime across different regions or time periods. It is based on published research (Prieto Curiel et al., 2017a).

5.1 Concentration of crime

The risk of suffering a crime is not uniformly distributed over a region (Johnson, 2010b) nor is it uniformly distributed across members of the same community (Grove et al., 2012), that is, some regions and some population

groups are more affected by crime than others (Farrell, 2015). In the case of burglary, for example, it has been shown that houses in deprived areas suffer a higher risk of being the target of a crime whilst other regions appear to be immune to that type of crime (Bowers et al., 2005). Just as there are places in which crime is concentrated (Freeman, 1996), there are population groups with a higher risk of suffering a crime (Farrell and Pease, 1993). Data shows that there is a high level of concentration of crime, whether the place in which crime is committed is being considered (Freeman, 1996), which can be as specific as a shop (Brantingham and Brantingham, 2010), the victims who suffer the victimisation, the criminals who execute crime or the time at which the crime takes place is considered (Johnson, 2010b).

As a result of the risk heterogeneity, crime is highly concentrated in certain population groups (O et al., 2017). For example, the analysis of data from the British Crime Survey (Bureau of Justice Statistics, 2016) shows that 2% of people who suffer the highest number of personal crimes, in fact, suffered 66% of the total reported for that type of crime (Pease, 1998). This has often been attributed to the *attractiveness* of a place or a person (Brantingham and Brantingham, 2010) and the interaction between potential offenders with potential targets, amongst many other reasons and theories (Stark, 1987).

The reasons why individuals experience different crime rates have been considered in depth by others (Tseloni and Pease, 2004) and explanations go from individual attributes, which cause an increase in the attractiveness; lifestyles that exposed certain people to a higher risk of victimisation (Hindelang et al., 1978); a boost on the probability of suffering a second crime after suffering an initial crime (Johnson et al., 2009). The focus of this study is the distribution of crime itself, so here different individual rates are assumed, without going any further into this topic.

5.1.1 Frequently used metrics for crime concentration

The traditional and most frequent measurement of the concentration of crime suffered by a population is provided by the average number of crimes suffered by a victim. Formally, we consider a population of size N and let V be the number of people within the population who suffered a particular type of crime during a given time period (usually one year) and let C be the number of times that a particular type of crime was committed. The *victimisation rate* (v), also known as *prevalence*, is defined as

$$\text{victimisation rate} = v = \frac{V}{N}, \tag{5.1}$$

and the *crime rate* (c), also known as *incidence*, is defined as

$$\text{crime rate} = c = \frac{C}{N}. \tag{5.2}$$

Based on these two measures, a frequently used measure is the *concentration of crime* (H), given by the ratio of the crime rate c to the victimisation rate v, that is

$$H = \frac{\text{crime rate}}{\text{victimisation rate}} = \frac{c}{v} = \frac{C}{V}. \tag{5.3}$$

Assuming that each crime is assigned to a single victim, $H \geq 1$, and so H is a measure which can be interpreted as the average number of crimes suffered by the victims of that type of crime. H has been used, for example, to measure the concentration of burglary (referred to in that case as the *burglary concentration*) (Tseloni et al., 2004). Although H is frequently used to measure the concentration of crime, it has severe flaws because it does not help us determine if crime is, in fact, more or less concentrated. Using a simple example it is possible to see how it is a poor summary of the crime suffered: consider two populations, A and B, both with a population size of $N = 100,000$, and both suffering the same crime rate of $c = 0.1$ and

the same victimisation rate of $v = 0.05$ so that in both populations there are the same number of victims ($V = 5,000$) and the same number of crimes ($C = 10,000$); hence, the same concentration of crime $H = 2$ is obtained, so that victims suffer an average of two crimes. However, consider the artificial situation in which for the population A each victim suffered exactly two crimes, but in the population B there were 4,000 victims suffering a single crime each and 1,000 victims suffering 6 crimes each. Clearly, this single measure of $H = 2$ does not differentiate the construction of these two distinct scenarios where, in the population B, the 1% of the population who suffers the highest amount of crime actually suffers 60% of the crime which is a completely different behaviour than that observed in the population A.

An adequate metric to determine the statistical dispersion of crime is to consider the Lorenz curve and its corresponding Gini coefficient G (Fox and Tracy, 1988) since it is a global metric that does not depend on an artificial cutoff point. In the artificial case that one individual suffers all the crime or all the events are concentrated on one street segment, or in the real case of the sexual offences in The Hague, the Gini coefficient gives a value close to one, indicating a high degree of concentration. However, there is still a major drawback of using the Gini coefficient directly from the number of crimes suffered by the population. The Gini coefficient is a valid metric with distinct values in the context where a variable, such as income, is distributed across most of the members of the population, but in the case of crime, the majority of the population suffers zero crimes and so the coefficient, computed directly from the number of crimes suffered by the population (or the number of crimes on each street segment), overestimates the level of concentration (Bernasco and Steenbeek, 2016). In the previous example of populations A and B, the Gini coefficient of the number of crimes is fairly similar given by $G_A = 0.95$ and $G_B = 0.97$ respectively, which reveals that crime is highly concentrated, but nothing more, providing little additional information to distinguish between the concentration observed in A and B.

When there are more individuals than crimes, as it is almost always the case, or more street segments than crimes, as it sometimes occurs, then an arithmetic adjustment to the Gini coefficient has been proposed (Bernasco and Steenbeek, 2016) which compares against a case of maximum equality (termed the generalised Gini coefficient denoted as G'). Although the generalised Gini coefficient is a clever way to correct the traditional Gini coefficient, it still has one major drawback. Consider two populations, B and C, where, as before, population B has a size of $N = 100,000$ individuals, and has 4,000 victims suffering a single crime each and 1,000 victims suffering 6 crimes each while the population C has only $N = 10,000$ individuals (that is only 10% of the size of B) where the population C has (just like population B) 4,000 victims suffering a single crime each and 1,000 victims suffering 6 crimes each. Populations B and C suffer crime under a different pattern and have a completely different concentration of crime since 95% of the population of B did not suffer any crime, whereas 50% of the C population suffered at least one crime. However, in this artificial but illustrative example, B and C have the same line of "maximal equality" and the same corrected Gini coefficient $G'_B = G'_C = 0.7$, precisely highlighting the main weakness of this arithmetic correction of the Gini coefficient: in the case of the population B, it *corrects* the traditional Gini coefficient by ignoring 90% of the population, but in the case of the population C it takes all of its individuals into account. In fact, any population with 4,000 victims suffering a single crime each and 1,000 victims suffering 6 crimes each, with a population size of $N \geq 10,000$ gives the same corrected Gini coefficient $G' = 0.7$ regardless of whether the population has only 10,000 inhabitants or millions. Thus, the arithmetic correction to the Gini coefficient creates another issue that the original Gini coefficient did not have (since $G_B = 0.97$ and $G_C = 0.7$).

Very recent developments, responding perhaps to issues raised by the *law of crime concentration* (Weisburd, 2015), have highlighted the need for more specific tools in the field of crime science. Unfortunately, current mea-

sures for the concentration of crime do not take into account the relative rarity of these events and so do not provide particularly useful information that would be necessary for policy design and decision-making. This results in misleading measures that underestimate crime concentration and which are not comparable over time, potentially leading to vulnerable groups being wrongly targeted while others are overlooked. Here, the Rare Event Concentration Coefficient (from equation 3.2, which in this thesis was used in the case of road accidents and volcanic eruptions), is used to measure of the concentration of crime suffered by a population which overcomes problems encountered when using other measures and descriptive statistics as metrics. This new measure can be used to compare different regions and it is also tractable across various time periods and therefore it can be used for purposes such as policy design, policing and crime prevention (Laycock and Farrell, 2003).

5.2 Victimisation and other concentrations of crime

5.2.1 Offending concentration

Measuring the number of crimes that a person suffers, quantifying its concentration and understanding the reasons why one person is victimised more frequently than others, has a counterpart in the number of crimes committed by potential offenders (Martinez et al., 2017). Relevant questions include, how many criminals are there in a region? And if crime increases, does it mean that there are more criminals, or that a few individuals have become more active? These questions have been of interest to criminologists for years (Nagin and Land, 1993) where the central question is how many crimes could be avoided by incarceration (Glueck and Glueck, 1950). For example, Wolfgang's classic study of a birth cohort in Philadelphia found that the majority of the population had no contact with the police, but at the same time there was a small group (less than 7% of the population) that was

responsible for the majority of the crimes committed by that cohort (Wolfgang et al., 1987). Also, by considering families and not only the individuals within a family, it was shown that less than 5% of the families account for more than 30% of the arrests (Farrington et al., 2001). Thus, crime is highly concentrated in the regions in which it is executed, in the population that suffers it and in the people that commit it.

5.2.2 Concentration of crime at places

The spatial concentration of crime is also a key element in the analysis of crime patterns, where the units of analysis might be households, street segments (Frith et al., 2017), areas (Lee et al., 2017), cities (Glaeser and Sacerdote, 1996; Oliveira et al., 2017) and maybe others (White et al., 2014).

In the specific case of a measure of spatial concentration of crime, a common technique is to consider the street segments in which the crime was committed (including, sometimes, the intersections) and then to determine the amount of crime concentrated in the top 5% of the segments (Andresen et al., 2016) or any other top $P\%$. A similar type of metric is often used when the number of crimes suffered by the most victimised people is reported (Pease and Ignatans, 2016; Pease, 1998), or the most criminal individuals (Wolfgang et al., 1987) or families (Farrington et al., 2001). This metric, however, has some severe issues, such as the lack of agreement on the percentage that gets reported (Fox and Tracy, 1988); the metric might not be comparable between different cities (Hipp and Kim, 2016); it might be the result of a certain degree of randomness (Levin et al., 2016) and it does not work as an adequate metric when the data is extremely sparse. Consider, for example, the number of street segments of The Hague and the number of sexual offences registered by the police between 2007 and 2009 in that city (Bernasco and Steenbeek, 2016). The extremely low frequency of this type of crime (only 430 cases) distributed over the large number of street segments (14,375 segments) means that taking the top 5% of streets is not even properly defined since, at most, 3% of the segments concen-

trate all the crimes. Taking the top 5% street segments, victims or criminals is a weak way of measuring crime concentration based on an artificial cut-off point which is blind, not only to the other 95% of the observations, but also, it is blind to the distribution inside the 5% being considered, which is extremely relevant when the events are rare, as crime usually is.

Other types of concentration also reveal patterns which help better understanding why crime occurs and how to prevent it and more, for example, the modus operandi, the type of crime or even aspects related to the police and its work. It was shown, for instance, that 1% of the Chicago Police officers were responsible for one-third of the lawsuits for a period of 6 years (Eck et al., 2017). From here onwards, the focus is placed on the victims only and the concentration of victimisation.

5.3 A probabilistic approach to the crime and victimisation rates

Assuming that the number of crimes suffered by the individuals within a population is independent and that suffering a crime does not affect the probability of a person being a victim in the future, then the number of crimes suffered by the i-th individual during a period of time (usually a year) might be modelled as a Poisson distribution with rate $\lambda_i \geq 0$. Although other distributions could be used for modelling the random component of suffering crime, such as a Negative Binomial (Park and Eck, 2013), the Poisson distribution allows to focus on a single parameter (the rate λ), and so it is frequently used in crime science (Maltz, 1996).

What is relevant about this approach for the distribution of the crime rates is that it is probabilistic: even when a person has a rate of $\lambda > 0$ of suffering a crime, the probability that he or she does not suffer any crime is not negligible, given by $\exp(-\lambda)$, therefore, even when a person did not suffer a crime, it does not necessarily mean that he or she has a crime rate of $\lambda = 0$, and vice versa, if a person suffers many crimes, it could be the

result of a small rate and bad luck. By focusing on the rates λ_i rather than on the observed frequency of crime, not only the population who actually suffered crime is being considered, but also the people who did not suffer any crime but were lucky, in the sense that given that their rate is greater than zero they were fortunate not to suffer any crime.

The assumptions required to model the crime counts as a Poisson distribution (independence of the crime suffered by individuals, independence between past and future victimisation and a constant rate) might be problematic, especially since past victimisation actually helps predict future victimisation (Tseloni and Pease, 2003), crime suffered by individuals might be strongly correlated, for example, if individuals live nearby and finally, certain types of crime are more likely to occur at specific times of the day. These assumptions are thus too strong and hence the results are not really useful to forecast the number of crimes that a person will suffer, for instance. However, the objective here is to construct a global metric of the concentration of crime and so these assumptions, although apparently unrealistic, help measure the concentration based on the least possible aggregated observations, usually from large populations but for which the number of crimes suffered is quite small.

Let λ_i be the *individual crime rate*, which represents the rate or "speed" at which the i-th individual suffers crime. The reasons why individuals experience different rates have been considered in depth by others and explanations go from individual attributes, which cause an increase in the attractiveness, to a boost on the probability of suffering a second crime after suffering a first one (Johnson et al., 2009). It is assumed that individuals might suffer a different crime rate λ_i and the distribution of the rates is referred to as an *inhomogeneous distribution* of the crime rates (the term "inhomogeneous" is used rather than "heterogeneous" for consistency purposes with other applications of a Poisson process).

The causal mechanism that leads to a population suffering an inhomo-

geneous rate has been studied before (Tseloni and Pease, 2004), but the focus of this study is the distribution itself, so here, different individual rates are assumed, without going any further into this topic. Considering a probabilistic approach means that the actual number of crimes suffered by the i-th individual is an observation from the Poisson distribution, so it is possible to analyse the individual crime rates rather than the observations. The expected number of crimes suffered by the population is simply the sum of the rates λ_i, from which

$$C = \sum_{i=1}^{N} \lambda_i, \tag{5.4}$$

which also means that the average rate $\bar{\lambda}$ is the population crime rate $c = C/N$.

The probability that the i-th person actually suffers a crime is $(1 - \exp(-\lambda_i))$, which means that the number of victims V follows a Poisson-Binomial distribution with parameters $p_i = 1 - \exp(-\lambda_i)$, with $i \in 1, 2, \ldots, N$. The Poisson-Binomial distribution is closely related to a Binomial distribution, in which each observation is allowed to have a different success probability (Chen and Liu, 1997). The expected number of victims is

$$V = \sum_{i=1}^{N} p_i = N - \sum_{i=1}^{N} \exp(-\lambda_i). \tag{5.5}$$

Since the number of crimes C and the number of victims V from the population is a fixed (observed) number, there are three restrictions for the rates λ_i:

Restriction I: $\lambda_i \geq 0$ for all $i = 1, 2, \ldots, N$.
Restriction II: $\sum_{i=1}^{N} \lambda_i = C$.
Restriction III: $N - \sum_{i=1}^{N} \exp(-\lambda_i) = V$.

The restrictions II and III refer to the expected value of the number of crimes C and victims V (so the left-hand side is not necessarily a whole number), which means that both are considered to be satisfied in an interval

around each integer C and V. Additional restrictions could also be considered by taking into account the observed number of people who suffered exactly two crimes, or exactly three crimes and so on against the theoretical (expected) outcome. Different distributions of the individual rates λ_i, with $i = 1, 2, \ldots, N$ which satisfy these restrictions could be the distribution of the crime rates over the whole population, and a goodness of fit test could help accept or reject the distribution of the individual rates λ_i.

A homogeneous distribution of the rates means that λ_i is the same for all $i = 1, 2, \ldots, N$, from which, due to the Restriction II, the value of $\lambda_i = c$. If (and only if) c and v are such that $(1 - \exp(-c)) = v$, then a homogeneous distribution of the individual rates might be accepted. However, this is rarely the case since crime is far more concentrated than random events would predict (Osborn and Tseloni, 1998); usually v is much smaller than $(1 - \exp(-c))$, meaning that a homogeneous distribution is far from being the observed one and other distribution of the rates λ_i needs to be considered.

Although the methodology presented here, by modelling the distribution of the victimisation rates, is designed for the number of crimes suffered by individuals, with certain precautions it could also be applied to other aspects of crime, for instance, the concentration of criminality (the observations could also be the number of crimes executed by a person) or the spatial concentration of crime (so the observations could be the number of crimes committed on a street segment). In the case of the crime committed by the population, the assumption that past and future events are independent is strong, since it has been found that the rate at which an individual executes crime tends to increase as they commit more crimes (Ferguson, 1952) meaning that a constant rate is dubious. In the case of the spatial distribution, crime suffered in street segments (Weisburd, 2015) or regions of a city (Mohler et al., 2012) tends to be highly concentrated and a hot spot pattern usually emerges, indicating a geographic clustering of crime (Short et al., 2008), thus, assuming independence between the observations might

not be adequate. The methodology presented could be used for measuring the concentration of crime in the other aspects, such as the criminals and the places of crime.

5.3.1 Inhomogeneous distribution of crime rates

The individual crime rates λ_i depend on many factors (Tseloni, 2000), such as the habits of the i-th person, their lifestyle, the region in which he or she usually commutes, physical attributes (such as gender or age), and perhaps that rate is of a similar value to other individuals who live under the same circumstances. To model the inhomogeneous distribution of the crime rates, it is assumed that the population can be divided into $k \geq 1$ distinct groups, where group j say, has Q_j members, all of whom suffer the same crime rate λ_j, with $j = 1, 2, \ldots, k$. Each of the N members of the whole population belongs to one and only one group so that $Q_1 + Q_2 + \cdots + Q_k = N$. The proportion of the population who suffer the crime rate λ_j is $q_j = Q_j/N$, so that $\sum_{j=1}^{k} q_j = 1$. To avoid ambiguous definitions, groups are ordered by their crime rate in increasing order, $0 \leq \lambda_1 < \lambda_2 < \cdots < \lambda_k$, so that groups are labelled according to their rate.

The distribution of the number of crimes that a random person from the population suffered can be expressed as

$$q_1 \text{Pois}(\lambda_1) + q_2 \text{Pois}(\lambda_2) + \cdots + q_k \text{Pois}(\lambda_k), \tag{5.6}$$

which means that the person is assigned into any of the k groups and suffers a Poisson distribution with the corresponding rate.

Assuming that the population can be divided into $k \geq 1$ distinct groups where all of the members of each group suffer the same rate is a common technique that simplifies the crime suffered by a population of perhaps millions of people into only a few parameters (Short et al., 2009; Brame et al., 2006; Nagin and Land, 1993).

The number of groups, k, is crucial for the mixture model (Böhning et al.,

1992), and distributions with a larger number of groups are less useful since for each additional group, its size and its rate need to be estimated, so this increases the number of parameters of the model. The (non-parametric) maximum likelihood estimator (*mle* or *npmle*) helps comparing between models and to pick the best (Böhning et al., 1998), since in the case of suffered crime, no prior information on the number of groups is considered (McLachlan and Peel, 2004). Although other techniques to estimate the number of groups are also available, for example, by using bootstrapping (Schlattmann, 2005), the *mle* is used here, which includes an estimate $\hat{k}$ of the number of population groups, easily computed using the statistical package CAMAN (Computer Assisted Analysis of Mixtures) by considering the observed number of crimes suffered by each of the individuals, C_i. A similar procedure, using a mixture model, has been used in different scenarios (Böhning, 1998), such as road accidents, the number of accidents in a factory and even the number of criminal acts from a set of persons considered to have deviant behaviour.

The results obtained using CAMAN and estimating the *mle* are:

- the optimal number of groups in which the population is divided $\hat{k}$,
- the size of each population group relative to the total population $\hat{q}_j$, expressed as a vector as $\underline{q}$, and
- the corresponding rate for each group $\hat{\lambda}_j$, also as a vector, $\underline{\lambda}$.

As an example, consider again the previous populations A and B, both with $N = 100,000$, $c = 0.1$ and $v = 0.05$. In population A (where 10,000 crimes are suffered uniformly by the 5,000 victims), the *mle* gives $\hat{k} = 2$ groups, with $\underline{q} = (0.937,\ 0.063)$ and $\underline{\lambda} = (0,\ 1.594)$, which means that 93.7% of the population has a crime rate of $\hat{\lambda}_1 = 0$ and 6.3% of the population has a crime rate of $\hat{\lambda}_2 = 1.594$. On the other hand, in the population B (where 6,000 crimes are suffered by 1,000 victims and 4,000 crimes are suffered by 4,000 victims), the *mle* gives $\hat{k} = 3$ groups, with

$\underline{q} = (0.409,\ 0.580,\ 0.010)$ and $\underline{\lambda} = (0,\ 0.069,\ 5.897)$, which means that the mixture model informs that indeed 1% of the population suffers crime with rate $\hat{\lambda}_3 = 5.897$, but also that 58% of the population suffers crime at a rate of $\hat{\lambda}_2 = 0.069$, so that if 15 individuals were randomly selected from that group, only one victim is expected. Results indicate that a large proportion of the population (more than half) suffers crime at a very small rate, but there is a particular group, (formed by only 1% of the individuals) whose members suffer a considerably large crime rate ($\hat{\lambda}_3 = 5.89$). In the population B, efforts might be much better oriented towards the small population group who expect to suffer $\hat{\lambda}_3 = 5.89$ crimes during that time period rather than the large population group who suffer the lower rate $\hat{\lambda}_2 = 0.069$.

The mixture model process depends not only on c and v, but it also changes based on the number of people who suffered exactly 0, 1, 2, 3 or more crimes. It gives a more accurate distribution of the crime rates, considering the random component of suffering a crime, hence it provides a better understanding of the distribution of crime in the population since it estimates that a group of relative size $\hat{q}_j$ suffers crime at a rate $\hat{\lambda}_j$. This distribution is useful since it allows us firstly, to compare different regions or different time periods, secondly, to simulate crimes in a population, and thirdly, to determine the expected departures that natural variability gives to the number of crimes suffered by the population. The mixture model is useful from a macro perspective giving an approximate distribution of the number of crimes over the whole population in terms of only a few parameters. It is not useful, though, from an individual perspective. For example, results for population B are that nearly 60% suffers a crime rate $\lambda > 0$ but it comes from a population where 95% suffered zero crimes, meaning that if a person suffered no crimes, it would not be possible to tell whether they belong to the group that suffers no crime or whether they belong to a group which suffers a small rate $\hat{\lambda}_2 = 0.069$ or even less likely, but not impossible, they belong to the group who suffers a large rate $\hat{\lambda}_3 = 5.89$ and yet they were lucky and suffered no

crime.

Two particular cases of the results of the mixture model are worth further comment. The first is when $\hat{k} = 1$ (which means that there is only one group, with rate $\hat{\lambda}_1 = c$), in which case the mixture model tells that crime is uniformly suffered by the population. The second case is when $\hat{k} = 2$ and $\hat{\lambda}_1 = 0$, which indicates that the population is divided into two groups, one of them of relative size $\hat{q}_1$ which does not suffer crime and the second group, of relative size $\hat{q}_2 = 1 - \hat{q}_1$, suffers all the crime within that population. This type of model is also known as a Zero-Inflated Poisson Model (Böhning, 1998), frequently used to model heterogeneity in the rates (Bushway and Tahamont, 2016) and for count data in which the number of zeros is frequent, such as here in which the count of the number of people who suffered zero crimes is considered. These two cases might result from data after fitting the mixture model, which means that the mixture model lets the data adjust to the most suitable distribution of the crime rates, without assuming anything about the uniformity of crime suffered by the whole population.

5.3.2 Immunity and chronic victimisation

As previously noted (Sparks, 1981; Hope and Norris, 2013), there is usually a population group which is immune to victimisation and the mixture model detects the existence of such a population group and determine its size. After analysing the data, if the results show that there are $\hat{k} \geq 2$ groups and $\hat{\lambda}_1 = 0$, then it means that indeed there is a group who is immune to crime and its relative size is given by $\hat{q}_1$. It is important to note that the existence of an immune group is the result of the data and the model rather than by an assumption. Equally, results might reveal that there is no immune group, in which case the smallest crime rate would be $\hat{\lambda}_1 > 0$.

Population groups that suffer chronic victimisation have also been noted previously (Hope and Trickett, 2008), where again, their existence might be tested using the results from the mixture model. Results from the mixture model might also show that a population group which suffers a rate higher

than $\lambda = 2$. Suffering two or more crimes per year is a persistent and perhaps habitual victimisation and so groups which suffer a rate $\hat{\lambda}_k \geq 2$ are considered to suffer *chronic victimisation*.

In the previous example of populations A and B (with $N = 100,000$, $c = 0.1$ and $v = 0.05$), for the population A (where 10,000 crimes are suffered uniformly by the 5,000 victims), results of the mixture model showed that 93.7% of the population has a crime rate of $\hat{\lambda}_1 = 0$, which means that a large portion of the population is statistically immune to crime. For the population B (where 6,000 crimes are suffered by 1,000 victims and 4,000 crimes are suffered by 4,000 victims), results of the mixture model showed that 40.9% are immune to crime, but also, 1% of the population has a crime rate of $\hat{\lambda}_3 = 5.89$, which means that they expect to suffer chronic victimisation of almost six crimes each year.

5.3.3 Concentration of crime metric

Although the distribution of the rates $(\underline{q}, \underline{\lambda})$ is powerful by itself, the *Rare Event Concentration Coefficient* $RECC$ (Prieto Curiel and Bishop, 2016a) is a new and standardised summary statistic from the mixture model, based on the Lorenz curve (Marsh and Elliott, 2008; Hope and Norris, 2013) and the Gini coefficient (Dorfman, 1979) of the distribution of the crime rates. It is important to note that it is not the Gini coefficient computed directly from the number of crimes suffered by each member of the population, as it has been previously used to measure the concentration of crime (Tseloni and Pease, 2005; Fox and Tracy, 1988; Bernasco and Steenbeek, 2016), but rather it is the Gini coefficient of the rate at which individuals suffer crime. The $RECC$ is given by

$$RECC = \frac{1}{2\sum_{i=1}^{\hat{k}} \hat{\lambda}_i \hat{q}_i} \sum_{i=1}^{\hat{k}} \sum_{j=1}^{\hat{k}} \hat{q}_i \hat{q}_j |\hat{\lambda}_i - \hat{\lambda}_j|, \tag{5.7}$$

which is the Gini coefficient of a stepwise distribution and can be interpreted in a similar way to how the Gini coefficient is used in the case of the distri-

bution of wealth: a smaller value of the $RECC$ means a more homogeneous distribution of the crime rates across the population, and a value closer to one means that crime is more concentrated in some population groups. The $RECC$ is comparable between different time periods and different regions and different types of crime.

Using again the example of populations A and B (both with $N = 100,000$, $c = 0.1$ and $v = 0.05$), for the population A (where 10,000 crimes are suffered uniformly by the 5,000 victims) the mixture model says that 93.72% of the population is considered immune to crime and so, the $RECC_A = 0.9372$. On the other hand, for the population B (where 6,000 crimes are suffered by 1,000 victims and 4,000 crimes are suffered by 4,000 victims) the $RECC_B = 0.7546$ which means that crime is suffered more homogeneously in the population B, perhaps an expected result since 59% of the population has a crime rate greater than zero.

5.3.4 Coefficient Interval

Is observing $RECC_A = 0.9372$ statistically different from $RECC_B = 0.7546$? An interval for the $RECC$ is constructed for each population based on a Monte Carlo method (Mooney, 1997) which incorporates a level of uncertainty. This method assumes that the distribution $(\underline{q}, \underline{\lambda})$ is the *true* distribution of the crime rates and so it is possible to simulate N individuals which suffer crime with the distribution given in equation 5.6. Each one of the N simulated individuals represents the number of crimes that a person taken at random from the population might suffer, given the true distribution of crime and by simulating N individuals, departures from the true distribution that the number of crimes the population could experience are considered. By computing the mixture model of the simulated crimes and then considering its corresponding $RECC_{sim}$ how low, or high, the $RECC$ could be is computed, given the exact same distribution.

Following the same procedure, a sufficient number of times (100 in this case) results in a simulated sample of potential values of the $RECC$. Sub-

sequently, the 95% interval is considered, to avoid the extreme simulated values (Greenland, 2004). The results, in terms of the simulated $RECC$, are given in Figure 5.1 in terms of the 95% lower and upper bound intervals.

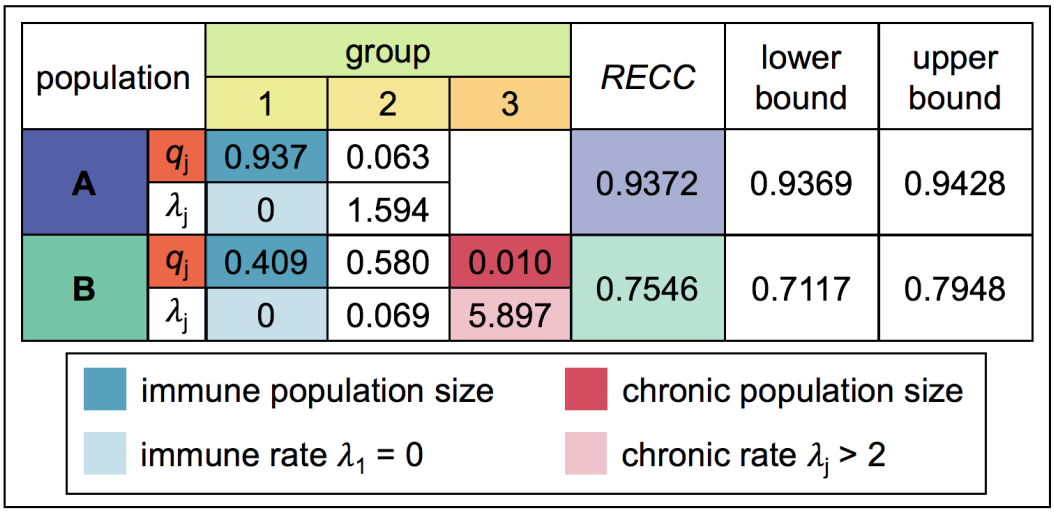

Figure 5.1: Group sizes $\underline{q}$, crime rate $\underline{\lambda}$ and intervals for the $RECC$ for the populations A and B.

These results show that with the true distribution observed in the population A, the $RECC$ does not achieve values as low as the ones obtained for population B. Therefore, with the simulated intervals, the null hypothesis that both populations have the same concentration of crime is rejected and so, thanks to the simulated intervals, a statistical justification that both populations suffer a different concentration of crime is obtained.

A relevant observation from the simulations is that the number of groups $\hat{k}_{sim}$ might change and also the sizes of the immune group and the chronic group might also change since, for example, suffering a small rate of $\hat{\lambda} = 0.01$ is almost the same as $\hat{\lambda} = 0$ but this difference would change the size of the immune group. Therefore, for comparing two different populations or comparing the same population over different time periods, a global metric, such as the $RECC$, provides more stable results.

Finally, it is important to note that after observing the X_i and executing the mixture model, not much can be said about the individuals (since what is obtained is a global behaviour). For population A, for example, there are

$N = 100,000$ individuals and 10,000 crimes suffered uniformly by the 5,000 victims; the mixture model gives $\underline{q} = (0.937, 0.063)$ and $\underline{\lambda} = (0, 1.594)$, so that, in this simple, illustrative model, 5% of the population actually suffers crime but then 6.3% of the population ($q_{2,A}$) belongs to a group who suffers crime at a positive rate ($\lambda_{2,A} = 1.594$). This group includes the 5% of the population who did suffer a crime and an additional 1.3% of the population who were the "lucky" ones since they have a rate $\lambda_{2,A} = 1.594$ but suffered no crimes. There is no way to distinguish the 1.3% of the population who suffered no crimes but have a rate of $\lambda_{2,A} = 1.594$. Results should be considered in a more global manner, in which the population might be best described as two groups, one immune to crime and a small group who suffers a rate of $\lambda_{2,A} = 1.594$.

5.3.5 Assumptions of unit and event independence

Assuming both, unit independence and event independence are strong assumptions and need to be carefully considered.

For unit independence, there is indeed a correlation between the number of crimes suffered by people who live in the same region or belong to similar social networks. However, there are two issues. First, the extremely low frequency of these events, so that most individuals experienced zero crimes during the period of a year. Secondly, and perhaps more importantly, the available information comes from a survey, so that those individuals who might have a correlation between their experience of crime may not necessarily have been surveyed and it is not possible to identify these people from the survey. Therefore, unit independence, when considering individuals, is the best possible option.

For event independence, perhaps for some particular individuals and for some selected types of crime, instead of having a constant yearly crime rate λ_i, which can be further divided into weekly or monthly crime rates $\lambda_i/12$ or $\lambda_i/52$, it is possible to consider the crime rate as a function of time $\lambda_i(t)$ or as a function of past victimisation and model the number of crimes as a

self-exciting point process (Mohler et al., 2012). However, due to the low frequency, rates are almost always so close to zero, that any refinement of a constant rate model does not significantly change the original result.

5.4 Case studies

5.4.1 Burglaries in Netherlands

To demonstrate the use of the victimisation profile and of the $RECC$, the number of burglaries suffered per household is considered, obtained from a victimisation survey called the 1993 Police Monitor in the Netherlands. The data has been used before as a test bed for analysis and to explore the level of concentration of burglary in that country and it contains a discussion on how data was obtained (Tseloni et al., 2004). The number of houses that suffered 0, 1, 2 or more crimes is displayed in Table 5.1.

burglaries	0	1	2	3	4	5	6+	Total
houses	36,632	2,548	448	134	51	19	17	39,849
frequency (%)	91.9	6.4	1.1	0.3	0.1	0	0	100

Table 5.1: Observed number of burglaries in Netherlands, 1993

Burglaries are highly concentrated. From the population surveyed, 91.9% suffered no burglaries, but then from the houses that in fact suffered a burglary, 21% suffered more than one. The mixture model applied to the number of burglaries suffered in the Netherlands gives as results $\hat{k} = 5$, with sizes and rates displayed in Table 5.2.

Group	1	2	3	4	5
Crime rate $\hat{\lambda}_i$	0	0.11	0.30	1.79	7.93
Relative size $\hat{q}_i$ (%)	58.2	24.3	15.8	1.7	0
Number of houses	23,188	9,679	6,282	688	12

Table 5.2: Results of the mixture model applied to the burglaries suffered in the Netherlands. $RECC = 0.7643$.

The results of the mixture model applied to this data show that the population can be divided into five groups, the largest one (58.2% of the pop-

ulation) suffers no crime ($\hat{\lambda}_1 = 0$), the second largest one (24.3%) suffers a rate of $\hat{\lambda}_2 = 0.11$, so that among any 9 houses of that group they expect to have experienced a burglary, and so on. There is a group (less than 3.3 out of 10,000 houses) which experiences a rate of almost 8 crimes per year ($\hat{\lambda}_5 = 7.93$).

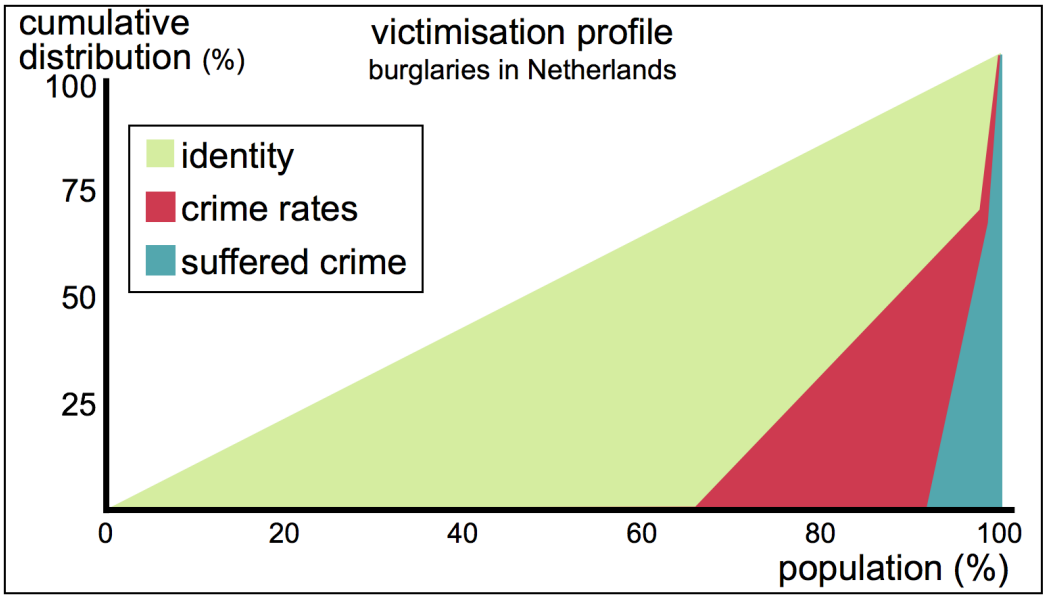

Figure 5.2: Lorenz curve of the individual crime rates and the number of crimes for the Netherlands in 1993.

The Rare Event Concentration Coefficient gives $RECC = 0.7643$ and the Lorenz curve of the observed number of crimes and the estimated rate is displayed in Figure 5.2 for comparison purposes. It is relevant to notice that the distribution of the rates $\underline{\hat{\lambda}}$ is much more evenly distributed than the actual crime and this is usually the case since, for example, the mixture model says that 24.3% of the population (Group 2) suffers a rate of $\hat{\lambda}_2 = 0.11$, so that within the population from that group, only one victim from every 9 houses is expected. In that group, the observed burglaries are highly uneven (for each house victimised there are eight houses not victimised), but the rates are uniformly distributed. The main element here is that the events considered are rare, so most of the observations (nearly 92% in the example from the Netherlands) are zero, but it does not mean that their rate is zero.

If for some reason, the number of households which suffer 6 or more burglaries drops to zero (a change in only 17 out of the 39,849 observations) the $RECC$ would be 0.7092, which means a difference of 0.0551. With the traditional approach to the concentration of a variable, if we compute the Gini coefficient directly to the number of crimes that each household suffered its value is 0.9362 and again, if for some reason the number of households which suffer 6 or more burglaries drops to zero, the Gini coefficient of the number of crimes suffered would be 0.9351, which only means a difference of 0.0011; an almost negligible change. Naturally, if the number of households that suffer 6 or more crimes drops to zero, the change in terms of the number of crimes might not be significant, but it is relevant in terms of the structural change in the way that crime is suffered, and it is a change that would not be detected by the traditional Gini coefficient. However, the $RECC$ allow detecting such change, even in the case when it occurred for such a small population group.

5.4.2 Robberies in Mexico

The case of Mexico is used to apply the mixture model and compute the $RECC$ to detect differences in crime concentration. Its territory is divided into 32 states with a wide variety in terms of population size —some states have a population of just above 700,000 inhabitants (a population size similar to Luxembourg), while at the same time, there is a state with a population size nearly 23 times larger, of more than 16 million inhabitants (a population size similar to the Netherlands)— and this sub-division also considers Mexico City as a separate state.

Data was obtained from an annual victimisation survey conducted yearly between 2011 and 2016 in Mexico. Thus, six years of data are available for measuring the concentration of crime across time. Although previous victimisation surveys were also conducted in Mexico, the survey used here provides comparable data over different years and also provides

an up-to-date data (from INEGI, 2016), with micro-data available on-line[1]. For each year, more than 80,000 surveys were conducted, and its sampling method allows separate data for each of the 32 states. The survey contains an expansion factor, used to establish an estimate for the number of people who are represented by each survey respondent so that every person older than 18 years in the country is represented by a single survey respondent.

For different types of crime, such as robbery of a person, car theft or burglary, the person is asked whether he or she suffered that type of crime and the number of times that it occurred during the previous year. The case of robbery of a person is used since it has the highest variability, from a crime rate as low as $c_{BCS} = 0.007$ (where the subscript denotes the state) to a crime rate as high as $c_{MEX} = 0.471$, nearly 68 times larger, so this particular type of crime allows detecting whether higher crime rates are also associated with a higher concentration of crime. Also, the analysis of the crime rates from the 32 states in Mexico allows comparing population groups, so detecting, for example, an immune group in the states with high crime rates implies that they live under better conditions, in terms of crime and security than some of the groups from states with low crime rates. Thus, living in a state with a lower crime rate is not necessarily preferable from an individualistic viewpoint.

5.4.3 Concentration of crime in Mexico

At a national level, the Table 5.3 gives the number of crimes suffered by the survey respondents in Mexico for 2016, as well as the national estimate, considering the expansion factor for each survey. The data shows that 91.9% of the population did not suffer a robbery of a person during 2015, but also, it is estimated that more than 100,000 persons suffered at least four robberies during that year.

Is the crime suffered the result of a homogeneous distribution? One state is used (Guerrero) and over one year (2016) to test against a random

[1] Available at `http://www.inegi.org.mx/`

	Number of crimes suffered						
	0	1	2	3	4	5	6 +
respondents	81,672	3,486	455	95	22	10	24
population*	75,992.6	5,604.5	769.5	193.1	72.7	23.1	9.8
%	91.9	6.8	0.9	0.2	0.1	0.0	0.0

Table 5.3: * - crimes suffered by the population, in thousands, estimated by considering the expansion factor for each survey respondent.

distribution of crime (Park and Eck, 2013). With the observed number of crime rate $c = 0.069$ and from equation 5.5, a victimisation rate of $v = 0.066$ would be expected under the null hypothesis and values between 0.065 and 0.068 would support this hypothesis. However, the observed victimisation rate ($v = 0.053$) is far from this interval so that, in this state, crime is far from being homogeneously suffered by the population and there are much fewer victims than the homogeneous distribution would indicate. Similar results occur for other states and so there is, indeed, a high concentration of crime.

The distribution of the rates for each state and for each year have been computed (R Core Team, 2014; Schlattmann et al., 2015) based on data from the victimisation surveys and the concentration $RECC$. The victimisation profile for the latest two years (2015 and 2016) is available in the Appendix and the victimisation profile for 2014 is in Figure 5.3. The results show that crime has a completely different pattern across the 32 states from Mexico. For example, in some states (Baja California Sur in 2015 or Aguascalientes in 2016), the population can be divided into just two groups, the immune and the victimised. However, in Morelos, for example, for 2015 and for 2016 the model gives 4 groups, which means that a more complex distribution of crime is needed.

states in Mexico 2014		group													
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
Aguascalientes	q_j	0.87	0.11	0.02											
	λ_j	0.00	0.15	0.45											
Baja California	q_j	0.79	0.21	0.00											
	λ_j	0.00	0.23	1.58											
Baja California Sur	q_j	0.80	0.20	0.00											
	λ_j	0.00	0.07	3.24											
Campeche	q_j	0.78	0.22	0.00											
	λ_j	0.00	0.15	1.93											
Coahuila	q_j	0.86	0.02	0.02	0.02	0.02	0.02	0.02	0.02						
	λ_j	0.00	0.12	0.17	0.22	0.26	0.32	0.43	0.73						
Colima	q_j	0.92	0.02	0.03	0.02	0.00									
	λ_j	0.00	0.17	0.20	0.22	2.44									
Chiapas	q_j	0.81	0.04	0.04	0.06	0.03	0.02	0.00							
	λ_j	0.00	0.14	0.19	0.21	0.23	0.26	1.42							
Chihuahua	q_j	0.78	0.02	0.03	0.03	0.03	0.03	0.03	0.02	0.02	0.01				
	λ_j	0.00	0.07	0.13	0.19	0.24	0.30	0.36	0.42	0.48	0.53				
Ciudad de Mexico	q_j	0.60	0.02	0.05	0.09	0.04	0.02	0.09	0.07	0.02	0.01				
	λ_j	0.00	0.09	0.17	0.24	0.32	0.40	0.48	0.57	0.66	0.74				
Durango	q_j	0.62	0.38												
	λ_j	0.00	0.10												
Guanajuato	q_j	0.75	0.02	0.03	0.05	0.02	0.04	0.05	0.02	0.02					
	λ_j	0.00	0.07	0.13	0.20	0.25	0.31	0.36	0.40	0.44					
Guerrero	q_j	0.82	0.02	0.03	0.03	0.03	0.03	0.02	0.02						
	λ_j	0.00	0.15	0.23	0.29	0.34	0.39	0.47	1.11						
Hidalgo	q_j	0.79	0.18	0.02	0.02										
	λ_j	0.00	0.14	0.79	0.93										
Jalisco	q_j	0.78	0.02	0.02	0.02	0.02	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.00
	λ_j	0.00	0.05	0.11	0.18	0.25	0.32	0.39	0.46	0.53	0.58	0.63	0.67	0.71	0.78
Estado de Mexico	q_j	0.50	0.15	0.02	0.06	0.04	0.08	0.13	0.03	0.00					
	λ_j	0.00	0.21	0.61	0.88	1.06	1.21	1.38	1.64	7.60					
Michoacan	q_j	0.82	0.05	0.04	0.08	0.02	0.00								
	λ_j	0.00	0.13	0.22	0.27	0.31	3.81								
Morelos	q_j	0.75	0.05	0.08	0.05	0.02	0.03	0.01	0.01	0.00	0.00				
	λ_j	0.00	0.20	0.26	0.34	0.53	1.02	1.27	1.33	1.36	6.67				
Nayarit	q_j	0.87	0.13	0.01											
	λ_j	0.00	0.13	0.78											
Nuevo Leon	q_j	0.79	0.02	0.02	0.03	0.02	0.02	0.03	0.02	0.02	0.02	0.00			
	λ_j	0.00	0.08	0.17	0.25	0.34	0.42	0.49	0.57	0.64	0.71	0.77			
Oaxaca	q_j	0.82	0.17	0.00											
	λ_j	0.00	0.28	2.29											
Puebla	q_j	0.72	0.03	0.07	0.05	0.06	0.04	0.02	0.01	0.01					
	λ_j	0.00	0.15	0.18	0.20	0.23	0.26	0.32	0.78	1.26					
Queretaro	q_j	0.66	0.35												
	λ_j	0.00	0.09												
Quintana Roo	q_j	0.80	0.07	0.08	0.05	0.00									
	λ_j	0.00	0.15	0.24	0.33	3.92									
San Luis Potosi	q_j	0.87	0.02	0.02	0.01	0.02	0.02	0.02	0.02	0.00	0.00				
	λ_j	0.00	0.08	0.23	0.42	0.58	0.69	0.77	0.83	0.90	4.81				
Sinaloa	q_j	0.83	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.00				
	λ_j	0.00	0.09	0.17	0.25	0.32	0.40	0.48	0.56	0.65	0.75				
Sonora	q_j	0.79	0.03	0.05	0.02	0.11									
	λ_j	0.00	0.07	0.11	0.13	0.15									
Tabasco	q_j	0.76	0.03	0.05	0.06	0.04	0.03	0.03	0.01						
	λ_j	0.00	0.09	0.17	0.24	0.31	0.39	0.50	0.64						
Tamaulipas	q_j	0.85	0.15	0.00											
	λ_j	0.00	0.19	4.20											
Tlaxcala	q_j	0.83	0.02	0.03	0.04	0.02	0.02	0.02	0.01						
	λ_j	0.00	0.15	0.17	0.18	0.20	0.22	0.25	1.26						
Veracruz	q_j	0.78	0.03	0.05	0.04	0.04	0.04	0.01							
	λ_j	0.00	0.08	0.13	0.18	0.23	0.28	0.34							
Yucatan	q_j	0.90	0.02	0.02	0.02	0.02	0.02	0.00							
	λ_j	0.00	0.04	0.10	0.17	0.28	0.40	0.52							
Zacatecas	q_j	0.78	0.21	0.00											
	λ_j	0.00	0.06	1.88											

immune population size
immune rate $\lambda_1 = 0$
chronic population size
chronic rate $\lambda_j > 2$

Figure 5.3: Victimisation profile for the 32 states in Mexico in 2014.

The outcome of the model is that there is usually a large group which have a rate $\hat{\lambda}_1 = 0$, forming the group which statistically is immune to crime and its relative size reaches a value as high as 95.6% in the state of Campeche in the year 2015. This means that during that year, in that state, less than 5% of the population actually expected to suffer crime, but with a rate of $\hat{\lambda}_2 = 0.853$, higher than most of the groups from all the other states. Actually, considering the population size of each state, this small group from Campeche suffered a higher rate than 96.8% of the whole country. The results obtained support the theory of the existence of an immune population group (Hope and Trickett, 2008) and its size on average through the six years considered, was 61.5% of the whole population.

Results also show that there are some states (15 out of the 32 states in 2015 and only 9 in 2016) with a group which suffers chronic victimisation, so they expect to suffer two or more crimes in a year. For example, in Estado de México in 2015, there is a small group (which represents approximately only 0.2% of the population), but which has a crime rate of $\hat{\lambda}_4 = 7.8$, which sadly means that they expect to suffer one robbery roughly every seven weeks.

Figure 5.4 shows the results from Mexico City during 2014 simply as an illustration of the crime rates suffered by the population. The upper diagram gives the rates from the mixture model so it provides the *victimisation profile* of Mexico City during 2014. The lower diagram, gives the cumulative rates and the Lorenz curve, where the area shaded in a bright colour is the distance to a uniform distribution of the rates.

5.4.4 Crime rates and crime concentration

An observation of the results is that lower (or higher) crime rates do not necessarily mean a lower (or higher) concentration of crime. For example, the state of Chiapas (CHIS) between 2015 and 2016 suffers a similar crime rate ($c_{CHIS,2015} = 0.037$ and $c_{CHIS,2016} = 0.034$ respectively) with an opposite pattern for the concentration of crime ($RECC_{CHIS,2015} = 0.897$ and $RECC_{CHIS,2016} = 0.228$).

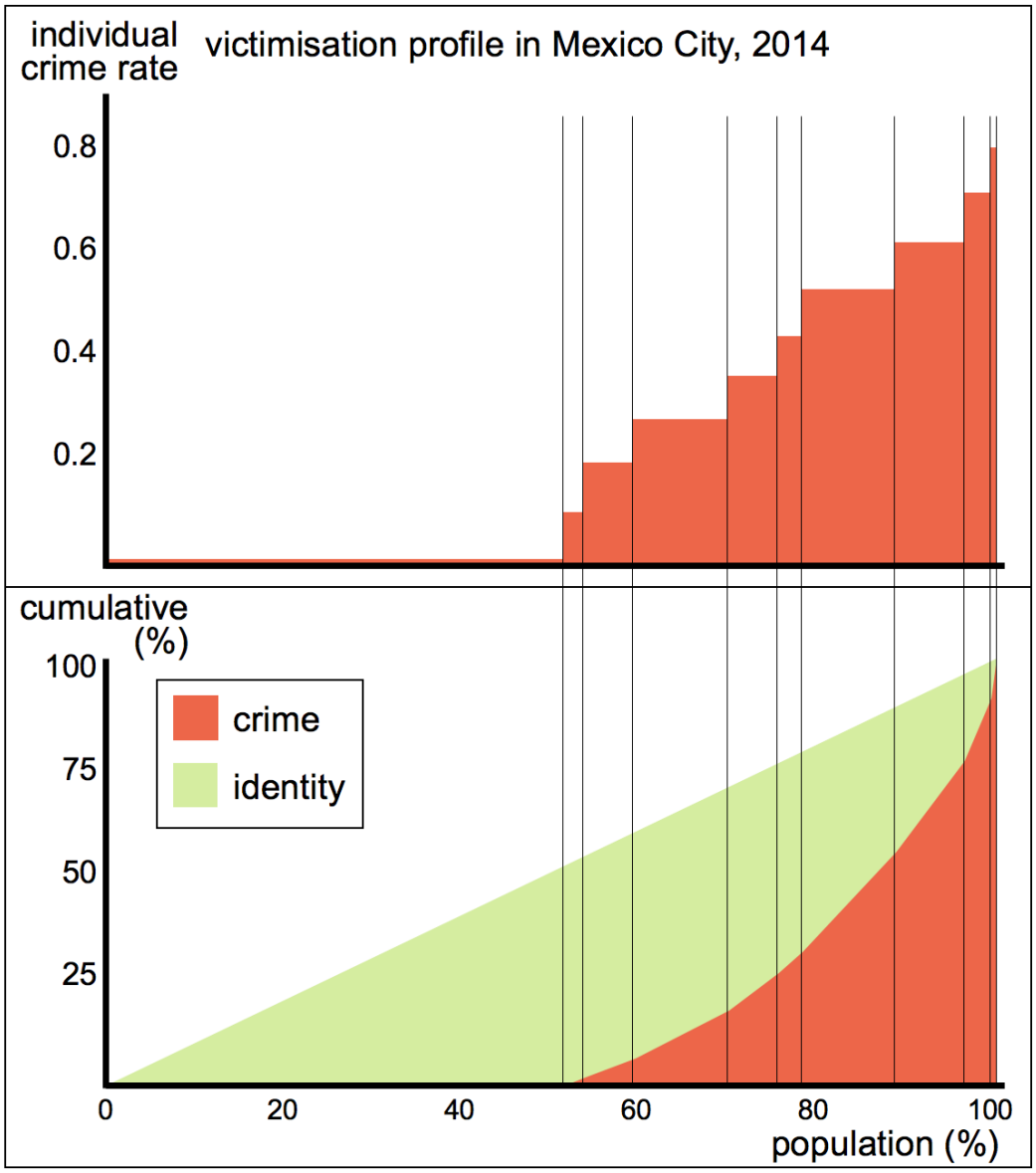

Figure 5.4: The victimisation profile of Mexico City in 2014: individual crime rates $\underline{\lambda}$ and group sizes $\underline{q}$ (above) and the Lorenz curve of the individual crime rates (below).

In general, there is a national decrease in the concentration of crime between 2011 and 2016. Between 2011 and 2013 the average $RECC$, weighted by the population size of each state, was nearly 0.8 but it has gradually decreased to an average of $RECC = 0.688$ by 2015 and $RECC = 0.563$ by 2016. At the same time, crime rates have increased slightly through this period, from a national rate near $c = 0.084$ to an average of $c = 0.105$.

Thus, for this particular type of crime, there are now more robberies which are being suffered by more people.

5.4.5 $RECC$ intervals and their interpretation

The $RECC$ and its corresponding intervals were computed for each state and for every year between 2011 and 2016. In addition, the victimisation profile gives a systematic way of simulating crime in a population and so, the particular results obtained in the 100 simulations for the Estado de México in the year 2014 is analysed. The specific case, by having the highest crime rates, also offers the highest possible variability of the estimated metrics. Results are the following:

1. k_{sim}: the *true* number of groups is $\hat{k} = 8$ and the simulations gave on average 7.87, with the extreme cases having either 7 or 9 groups, so, the number of groups does not change significantly.

2. $q_{immune,sim}$: the *true* size of the immune population is $\hat{q}_{immune} = 0.4953$ and the simulations gave on average $\bar{q}_{immune,sim} = 0.4569$, with an interval of $q_{immune,sim} \in [0.4296, 0.4825]$.

3. $q_{chronic,sim}$: the *true* size of the chronic population is $\hat{q}_{chronic} = 0.00233$ and the simulations gave on average $\bar{q}_{chronic,sim} = 0.00654$, with $q_{chronic,sim} \in [0.00168, 0.03247]$, so, the size of the chronic population changes.

4. $RECC_{sim}$: the *true* $RECC$ is $\hat{RECC} = 0.66012$ and the simulations gave on average $\bar{RECC}_{sim} = 0.6580$, with $RECC_{sim} \in [0.6542, 0.6628]$, so, the $RECC$ actually provides a very stable measurement for the concentration of crime. It does not change drastically from one simulation to the next, even when the number of groups changed, the size of the immune or the rate of the chronic group changed, the $RECC$ is a much more stable measure.

As a result of the Monte Carlo procedure, there are three relevant results. Firstly, the number of groups k might change, but it does not significantly change the profile of the crime rates since even with more or fewer groups, the simulated distribution still behaves like the true distribution. Secondly, the size of the immune group or the chronic group might vary a bit since observing a group with a rate equal to zero and a group with a rate equal to $\lambda_i = 0.01$, for example, might not be that different, but this does change the size of the immune group. Finally, the $RECC$ gives a stable and robust measurement of the concentration of crime rates as its variation is relatively small. The $RECC$ takes into consideration the behaviour of all the groups and so even when the result of one of the simulations gives fewer groups, a larger chronic group or a smaller immune group, the $RECC$ is a global measure and so it provides a stable result.

From the 100 simulations carried out for the Estado de México in 2014, a (simulated) victimisation profile, could also be obtained, which are based on the *true* distribution and so, for a given percentile, departures from the rates are obtained. For instance, in Figure 5.5 (upper panel) the 100 simulated profiles are displayed simultaneously and, for example, the rate for the 30th percentile is always zero, the rate for the 60th percentile goes between 0.3 and 0.42 and so on, for different percentiles. It is possible also to obtain the mean victimisation profile, displayed in 5.5 (lower panel), where now the interface between groups is not step-like, but closer to a smooth curve. This shows that there are indeed different groups (the immune ones, which in all the simulations are nearly half of the population, the ones who suffer a relatively small amount, the group with chronic victimisation and so on).

Unfortunately, although this method of taking the average of the 100 simulated profiles gives a much more “stable” victimisation profile, it has two major drawbacks and therefore it is not considered the victimisation profile of the population. Firstly, the results are no longer a function with only a few parameters and, for instance, for the year 2016, with only 5 free parameters

on average, it is possible to describe the whole victimisation profile for the states, meaning that the distribution of crime in a state in Mexico is correctly described by only 5 free parameters. The mean profile has a number of free parameters with an order of magnitude of the size of the population being simulated (a different rate for each individual). The second, and much more relevant weakness, is that the mean profile is based on the *true* distribution of crime and so it includes additional noise from the simulations and there is no theoretical reason why the mean profile gives a better description of the crime suffered by the population other than the fact that it generally gives smooth transitions between groups.

The Monte Carlo procedure is time-consuming. To obtain intervals for the $RECC$ for the 32 states and for the 6 years analysed, more than 2.8×10^{16} Poisson distributions were simulated.

Assuming that there exists of only a few homogeneous groups, in the case of crime suffered, is quite a simplification of a much more complex reality and perhaps the rate profile of a population looks more like the one at the bottom of Figure 5.5 where, by ranking the population from the one who suffers the lowest crime rate to the one who suffers the highest crime rate, a more detailed curve than a piecewise constant function is obtained and so, the interface between one group and the next one is never a drastic increase as the one assumed.

5.4.6 Rationale for case selection: Mexico

The case of robbery of a person in Mexico allows detecting higher degrees of concentration, with certain groups suffering up to eight (or even more) crimes per year. In terms of the data, it provides a large variation at a state level, from the highest crime rate, observed in Estado de México in 2014 $c_{MEX} = 0.471$ to the lowest crime rate, observed in Baja California Sur in 2011 $c_{BCS} = 0.007$ (so a 68:1 proportion of crime rates). Therefore, by considering the case of Mexico, the concentration of crime in states with a crime rate lower than some countries in the developed world to a crime rate

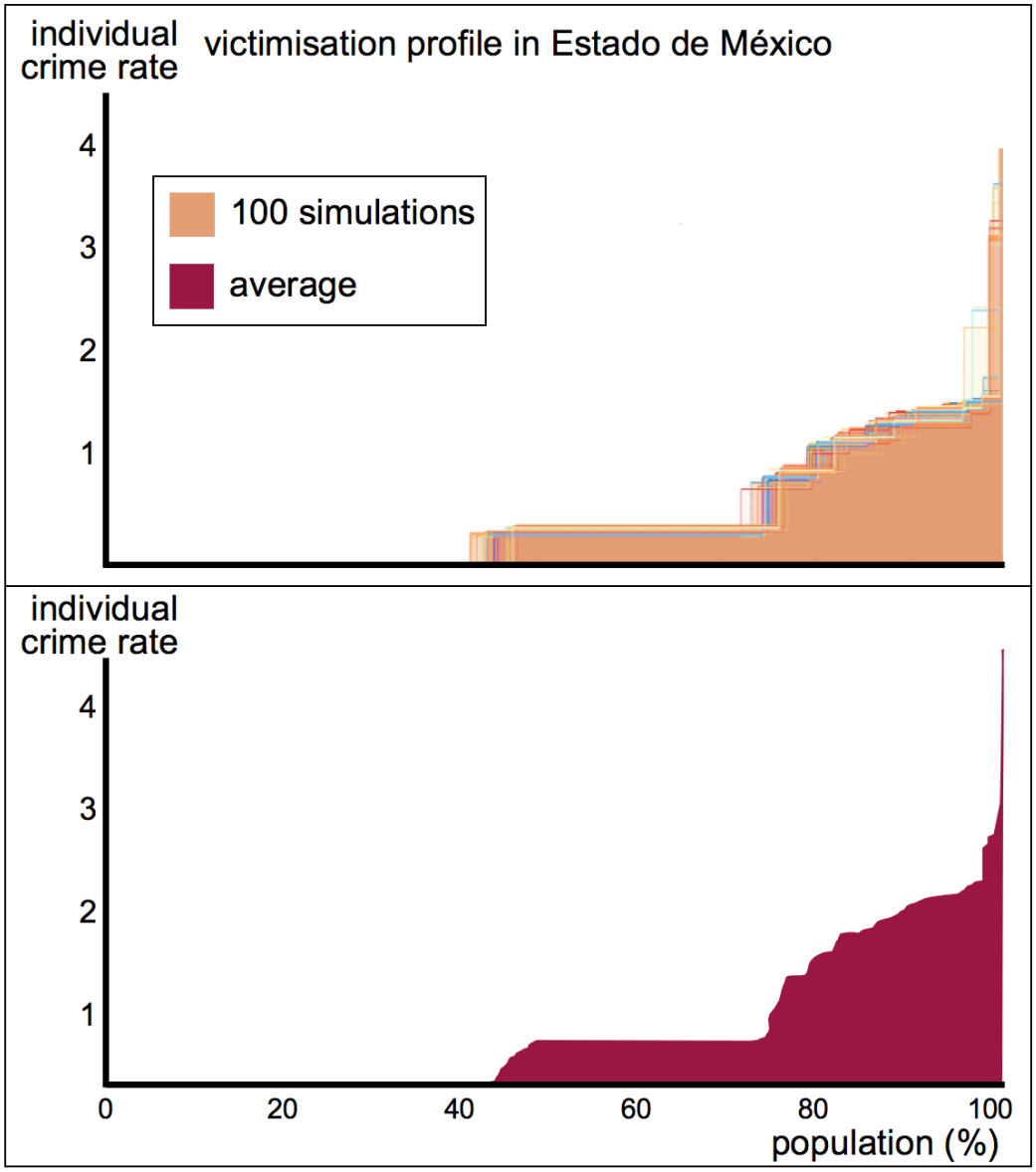

Figure 5.5: 100 simulated victimisation profiles for Estado de México in 2014 (above) and mean simulated profile (below).

so high that almost represents half of the population has been compared. Therefore, it considers a wider spectrum of crime and concentration rates.

In addition, Mexico has several cities, Mexico City is a mega-city but there are more than 10 other metropolitan regions with more than one million inhabitants. Thus, the case includes mega-cities, cities, towns and other population units of different sizes. Besides, Mexico also presents a large variation at the state level. Some states, like the Estado de México for in-

stance, have a similar population to the Netherlands, while other states like Colima and Baja California Sur have a similar population to Luxembourg. Thus, there is a 23:1 size variability in the observations considered so that our metric does not depend on small (or large) populations to work.

Finally, it is a country with increasing crime rates, where security policies have not been clearly successful in the reduction or control of crime nor have they been successful in improving the perception of security. In addition, Mexico offers different and contrasting scenarios when considering its 32 states. On the one hand, the Estado de México has nearly the same population as Chile or Ecuador, and nearly 30% of its population suffered at least one robbery of a person during 2015, and yet, on the other hand, in the state of Zacatecas, less than 1.5% of the population suffered at least one robbery of a person in the same year. Thus, Mexico is a country in which its size and diversity includes different scenarios for the crime concentration at the same time.

Therefore, due to the diversity of crime rates, large population groups and the potential policy implications of the results, Mexico provides a good case study for considering the concentration of crime.

5.4.7 Type of crime: robbery of a person

Between 2011 and 2016 Mexico has seen a slight increase in its crime rates for the particular case of robbery of a person accompanied by a drastic change also in the concentration rate $RECC$. Therefore, using this type of crime also gives observations with variance and dynamism.

The situation of organised crime in Mexico is clearly relevant for some crimes, but the case of robbery of a person is arguably less related to organised crime. Robbery of a person is a more opportunistic type of crime, based on the chances of a criminal meeting a potential victim particularly in the absence of a police presence or CCTV coverage. Hence, robbery of a person provides a good test for new measurements of crime concentration.

The technique could be used in other regions or for other types of crime.

For example, it could help us gain a better understanding of whether the changes in the global distribution of victimisation are indeed the result of a higher degree of concentration of crime (Pease and Ignatans, 2016).

For crimes that are less frequent, it is possible to carry out further analysis of the rare events. For many types of crime, suffering more than one incident is extremely rare (for example, only 43 survey respondents out of the more than 85,000 respondents in 2016 had their car stolen more than once). Results in the case of such a low-frequency type of crime often (although not always) show that a large group suffers zero crimes (so, the existence of the immune group holds for burglary, for example) and then a single group which suffers all the crime, with some rate $\lambda_j > 0$.

Results for the 32 states in Mexico for Car Theft, for example, shows no group with chronic victimisation (as expected since almost no one suffers two or three of this type of crime) and the behaviour in 17 states are best modelled as a Zero-Inflated Poisson distribution, meaning that there is only one group who suffers this type of crime.

5.4.8 Extensions of the Rare Event Concentration Coefficient

The $RECC$ is a metric designed to detect changes in the concentration of crime based on the number of crimes suffered by the individuals. It considers that the observed numbers might be the result of luck and that fluctuations might be present so that it gives a probabilistic approach to the concentration of crime. It is based on some assumptions: individuals suffer crime independently, at a constant rate and suffering a crime does not affect future victimisation. Based on these assumptions, the number of crimes suffered by an individual might then be modelled as a Poisson distribution.

However, looking at the individual observations, such as the crimes suffered by a particular person, the crimes executed by an offender or even its arrests, is not consistent with a constant rate (Bushway and Tahamont, 2016). Also, looking at a more aggregate type of data means reducing the

number of observations, from millions of individuals into a few thousand street segments or regions and it reveals a pattern in terms of the space, the time or both, in which crime occurs (Osgood, 2000). For more aggregate data, assuming independence or a constant rate for each observation might not be appropriate, as most of the observations are not zero and so crime, within this more macro data frame, might not even be a rare event. Therefore, the $RECC$ is not convenient to determine the concentration of crime within this context and more adequate tools for estimating the rate of crime of each of the k observations exist, for example, by considering the rate as a function of time t which might consider space, past victimisation, or even the topology of the street network (Rosser et al., 2016).

Provided that the rate of the i-th region (or a street segment, for instance) is modelled as a function of time $\hat{\lambda}_i(t)$ then measuring the concentration of crime within a more macro data set can be carried out by considering the *Event Concentration Coefficient* ECC (Prieto Curiel and Bishop, 2016a) which is constructed similarly, by computing the Gini coefficient of the k individual rates, even in the case in which they were estimated using a different model and under different assumptions. In this case, the ECC is constructed as

$$ECC(t) = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k}|\hat{\lambda}_i(t) - \hat{\lambda}_j(t)|}{2k\sum_{i=1}^{k}\hat{\lambda}_i(t)}, \tag{5.8}$$

which would also be a function of time. Thus, the $ECC(t)$ gives a measure of the concentration of the rates under a different data frame and with different assumptions, but its interpretation is the same as the $RECC$, so values closer to 1 means a higher concentration of the rates and values closer to zero means a more homogeneous distribution. With the $ECC(t)$, periods of the week in which crime is more evenly distributed or more concentrated in a few places, could be detected.

5.4.9 Should crime be less or more concentrated?

The estimated distribution of crime and its corresponding $RECC$ describe the degree of concentration of crime and, using data from the victimisation survey, gives quantitative results for the different states in Mexico during the years considered. What is not clear is what degree of concentration of crime is preferable.

Crime prevention strategies might result in some displacement of crime (Guerette and Bowers, 2009; Johnson et al., 2014) from one place to the other, to an alternative victim (which is referred to as *target displacement*), to different times of the day, to a different tactic or to a different type of crime (Bowers and Johnson, 2003), which has an effect on the levels of concentration of crime, but then this promotes the question: is it desirable to have less concentrated crime? Clearly, a population with overall less crime is desirable, but what about two populations with the same number of crimes? On the one hand, a high degree of concentration of crime means that fewer people suffer crime, that is, fewer victims, but those victims usually suffer much more than a single crime. With a high degree of concentration of crime, resources might be better targeted to those who suffer most crime in terms of prevention and victim support. On the other hand, a low degree of concentration of crime means more victims, which makes policies less efficient, and it might deteriorate the perception of security in a particular region. Suffering a crime under a low degree of concentration becomes a matter of *bad luck* and not a matter of being socially deprived, a minority, a female or any other attribute which perhaps increases the chances of suffering a crime, therefore, it could be considered as a more fair distribution of crime (Bowers and Johnson, 2003) as opposed to a population with a high degree of concentration of crime.

5.4.9.1 Less or more concentration in Mexico City

The particular case of Mexico City between 2011 and 2012 is analysed. Attention is placed on this particular example for three reasons. Mexico

City is different from other states as it is the only one which is also a single metropolitan area so that any security programs are easily identifiable within its constraints. The majority of other states have three levels of police officers (federal, state and local) but in Mexico City, all security efforts are coordinated by the state police. Secondly, between 2010 and 2012, Mexico City started a security program which consisted of installing more than 13,000 CCTVs across the city and utilising several hundreds of police officers to perform surveillance[2] with a real-time police allocation strategy, investing nearly 500 million dollars in this program alone[3]. Thirdly, it has one of the most drastic changes of the $RECC$ between two consecutive years meaning that the concentration of crime rates across its population significantly changed between these two years.

The most frequently used metrics for the concentration of crime in Mexico City were computed (Figure 5.6) which show that with most of the traditional metrics, almost the same results are obtained. For example, the Gini coefficient is $G_{CDMX,2011} = 0.898$ and $G_{CDMX,2012} = 0.872$ respectively, which only highlights the high concentration of crime, but a barely perceptible change and the reason is that in both years more than 85% of the population did not suffer any crime. The arithmetic correction of the Gini coefficient G' and the average number of crimes suffered by the victims H does show a slight difference between the two years. However, the victimisation profile for these two years are displayed in Figure 5.7 where a drastic change is noticeable.

Using the method described here, for 2011 it was estimated that around 5% of the population suffered a rate of $\hat{\lambda} = 1.35$ and, perhaps due to a policy-oriented to reduce the crime suffered by that population group, for 2012 this high rate was reduced to $\hat{\lambda} = 0.32$, which appears to be a good result. However, during 2011 more than 50% of the population were immune to crime, because their rate was exactly zero, but by 2012, only 28% of the

[2]More at `https://bit.ly/24GlUtw`
[3]`https://lat.ms/2Jm7jga`

metric	year		ratio 2011 / 2012
	2011	2012	
victimisation rate v	0.127	0.148	0.860
crime rate c	0.170	0.173	0.982
standard deviation σ	0.507	0.514	0.986
coefficient of variation	2.990	2.9676	1.005
concentration H	1.330	1.166	1.141
Gini coefficient G	0.898	0.872	1.031
generalised Gini coefficient G	0.401	0.257	1.560
$RECC$	0.680	0.367	1.853

Figure 5.6: Metrics of the concentration of crime observed in Mexico City in 2011 and 2012.

population was immune to crime, and in fact, the global crime rate increased from $c_{CDMX,2011} = 0.170$ to $c_{CDMX,2012} = 0.173$.

Between 2011 and 2012, some people suffered a lower crime rate while some suffered a higher rate, which means that comparing individual rates does not provide much information (Figure 5.7). However, for 2011 the $RECC_{CDMX,2011} = 0.680$, shows a much higher degree of concentration than for 2012, where the result was $RECC_{CDMX,2012} = 0.367$ and Figure 5.6 shows that the values are statistically different. Thus, between 2011 and 2012, the results obtained in Mexico City indicate that there was not a reduction in the crime rates, but rather it was merely a target displacement with a lower level of concentration of crime. In this period, the crime rate did not change, in fact, it increased slightly, meaning that the surveillance program with a real-time police allocation strategy might have simply induced a target displacement.

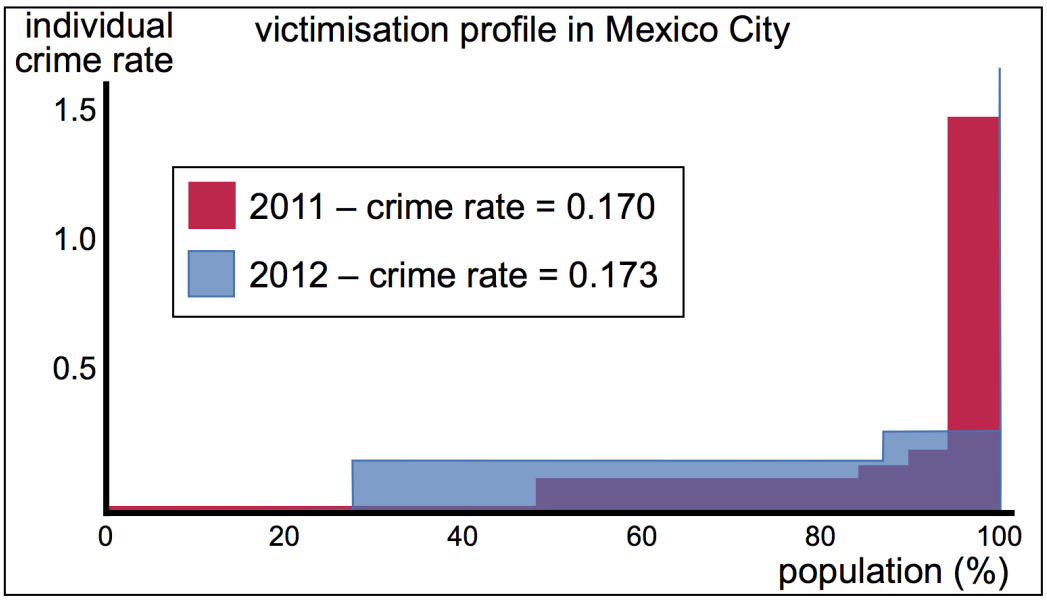

Figure 5.7: Victimisation profile (individual crime rates $\underline{\lambda}$ and group sizes $\underline{q}$) for Mexico City between 2011 and 2012.

5.5 Remarks

Although this probabilistic approach to studying crime and victimisation rests on limited assumptions (i.e., that crime occurs independently, and that being victimised once does not affect the likelihood of being victimised again), by considering the number of crimes that each person suffers and modelling its random component, pieces of valuable information about the crime problem are gained. Firstly, reject the notion that every person suffers the crime at the same rate (the homogeneous distribution of crime in the population is far from being observed).

Secondly, the mixture model provided a distribution of crime rates for the entire population. The model has only a small number of parameters, and so the distribution is useful for simulating the number of crimes that a population might suffer.

Thirdly, the approach allows detecting whether a group exists which expects to suffer more than two crimes in a year, so they suffer from chronic victimisation and it also allows detecting the existence of a group which is immune to crime. The existence of this immune group and the chronic group

is based on the distribution of the crime rates and not directly on the number of crimes so, for example, a person who suffered no crime during a year is not necessarily immune to crime. Instead, they might have a positive rate of suffering crime but it was just lucky that they experienced no crime. A similar reasoning applies to a person who suffered two or more crimes, which could also be observed with a small rate.

5.5.1 Victimisation profile

The model presented here captures the general behaviour of the distribution of crimes. It is easily displayed through the victimisation profile, a novel way to graphically display the distribution of the crime rates suffered by the population, which is comparable between different populations and over a different period and so it also provides a versatile tool for dissemination purposes.

A global measure for the concentration of crime (the $RECC$) is constructed, which is a comparable metric between different periods or different populations. The $RECC$ is an adequate tool for measuring the concentration of crime and it takes into account that crime is rare, highly concentrated and depends on random elements.

By using data from a victimisation survey in Mexico, and considering a different distribution of crime over its 32 states over six years, this study revealed that in most cases there is a considerably large population group which suffers a crime rate equal to zero so that they are statistically immune to crime. Also, some states have a small population which suffers chronic victimisation, meaning that they expect to suffer two or more crimes. These sorts of questions could also be answered using other data sources, for instance, the National Crime Victimization Survey (Bureau of Justice Statistics, 2016), the Crime Survey for England and Wales (Office for National Statistics, 2016) or perhaps by examining crime calls at addresses or street segments (Hipp and Kim, 2016).

Using the specific case of the robbery of a person suffered in Mexico

City between 2011 and 2012 showed a real scenario for which the traditional tools for measuring crime concentration failed to detect any major changes between one year and the next one, but the victimisation profile and the corresponding $RECC$ showed a drastic change in the observed pattern in which that crime is suffered.

5.5.2 Beyond the victimisation profile

A similar analysis could be conducted by considering the number of crimes committed by the population. With proper data, the rates at which the population commits crime could be constructed and divide the population into the non-criminal group, the one-time offenders and the chronic offenders (Wolfgang, 1983) which would be helpful to determine the effect of a preventive program.

This work goes towards the provision of adequate tools in the field in response to the need highlighted by the *law of crime concentration* (Weisburd, 2015). Having adequate tools to measure the global concentration of crime, both from the crime that is suffered and the crime that is executed by the population, or whether is a spatial metric or based on the time of the crime, can help researchers and policymakers better understand the crime problem and what can be done to fix it.

5.5.3 Policy impact of crime concentration

Crime and security-related issues are not constrained to a particular city, state or even country. Not only economically speaking countries are regionally similar, but also, shared history, urban structures and many other social attributes might have an impact upon the way crime is suffered, the reactions from government and society to fight against it and the perception of security/security which it causes.

In terms of a policy design, the degree of concentration of crime actually poses a relevant question. Assuming two populations with the exact same crime rate, having crime more or less concentrated might be the result or

even the objective of a certain policy. One the one hand, having a higher concentration of crime reduces the number of people who actually suffer crime, which sounds desirable and perhaps improves the quality of the attention provided to the victims, but on the other hand, a higher concentration of crime as a result of a policy is unfair to the victimised groups, which tend to be minorities and people with social deprivation. This poses the question of desirability. Even with a metric to determine the levels of concentration of crime, its effect might go beyond crime itself. A higher or lower level of concentration of crime might encourage social segregation, might increase the social deprivation from the already highly victimised groups and might deteriorate the perception of security from a region. These effects are investigated in subsequent chapters.

Chapter 6

Regional fear of crime

Fear of crime is an opinion, or an idea, result of suffering a crime but also, many more elements, including the fear of others, media, demographic and regional aspects and more. Quantifying fear of crime poses serious challenges, as it implies mapping human feelings into numbers, ordering then, saying who is more fearful and which region is perceived as the most dangerous. Although quantifying fear has some weaknesses, it gives results in terms of why a person fears crime, how is it updated, how does it emerge in a society and so it gives valuable results which can then be applied to the security policies of a city.

This chapter quantifies the perception of security of a region and gives a metric comparable between different regions. It is then applied to different regions of Mexico and compared against observed crime, and so a weak relationship between different types of crime and its fear is established. It is based on published research (Prieto Curiel and Bishop, 2016b).

6.1 Defining fear of crime

Fear of crime is a feeling, perhaps even an instinct, which is based on the self-perceived risk of suffering a crime. This fear encourages people to avoid certain streets, or to walk faster during the night and motivates individuals to spend more than one month of their lives locking and unlocking cars

and buildings (Anderson, 1999) or taking other security measures as part of their daily routine (Jackson and Gray, 2010). Fear of crime leads to those who are more prosperous to protect themselves and their property, possibly displacing crime to those less privileged (Box et al., 1988). Fear can also transform some public places into no-go areas (Morgan, 1978) which has a severe impact on the local prosperity and economy (Brands et al., 2015). It has been argued that the perception of insecurity emerges as a social problem (Austin et al., 2002) and it is now becoming one of the main concerns for residents and administrators in almost every large city (Carro et al., 2010). Falsely based perceptions affect the efficiency of the security systems, since governments are encouraged to spend resources, such as an increased number of police officers, or even introduce urban interventions, in places where people are more concerned, but not necessarily where action is most needed or where it could have the greatest impact (Grogger and Weatherford, 1995).

It is the fear of crime, not necessarily crime itself, that motivates people to take healthy precautions, such as locking the doors of their home (Jackson and Gray, 2010), and creates awareness and caution (Skogan, 1987), but, in extreme cases, fear has a severe impact on the quality of life, it might cause paranoia (Ruijsbroek et al., 2015), tension, social isolation (Jackson and Gray, 2010) and fear is the reason why millions of people have been displaced from their own home (Albuja, 2014; Cantor, 2014).

The study of the fear of crime and the perception of security has many issues, which begin with the concept itself. Fear of crime and perception of security represent conceptually distinct constructs, although they have some similarities (Wilcox Rountree and Land, 1996), especially in the way both concepts have been studied (Farrall et al., 1997). A person, for example, might perceive his or her neighbourhood as being insecure, but at the same time might not be afraid of crime since the person does not consider themselves to be a potential target or considers that they have taken suf-

ficient precautions to avoid being the victim, however, this is not the most common scenario. From here onwards the term *perception of security* and *fear of crime* are used to refer to the answers provided in the surveys.

Quantifying the perception of security, or the fear of crime, poses a serious challenge. Firstly, the concept itself. Fear of crime has acquired many divergent meanings and it is almost never defined, but it is implicitly considered as the perception that a person is likely to be the victim of a crime (Ferraro and Grange, 1987). Secondly, the fact that two people feel a particular place to be *insecure* does not mean that they fear the same things or to the same magnitude. However, by imposing a metric for the perception of security and quantifying the fear of crime, many factors affect which affect the levels of fear are detected, such as demographic variables (e.g. age, race and gender (Carro et al., 2010; Brunton-Smith and Sturgis, 2011)), physical factors (the location of a person or a house or vandalism in particular streets, etc. (Pantazis, 2000; Tseloni, 2007)), past victimisation and more.

It has been found that the perception of security is related to age, gender and income (Wilcox Rountree and Land, 1996); minorities tend to have an increased fear of crime (Kershaw and Tseloni, 2005); and even the physical condition of the neighbourhood has an impact on how secure it is perceived (Tseloni, 2007). A person might consider a place to be either secure or insecure based on their previous experience in that place, what they have heard or seen, the fact that it is dark or rundown (Austin et al., 2002) and other factors. To protect their person and belongings, they may choose a different route to work, cross the street when the lighting is poor, change job or even choose to carry a weapon.

Up to now, the standard way to assess the perception of security has been to use victimisation surveys (Gottfredson and Hindelang, 1981). Victimisation surveys are a powerful tool to understand the crime suffered by a population and have been available for more than 40 years in some coun-

tries. Victimisation surveys, such as the Victimisation Survey from Mexico (from INEGI, 2014), the National Crime Victimization Survey in the United States (Bureau of Justice Statistics, 2016) or the Crime Survey for England and Wales (Office for National Statistics, 2016), which allow agencies and governments to obtain precise information about the fear of crime and they are widely consulted by policymakers and researchers. Using these surveys, it is possible to estimate the rate at which crime is suffered by a population, including the crimes that are not reported to the police, monitor the fear of suffering a crime and more. Using victimisation surveys it has been shown, for example, that people with a lower income tend to have more fear of crime than those with a higher income (Pantazis, 2000); more than half of the population said that they had been worried about being the victim of a crime sometime in the past (Farrall et al., 1997); an increased fear of suffering a burglary was in fact correlated with the risk of suffering one (Borooah and Carcach, 1997); minorities tend to be more fearful (Brunton-Smith and Sturgis, 2011) and that older people tend to feel less secure (Carro et al., 2010).

6.1.1 Media and the fear of crime

Another relevant factor to consider is the role that media (and perhaps also social media) plays in the fear of crime.

Regarding the content of media, by analysing what is being published by newspapers, magazines and others, (Berelson, 1952), the focus of the media can be assessed and any bias that exists between what is published and the reality can be detected (Ditton et al., 2004). For instance, it is known that there is more coverage of violent crimes than other crimes (Ditton and Duffy, 1983), with tabloids tending to publish more sensationalist news items (Dickinson, 1993), meaning that what is published in the news differs from reality in some sense. It is estimated that less than 1% of the crime features in the newspapers (Chadee and Ditton, 2005) and that crimes of a sexual or a violent nature have a much higher probability of appearing in the news

(Ditton and Duffy, 1983). In fact, even though murder is one of the least frequent crimes, it makes up for nearly one-third of the crime stories in the newspapers (Liska and Baccaglini, 1990).

In terms of the audience, readers choose which articles to read and therefore those readers who pick crime stories already are prepared to approach the topic (Lane and Meeker, 2003); also, the impact of a specific report depends on whether or not justice is restored at the end of the crime stories, not the actual number of crimes (Heath and Gilbert, 1996); and finally, the individual interpretation, often referred to as the "reception" of media, alters how two individuals might read the same news and obtain a different perception as a result (Ditton et al., 2004).

There are some very weak results about fear of crime and the media (Hollis et al., 2017) and it is not clear whether people who read more newspapers or listen more frequently to the radio, for instance, tend to have a lower or higher fear of crime.

Therefore, with no clear impact from the media in terms of the fear of crime, its relevance is omitted from the analysis and from any further modelling.

6.1.2 Defining the regional fear of crime

A person's perception of the level of security at a specific location depends on many factors, including past experiences in that location, the actual crime suffered by the population and more. Thus, when the individual perception that a location is insecure becomes the general rule or an 'agreement' from its population is when the perception of security becomes an attribute of the region rather than the fears of some of its individuals, hence the relevance of aggregating individual perceptions of security into a single *regional* perception of security.

In a particular region, people's perception of the level of crime is formed by a number of factors where the first and most clear one is whether they were personally the actual victims of a crime: past victimisation almost dou-

bles the odds ratio of a person having fear of crime (Tseloni, 2007). However, past victimisation is not the only factor that contributes to a person fearing crime or perceiving a region as insecure and one of the main reasons is that, fortunately, crime is a rare event. For example, the International Crime Victims Survey shows that less than 3% of the population from the surveyed countries experiences a theft from person during the period of one year (Tseloni et al., 2010), and rates are similarly low for other types of crime. Hence, more frequently, a person experiences indirect victimisation via interactions with friends, neighbours or through media rather than experiencing actual crime (Gilchrist et al., 1998).

There are many other reasons which affect whether a person feels secure or not in terms of crime. It was shown, for example, that some area characteristics, such as the economic level or the number of people of age between 16 and 24, might act as a better predictor of fear of crime, even better than the actual crime rates (Kershaw and Tseloni, 2005).

The perception of security and the fear of crime have been studied from many angles: from its social and psychological construction (Farrall et al., 2000); its relation with the environment (Brunton-Smith and Sturgis, 2011); the impact of past victimisation to the current perception (Hale et al., 1994); to more methodological aspects such as how the fear of crime is measured (Farrall et al., 1997).

6.1.3 Quantifying the regional fear of crime

To understand the perception of security attached to a particular region, consider the number of people surveyed that perceives the region to be either secure or insecure, where the term *region* here might be as specific as a park or as general as a state. Typically in crime surveys, an individual might be asked whether a region is secure or not, so a binary answer is usually recorded. Other studies about the fear of crime or the perception of security have been conducted, and the techniques used strongly depend on the type of data that is used. For example, the Crime Survey for England and Wales

(Office for National Statistics, 2016) considers questions such as "*How safe do you feel walking alone in this area after dark?*", and the respondent has four different options: 1-Very safe, 2-Fairly safe, 3-A bit unsafe, and 4-Very unsafe; these responses provide an ordinal variable (where the order matters) and, in order to combine different questions and answers into a single number, a common technique is to assign a number to each response and sum them into a single *fear rate* (Kershaw and Tseloni, 2005), which might then be used in a statistical model (Tseloni, 2007).

Expressing the number of people who consider a region to be insecure as a ratio to the total number of people, produces a number between 0 and 1, where a value close to 0 means that the region is considered to be secure, and a value of close to 1 means that the region is considered to be insecure. Dividing the space of analysis (urban or otherwise) into a number of non-overlapping regions $R_1, R_2, \ldots, R_n$, gives the perception of security of the general region R_k by this ratio, denoted by s_k. Thus, s_k represents the mean perception of the surveyed population from a region and it provides an estimate of the probability that a randomly selected person from the region R_k, considers that region to be insecure. If enough surveys were conducted in the region R_k, then that number s_k is a reasonable estimate of that probability.

This is a simple way to quantify the perception of security of a region, however, this number s_k, by itself, is not particularly useful in isolation, since the norm is unknown, so for example, if a region gives $s_k = 0.4$, is that considered to be high or low perception of security? It is much better to compare this number against the equivalent estimates in other regions or compare this regional measure over time to see the effect of either changing crime numbers or changing police strategies.

The process of dividing into non-overlapping regions and taking into account the perception of security as a combination of the opinions of individuals is schematically drawn in Figure 6.1.

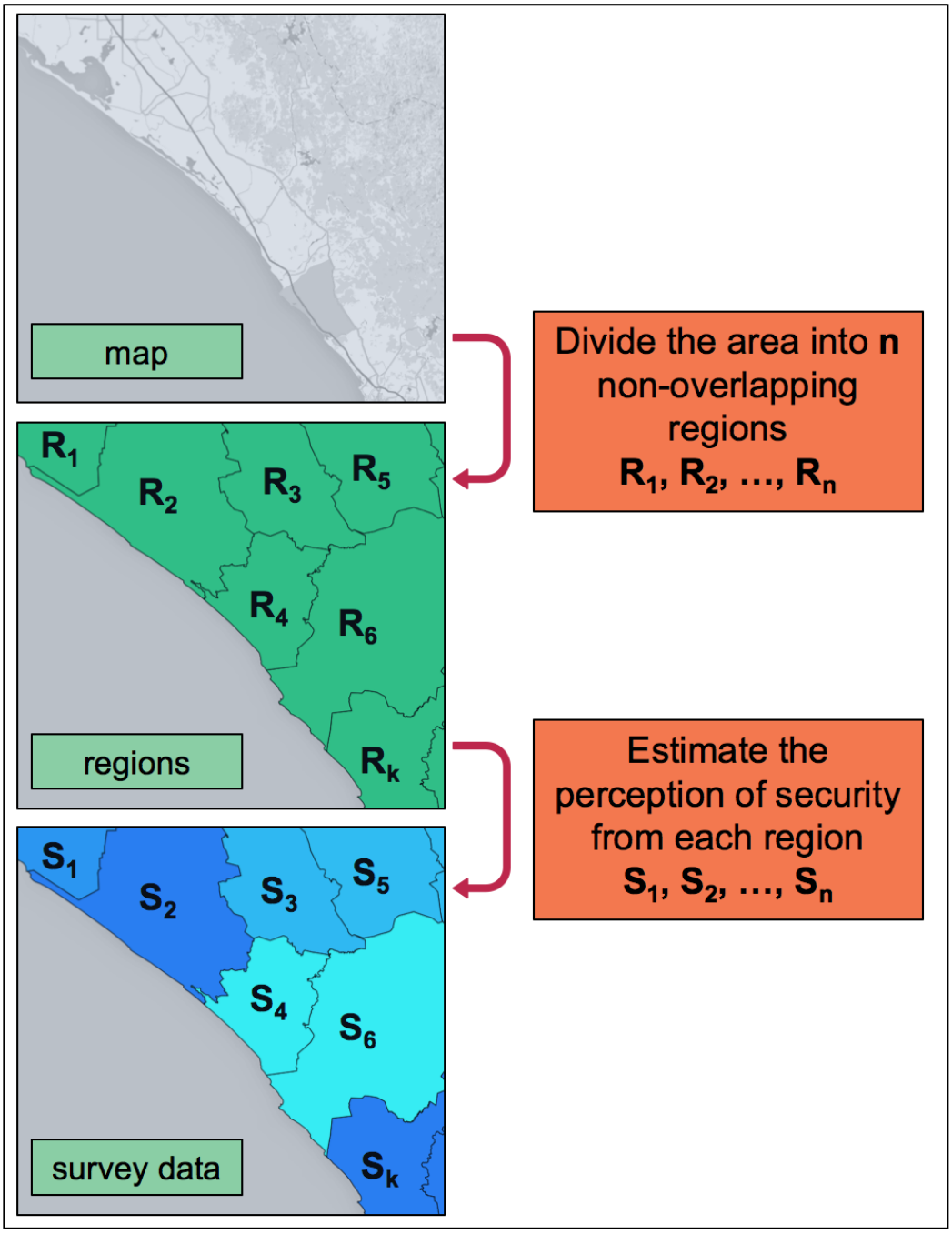

Figure 6.1: Process of dividing the area into non-overlapping regions and obtain their perception of security from survey data.

Comparing the perception of security between two different regions is the crucial point since it allows quantifying the perception in a mathematical way and then to use it in a modelling context.

6.2 Case study - fear of crime in Mexico

6.2.1 Data description for the fear of crime

The national victimisation survey conducted in Mexico in four yearly periods from 2011 to 2014 is used (from INEGI, 2014), in which they ask, amongst others, the following questions:

- in terms of crime, do you consider your locality to be secure or insecure?
- in terms of crime, do you consider your county to be secure or insecure?
- in terms of crime, do you consider your state to be secure or insecure?

These questions help measuring the perception of security from people at three different geographic levels: locality, county and state, where the county is formed by a set of localities and a state is formed by a group of counties. The answers are binary, so the person being interviewed was only allowed to answer if he or she considered the region to be either secure or insecure in terms of crime.

Familiarity with the area reduces the worry about suffering a crime (Gilchrist et al., 1998) which implies that smaller regions have a tendency of being perceived as more secure than larger regions, and so the state level is too general and gives non-comparable observations, taking in account that there are some states (such as Chihuahua) which are larger in area than England, but there are some states (such as Morelos) which are smaller in area than Cyprus. Hence, the measurements at a state level are dismissed. The level of locality is, on the other hand, perhaps too specific and it does not provide a clear distinction in largely populated areas (such as Mexico City, which is divided into a few thousand localities). Hence, the region of measure of a county is used in this study.

The way in which the survey was conducted considers the number of people from the population represented by each of the respondents based

on their demographics so that the data contains an expansion factor, which is interpreted as the number of people represented by each observation and it goes between a few dozens to a few thousand. To avoid considering observations with only a small amount of respondents to the surveys, only counties with more than 300 people answering the survey in 2014 are considered, resulting in a total of 53 counties. Summary statistics for the perception of security data obtained for the four years considered is displayed in Table 6.1.

	surveys	perception of security			
year	considered	mean	min	max	std. deviation
2011	32,402	0.649	0.209	0.947	0.190
2012	42,033	0.631	0.169	0.932	0.194
2013	42,527	0.656	0.270	0.903	0.160
2014	34,544	0.660	0.246	0.922	0.153

Table 6.1: Number of surveys considered and summary statistics for the perception of security, where a value closer to 0 means that in terms of crime people feel more secure and a value closer to 1 means that people tend to feel less secure.

The perception of security is a metric which allows differentiating counties based on the responses obtained from the survey. It shows, for example, that within the four years and in the 53 counties considered, the perception of security went as high as $s_k = 0.947$ (Ciudad Juárez in 2011), which can be interpreted as the probability that taking a person randomly from that county, he or she considered it to be insecure, with the answer being surprisingly high. The lowest perception of security (which means that the majority of the county considered it to be secure) was $s_j = 0.169$ (Mérida in 2012), which is a region of the country considered traditionally as being the most secure. This perception of security works as a quantitative measurement for the way in which a county is perceived and is not based on a single individual, so it gives a robust estimate for the regional perception of security; it helps differentiate counties which are generally perceived as being secure to those counties which tend to be perceived as insecure. Figure 6.2 represents the

perception of security from each county in Mexico during 2014. The perception of security is reasonably homogeneous, in that a county which is perceived as secure has similarly perceived counties as neighbours.

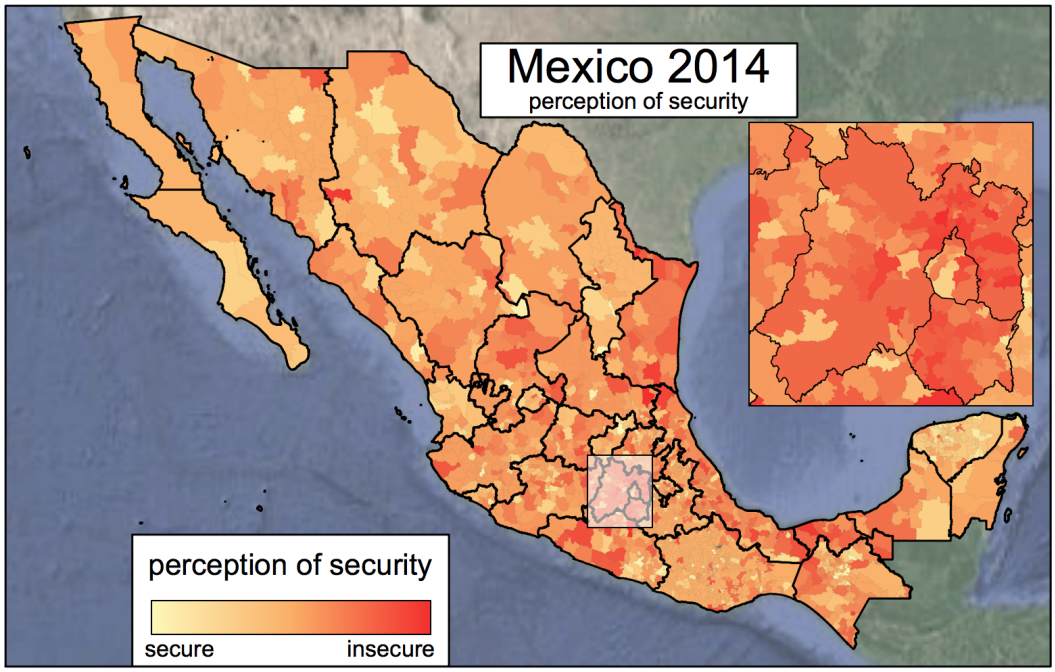

Figure 6.2: The perception of security in Mexico by counties. For counties with no information or just a small amount of survey respondents, the average perception of security, obtained as the average of the whole state is plotted, only for display purposes.

6.2.2 Victimisation rates

The objective is to relate the obtained metric for the perception of security to the victimisation rates. Eight different types of crime are considered:

1. robbery of a person
2. car theft
3. partial car theft
4. burglary
5. vandalism
6. kidnap
7. murder

8. missing person.

The same survey that was used to construct the measure of the perception of security is used to estimate the victimisation rates. From the same survey, they ask, amongst others, the following questions:

- During the previous year, did any member of this home, including yourself, suffered a *type of crime being considered*?
- How many times did you or any member of this home suffered a *type of crime being considered* during the previous year?

These two questions are asked for each type of crime and help quantify the different victimisation rates in each of the regions. In the case of robbery of a person the question does not include other members of the home, and in the case or murder, missing person and kidnap the question is asked only with regards to the members of the home. All cases are self-reported, so it depends on the ability of the survey respondents to recall their experiences in terms of crime (Farrall et al., 1997), and so there may be sources of error, nevertheless, they lead to valuable information. This part could be computed or merged with police recorded crime data, however, in Mexico, less than 8% of the crimes get recorded by the police (from INEGI, 2014).

An estimate of the number of victims of each type of crime during a year is obtained, expressed as $v_k^{(i)}$ for the i-th type of crime and for the k-th county. Reported victimisation rates for the eight types of crime considered are reported in Table 6.2.

6.3 Crime and its fear

Due to many reasons (such as the lack of information, the fact that crime is a rare event and the role that the media plays) a person might not be able to correctly predict their own chances of suffering a particular type of crime and therefore, fear of crime might not be related to the risk that a person has or their past experiences. From a regional level, it might be assumed

crime	mean	min	max	std. deviation
robbery of a person $v^{(1)}$	0.091	0.007	0.45	0.066
car theft $v^{(2)}$	0.030	0.002	0.096	0.019
partial car theft $v^{(3)}$	0.125	0.025	0.250	0.054
burglary $v^{(4)}$	0.066	0.023	0.149	0.025
vandalism $v^{(5)}$	0.107	0.014	0.222	0.053
kidnap $v^{(6)}$	0.004	0	0.014	0.003
murder $v^{(7)}$	0.001	0	0.007	0.001
missing person $v^{(8)}$	0.002	0	0.006	0.001

Table 6.2: Victimisation rates, range and standard deviation for the eight types of crime considered in Mexico during 2014. Data obtained from the victimisation survey.

that fear of crime is a direct result of the crime in a region and therefore, if a region is deemed as being insecure, then the region has a criminal problem that should be solved. Although it seems reasonable that fear of crime is a consequence of actual crime, if this relationship is weak or if many more factors are related to fear of crime rather than just crime, then any policy oriented to reduce fear of crime, even if it is successful at reducing any actual victimisation, might not have much impact on the perceptions of the population.

Having a metric for the regional fear of crime and a metric for the victimisation of that region allows comparing them (Figure 6.3).

6.3.1 Ranking the regional fear of crime

Given a method to establish the regional perception of security allows ranking all regions in the space. Assume for the moment that the perception of security from each region is different, and let $S_k \in \{R_1, R_2, \ldots R_n\}$ be the unique region which occupies the k-th position on the ranking of the perception of security, so S_k reflects the perception of security of that region when compared to the other $n-1$ regions. The region R_j such that $S_1 = R_j$ is considered the most insecure since it occupies the first place in the ranking, and the region R_l such that $S_n = R_l$ is considered the most secure and is

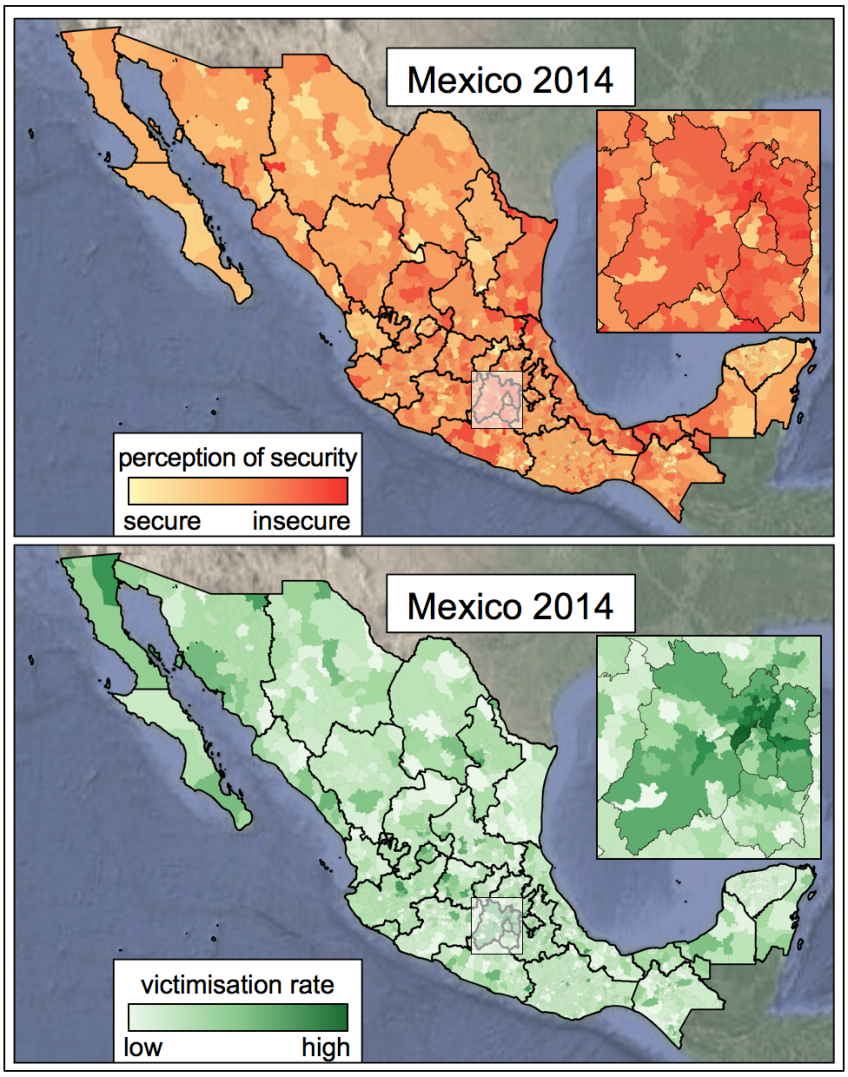

Figure 6.3: The regional perception of security (above) exhibits a spatial pattern. Specific metropolitan areas (such as Mexico City and some cities along the border with United States) are considered less secure. Victimisation rates (below) also show a spatial structure, perhaps, only slightly related to fear of crime and so the most victimised regions are not necessarily the most feared regions.

positioned in the last place of the ranking. Let $S = (S_1, S_2, \ldots S_n)$ be the ranking obtained from the security perception, starting from the one considered to be the least secure up to the one considered the most secure. Now, if the survey considers a different type of question (or questions) to determine the perception of security from the survey respondent, it is also possible to either transform the response into a binary variable or to assign a number based on the order of the response (Kershaw and Tseloni, 2005). By either

recoding the response or by considering the sum, a similar ranking of the perception of security S might be obtained.

The perception of security between two different regions is comparable and it is possible to state that, if in the ranking S, the region R_k is listed before R_j then this means that the region R_k is considered to be less secure than R_j, so $s_k > s_j$. Thus, with n regions there are $n(n-1)/2$ comparisons by taking each pair of regions and selecting which of the two is considered to be the most secure and which the least.

Let v_k be the rate in which the region R_k is victimised. This has many interpretations and depends on the type of crime and the period considered, but suppose that v_k represents the probability that a person suffers a particular type of crime, such as robbery of a person or burglary, at least once, in a yearly period. Then v_k provides information to compare two different regions and if $v_k > v_j$ it means that the population in the region R_k suffers a higher probability of being the victim of a crime than the region R_j for the type of crime considered. If the victimisation rate is different in every region, then it also provides a unique way to rank the n regions. Let $V_k \in \{R_1, R_2, \ldots R_n\}$ be the unique region which occupies the k-th position on the ranking of the victimisation rate. The region R_j such that $V_1 = R_j$ has the highest victimisation rate and the region R_l such that $V_n = R_l$ has the smallest victimisation rate. Let $V = (V_1, V_2, \ldots, V_n)$ be the ranking obtained from that victimisation rate, as an ordered list of the n regions, from the one with the highest victimisation to the one with the lowest victimisation rates.

Consider now the victimisation rate and the perception of security as the ranking obtained when the regions are sorted from the one with the highest victimisation to the one with the lowest victimisation rate as V and from the one perceived as less secure to the one perceived as the most secure, as S. The objective is to analyse the relationship between both rankings and so all possible scenarios are considered. For example, in the case of a tie, that is, if there are two regions such that $s_k = s_j$, then the ranking of the perception

of security would not be unique, and a similar situation happens if there are two regions such that $v_k = v_j$. Although ties are very unlikely to occur, the index of the regions j and k is used, and if $j < k$ then the R_j appears first in the ranking, and the same criterion is applied to the victimisation rate. The result is a unique ranking of the regions based on the perception of security and a unique ranking of the regions based on the victimisation rate (Figure 6.4).

Robbery of a person is used as an initial approach to the victimisation rates hence $V^{(1)}$ is the ranking from the 53 counties, from the one which suffers the highest amount of robbery of a person, to the one which suffers the lowest. The ranking metric based on the permutations is displayed in Figure 6.4, where the first column is the victimisation rate ranking, $V^{(1)}$, with the counties that suffer the higher rates in the upper part, and the second column displays the perception of security ranking, S, with the counties perceived as the least secure in the upper part. A line is drawn between the same county in the two rankings. Hence, a horizontal line indicates that a county is ranked in the same place in both $V^{(1)}$ and S, and intersections between any two lines mean that the corresponding counties are not in the same order in both rankings. Perhaps as expected, in Figure 6.4 we can identify that it is not usual for a county to have a high (low) victimisation rate and to be identified as secure (insecure).

6.3.2 Compare two rankings

From the perception of security and from the victimisation rate, two rankings are obtained (S and V), which may, or may not, be associated with each other and the objective is to quantify the degree of association. If the list consists of n elements, i.e., n different regions, then there is a total of $n(n-1)/2$ possible permutations, or distinct rankings in which the regions might be ordered. To determine the *degree of similarity* between the two rankings a metric is constructed, based on how far away is one ranking from the other, that is, how many movements would it take to go from one ranking

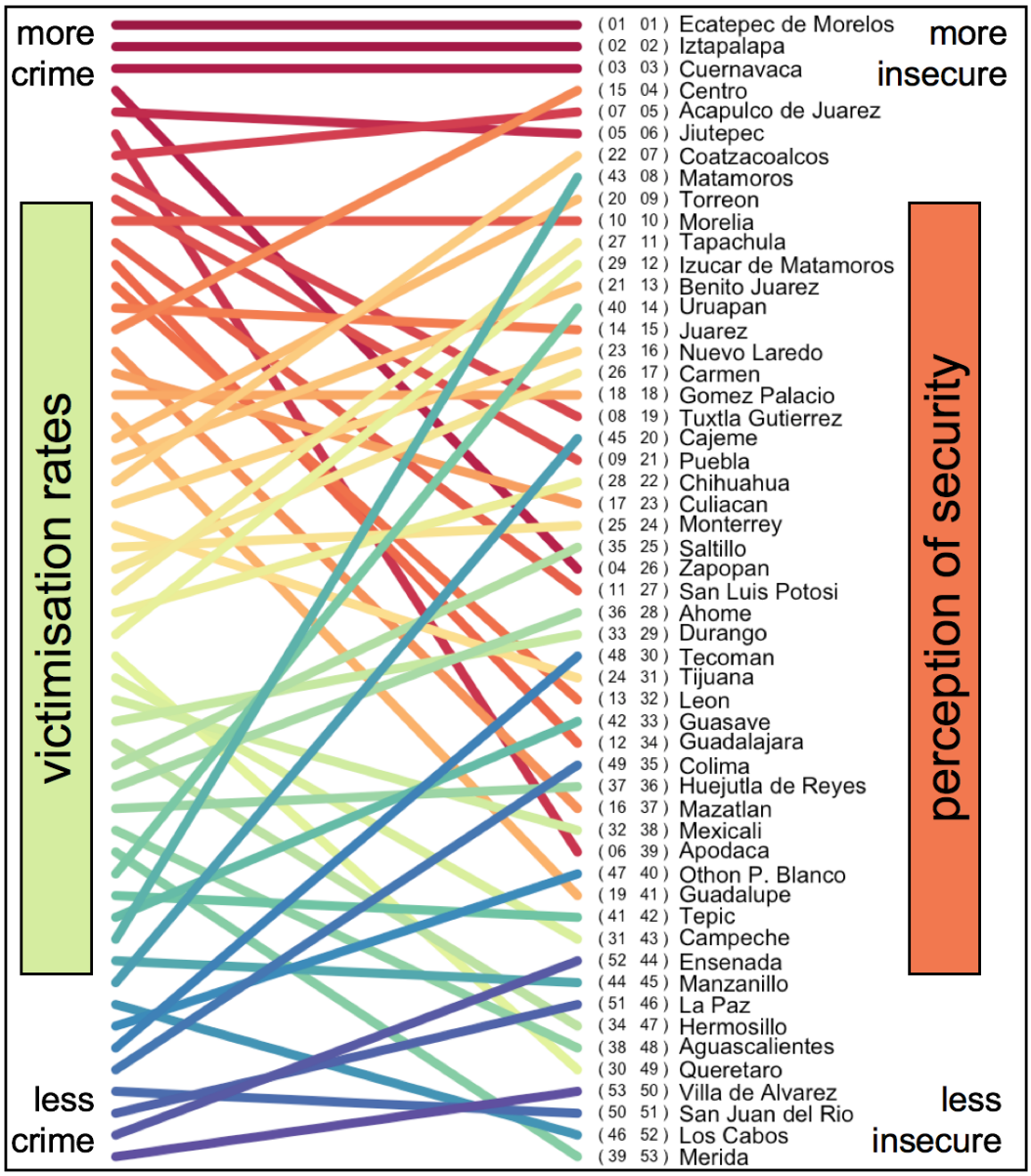

Figure 6.4: The ordering in the first column shows the victimisation rate ranking of robbery of a person obtained from the 53 counties in Mexico and the second column is the ranking of the perception of security/insecurity from that county. Places with a low level of victimisation are usually ranked as secure. The numbers inside the brackets (in small font) are the actual rankings from each of the regions on both of the lists, that is V and S.

to the other. More formally, define a *swap* to be the permutation of any two neighbouring elements on the list, and a metric to compare S and V can be constructed by counting the minimal number of swaps required to go from one ranking to the other. For example, if the two rankings are identical, then

it means that no swap is needed; if only the first two elements of S and V are in reversed order, then only one swap is needed and so on. The maximum number of swaps would occur in the scenario where S and V provide the same order, but reversed, so that the first place of S is the last place of V and so on, and in that case $n(n-1)/2$ swaps would be required to go from one ranking to the other.

Let p be the minimal number of swaps required to go from ranking S to ranking V. The *ranking metric* $P(S,V)$ is then given by

$$P(S,V) = \frac{n(n-1)-4p}{n(n-1)}, \tag{6.1}$$

which measures the number of swaps required to go from S into V and compares it against the maximum number of swaps. Since $p \in \{0,1,\ldots,n(n-1)/2\}$, then $P(S,V) \in [-1,1]$. Thus, when $p = 0$ then $P(S,V) = 1$ and it means that rankings S and V are identical; a small value of p means that it only requires a few swaps to go from one ranking to the other, and so $P(S,V)$ is close to one. When p is closer to $n(n-1)/2$ it means that it requires most of the possible permutations, so S and V also have a relationship, only they provide a reversed order, and in this case, $P(S,V)$ is close to -1. Finally, when the number of swaps required to go from one ranking to the other is closer to $p = n(n-1)/4$, which is the middle between the largest and the smallest amount of swaps, then $P(S,V)$ is close to 0.

A similar problem arises when trying to compare the ranking provided by two different search engines (Kumar and Vassilvitskii, 2010). This metric is also known as Kendall Tau Rank Distance (Shieh, 1998) or the Kendall Rank Correlation Coefficient.

An alternative way to interpret such a metric is based on the multiple comparisons which the rankings S and V allow. Select a pair of regions, R_k and R_j and compare their position in both S and V, then there are two possible scenarios, either both regions have the same order in both rankings, which means that in that comparison, the region with the highest victimisa-

tion rate is perceived as being less secure. The second scenario is that they do not preserve the same order, meaning that the region with more victims is considered to be more secure. The metric is defined as the number of times that a comparison preserves order in both of the rankings, S and V, against the total amount of comparisons that can be made by taking two different regions.

6.3.3 Perception of security and victimisation rates

Using the same data, both rankings $V^{(1)}$ and S might be displayed as an upper triangular matrix T_{ij}, with each of the counties as the rows and columns of T_{ij} and the result of the comparison between the i-th and the j-th county as the entry (i, j), where a value of 1 is assigned if these two counties have a different order in both rankings and a value of 0 otherwise (so that T_{ij} identifies every pair of counties in which the one with the smaller victimisation rates is perceived as being less secure). The sum of all the entries of T_{ij} gives p, the number of swaps required to go from one of the rankings to the other.

Using different colours to identify the entries of T_{ij} which are either 1 or 0, the results of the comparison between each pair of counties is displayed in Figure 6.5. The column on the right-hand side displays the percentage of counties which have a different order against the corresponding county, and it reveals that the counties with the lowest number of differences are those in which their victimisation level is either so high or so low that they are easily identifiable as secure or insecure. However, the counties which have a higher number of differences are the ones in which the victimisation does not correspond to the perception of security. Two examples are the counties of Apodaca and Guadalupe, both in the Nuevo León state, which have a high victimisation level but are perceived as relatively secure.

When investigating the relationship obtained between the perception of security and the victimisation rate, the value of the metric and its display might, statistically speaking, be the result of randomness. However, there

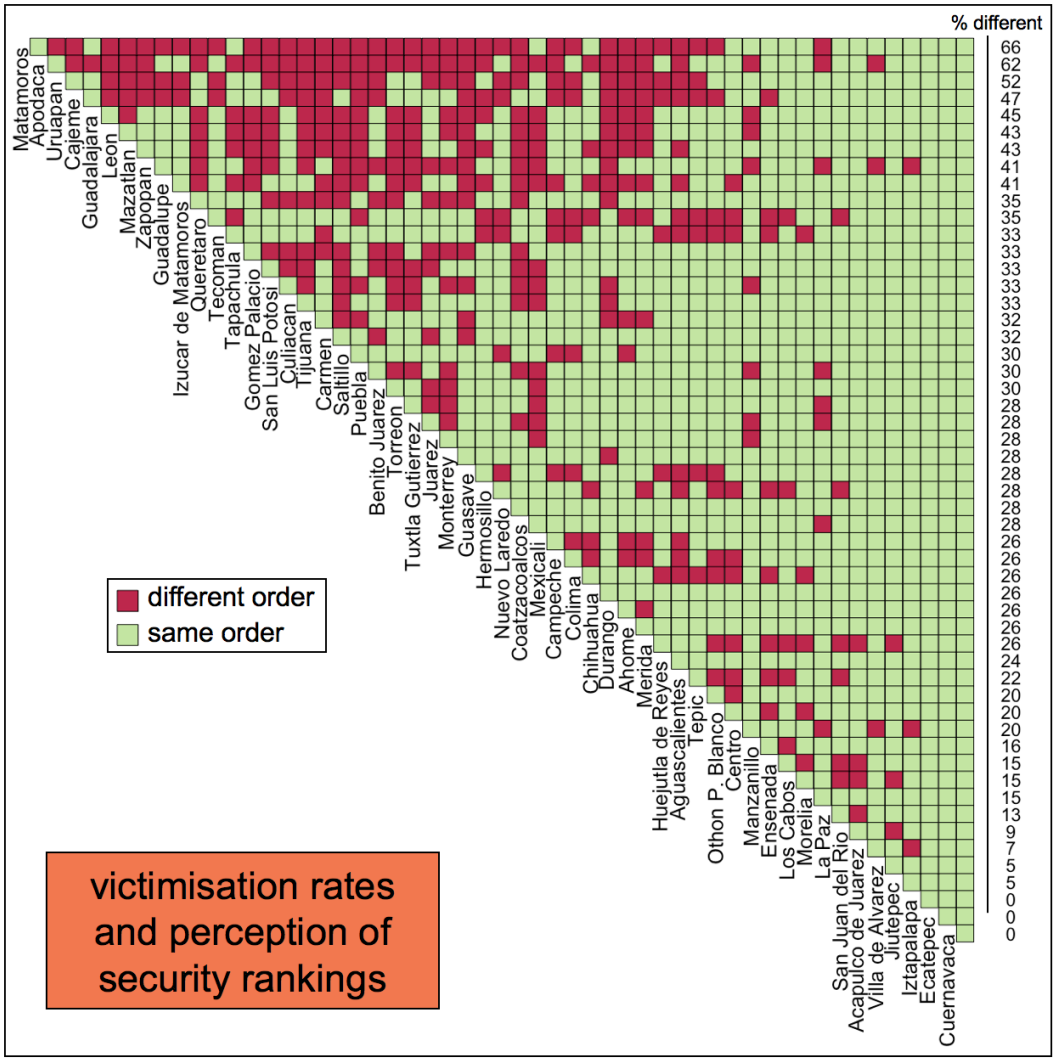

Figure 6.5: Displayed are the 1,653 comparisons that were made from the selected 53 counties in Mexico, using a red mark to highlight a comparison that does not preserve order, meaning that places with a higher victimisation of robbery of a person are nonetheless considered more secure.

is strong evidence to counteract this argument. A null hypothesis of no correlation between the rankings S and $V^{(1)}$ can be considered, in which the metric $P(S,V^{(1)})$ has an expected value of zero. For a very small number of elements in the rankings, it is possible to compute the exact distribution and obtain a confidence interval for the expected value of $P(S,V^{(1)})$ under that hypothesis, whilst for a large number of elements, the variance can be

approximated (Kendall, 1948) by

$$Var(P) = \frac{2(2n+5)}{9n(n-1)}. \tag{6.2}$$

However, 53 counties is not small enough to compute the exact distribution, nor is it large enough to trust the approximated variance. Instead, two random variables with a length of 53 cases each are simulated and both rankings are computed. The result is not usually an ordered structure, compared to the one observed in Figure 6.4, but is more similar to a completely disordered graph, with tangled lines, as the one displayed on the left-hand side of Figure 6.6.

From the 53 different counties, the value of the $P(S,V^{(1)}) = 0.44$, but, is that value enough to reject the null hypothesis that the perception of security ranking and the victimisation rate ranking are not related? A simulation of 2,000 random rankings and its 95% interval helps rejecting the null hypothesis and therefore, these variables are indeed related (Figure 6.7).

The Ranking Metric might also be used to track the changes in the perception of security over different time periods. Let S^{2013} be the perception of security ranking for the year 2013 and S^{2014} for 2014, and measure $P(S^{2013},S^{2014})$ which gives the ranking metric of the perception of security for the two consecutive years. Similarly if the $P(V^{2013},V^{2014})$ between two victimisation rate rankings is considered. Results for the ranking metric between the perception of security and victimisation rates for different years is displayed in Table 6.3.

The perception of security ranking tends to be more closely related between any two consecutive years, as expressed in Table 6.3, but this is not always the case since, for example, the S^{2014} ranking is more similar to the S^{2011} than it is to the S^{2012}. In general, the perception of security is set for a considerably large amount of time, and so a region that in the past was perceived as being secure (insecure) has a tendency of being perceived secure (insecure) in subsequent years. There is, after all, a memory in the system.

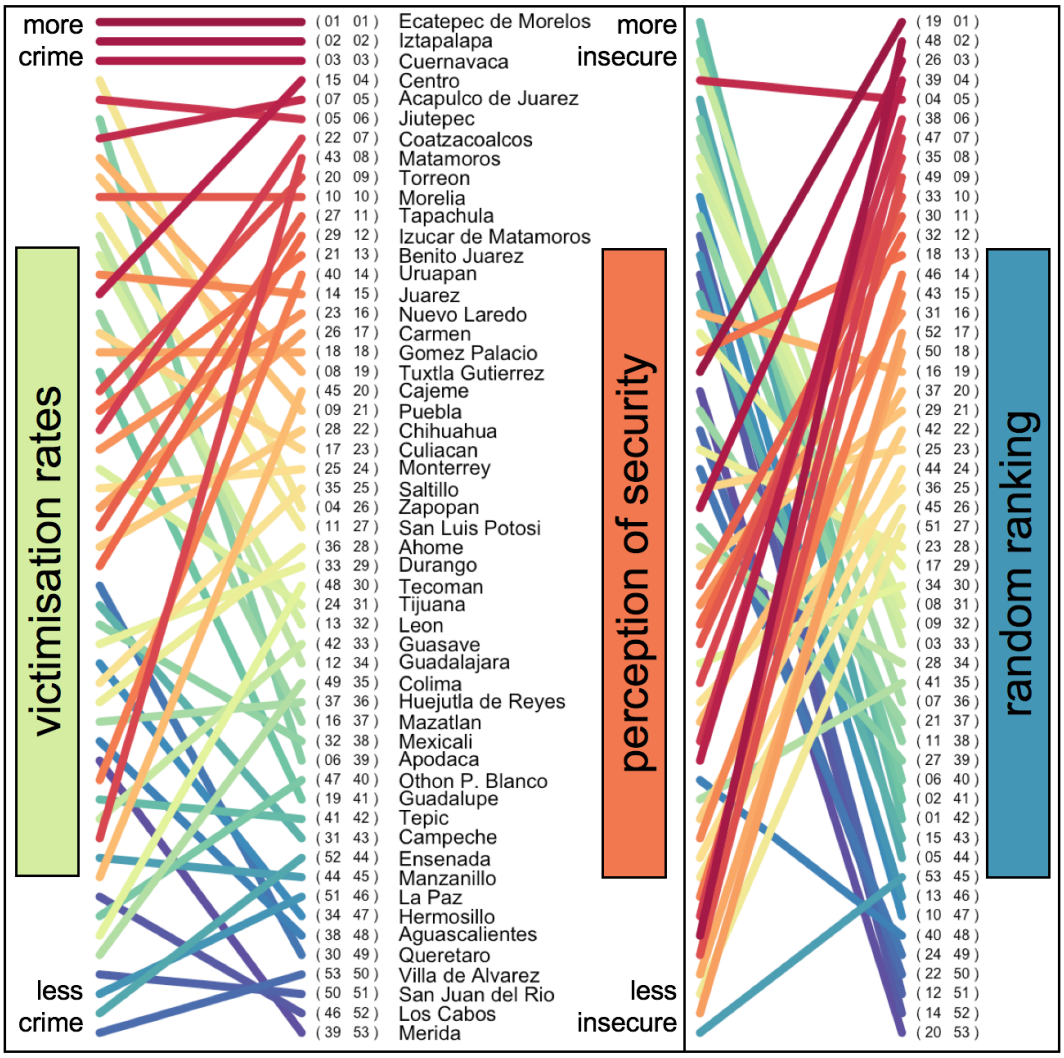

Figure 6.6: Comparison of random variables with no relationship between each pairing on the right-hand side, and the relationship between robbery of a person and perception of security obtained through victimisation surveys, on the left-hand side. As observed, random rankings create a more tangled pattern.

The victimisation rankings follow a different pattern, for example, the ranking of the year 2011 is more similar to the 2014 ranking than to the 2013 ranking. There are some counties with low (high) victimisation rate in 2011 which had a higher (lower) victimisation rate for the year 2013 but then went back in the year 2014.

The perception of security ranking has a small degree of variability between consecutive years, (Figure 6.8), particularly for the counties which are considered secure. Two cases are Ciudad Juárez and Chihuahua, both

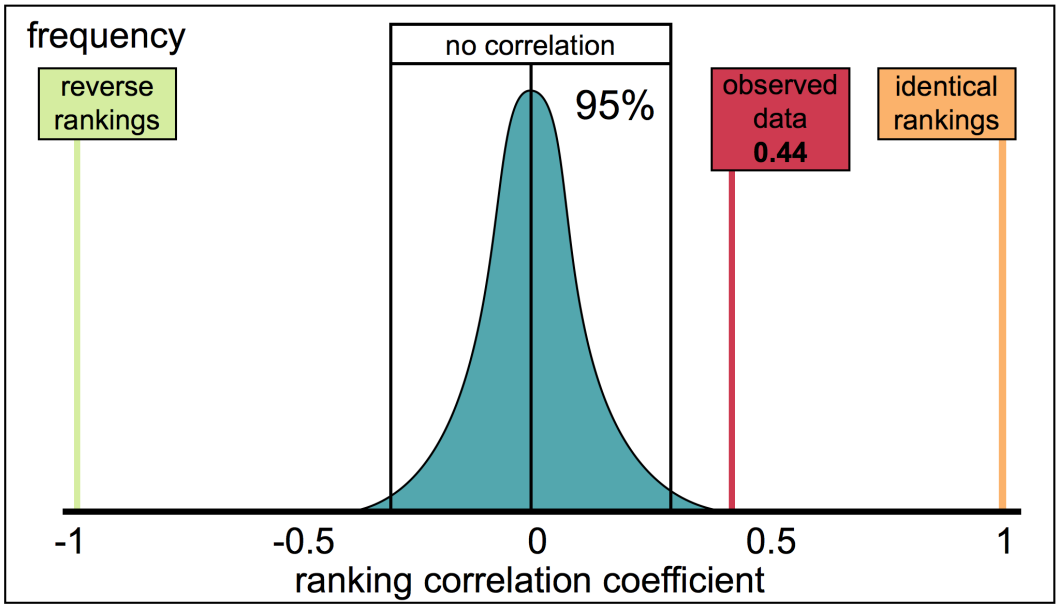

Figure 6.7: From the 2,000 random rankings, 95% were found to have a ranking metric that lies between -0.182 and 0.182, where any value above that interval, as in the observations from the counties in Mexico, is interpreted as a clear relationship between the two rankings. Observed values below the (-0.182, 0.182) interval obtained through the simulation are also considered as a clear relationship between the rankings, only having a reversed order (which is what would be obtained if instead of ranking counties from the one with the most victimisation, the ranking started from that one with the least victimisation).

	perception of security ranking S			victimisation ranking V		
year	2012	2013	2014	2012	2013	2014
2011	0.6286	0.5604	0.5240	0.6604	0.5806	0.6546
2012	–	0.5720	0.5008	–	0.5864	0.6430
2013	–	–	0.6996	–	–	0.6648

Table 6.3: Ranking metric between the perception of security on the left, and the victimisation rates on the right, between 2011 and 2014. A value closer to 1 means that the rankings have a higher degree of correlation, a value closer to 0 means no correlation between the rankings, and a value closer to -1 means that the rankings provide a reverse order.

located in the state of Chihuahua, which occupied the 1st and 8th place as the most insecure counties in 2011 and they are located in the 2014 survey in the 15th and 22nd place respectively. The perception of security from these two counties has improved considerably as compared to the rest of the counties considered.

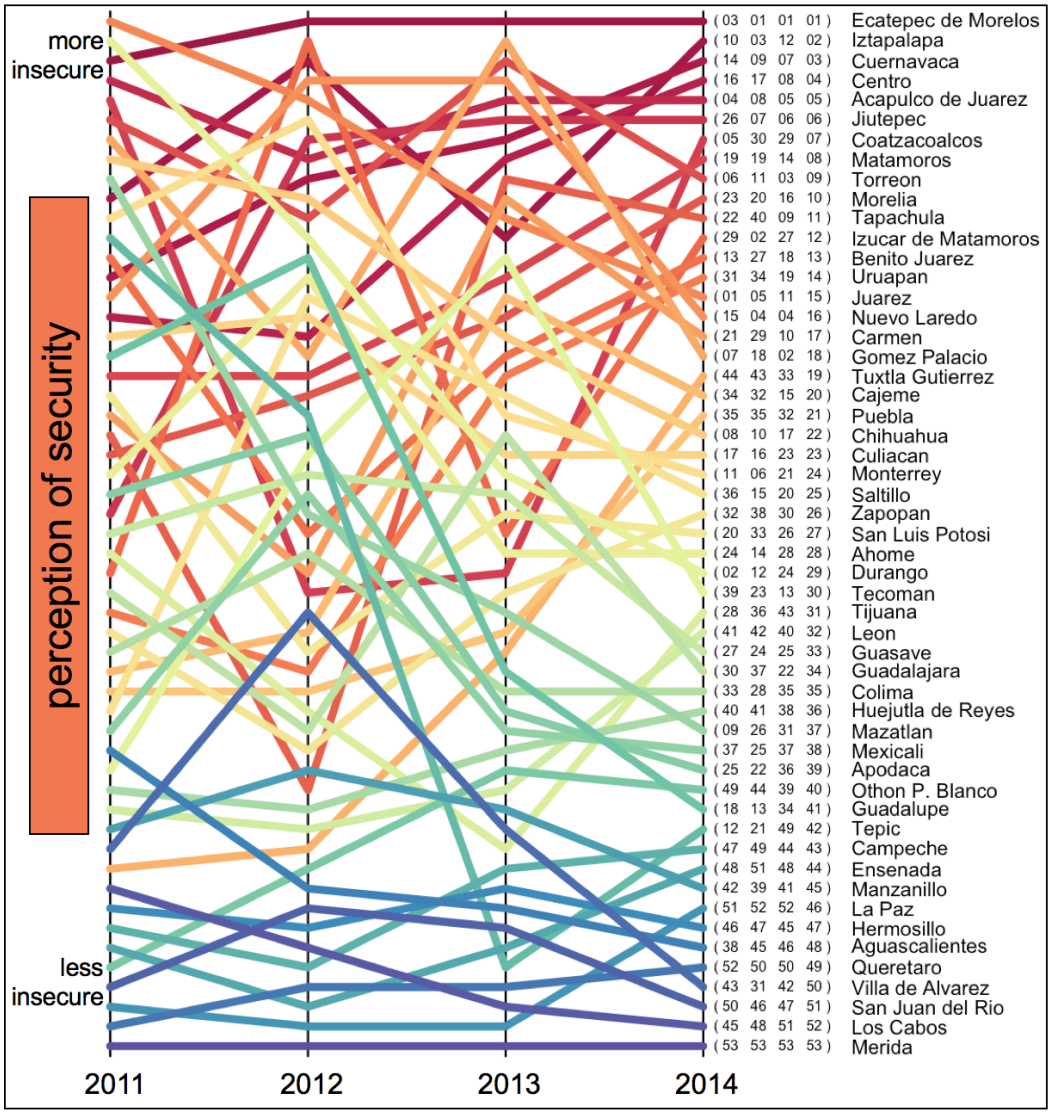

Figure 6.8: The perception of security ranking S of the 53 counties being considered in Mexico from 2011 to 2014. In brackets is the four consecutive rankings from that region. It shows that it is common for a region to have similar rankings in consecutive years.

The ranking metric $P(S,V^{(i)})$ can be computed using data for different types of crime and, since the ranking obtained by sorting the counties based on the victimisation rate of robbery of a person is different to the ranking of the other types of crime, different results for the $P(S,V^{(i)})$ metric are obtained. The ranking metric $P(S,V^{(i)})$ obtained by ranking the different types of crime alone and the results are displayed in Table 6.4.

At first glance, the results in Table 6.4 show that the ranking of most

crime	ranking metric $P(S,V^{(i)})$
robbery of a person	0.4368*
car theft	0.1292
partial cat theft	-0.0522
burglary	0.0538
vandalism	-0.0290
kidnap	0.2510*
murder	0.1002
missing person	0.1902*

Table 6.4: Ranking metric between the perception of security and the different types of crime. A value closer to 1 means that the rankings have a higher degree of correlation, and a value closer to 0 means no correlation. Values that are statistically different to zero are marked with an asterisk (*).

of the types of crime is statistically not related to the perception of security ranking, since again, most of them lie between the $(-0.18, 0.18)$ rejection interval obtained through simulation. Ranking the 53 counties based on their Murder rates (from the ones with the highest rate to the ones with the smallest rate) and comparing that ranking to the perception of security ranking, they display no relation. Murder is a rare event, so the counties usually have a rate so close to zero that the murder ranking becomes almost irrelevant.

There are some counties, such as Uruapan and Matamoros, in which the perception of security does not match the victimisation rates, and they tend to be the ones with the highest number of comparisons with different order between the victimisation ranking and the perception of security ranking (Figure 6.5). This means that these counties may have particular situations, perhaps, such as organised crime, or a high victimisation background.

There are, on the other hand, counties such as Apodaca and Guadalupe, both in the state of Nuevo León, as well as Guadalajara and Zapopan in the state of Jalisco, as well as Mazatlán in the state of Sinaloa, which are perceived as being more secure than expected based on their victimisation rates. These four counties were perceived as much less se-

cure a few years ago (Figure 6.8), particularly Mazatlán, which was in the 10th place as the least secure county in 2011 but in 2014 occupies the place 37th. It is possible that these highlighted cases might be counties where organised crime and other types of crime such as extortion have had an impact on the society and its perception of security, without crime itself being reflected in the eight victimisation rates considered in the model.

6.4 Remarks

The methodology presented constructs a regional metric for the perception of security which can be easily interpreted as the probability that a person perceives that region to be insecure. Even when the perception of security is based on the impressions and fears from a surveyed population, with these feelings representing a generalised behaviour in the population, the perception of security might be considered an attribute of the region rather than the impressions and fears of some of its individuals, hence the validity of a regional approach to the perception of security.

A regional metric for the fear of crime was constructed by considering the average opinion of individuals. This fear of crime metric can be compared between different regions and the relationship between different types of crime and the metric for the fear established.

6.4.1 Why rankings?

Similar results, comparing the regional fear of crime and the crime rates, could be obtained by using traditional association metrics, such as the Pearson's correlation coefficient. However, quantifying fear of crime, is perhaps, one of the most significant challenges in terms of perceptions and so, a novel technique by considering the rankings of the fear of crime S rather than the average fear, opens the possibility to use other types of data for its analysis.

Three examples of different types of data which are often encountered in the analysis of the fear of crime. Firstly, a person is asked to decide

between two locations, R_i and R_j, which they consider to be more secure. Secondly, a person is requested to rank between the whole set of regions R_i from the one they consider to secure to the one they think is less secure. And thirdly, a person is required to pick the top 5 or m regions which they consider insecure. In the three examples, instead of a regional fear, collating the answers between different individuals would result more naturally in a ranking than on a continuous variable.

The analysis of fear of crime depends strongly on the available data and so, having tools for dealing with a broader variety of data structures opens the field for more specialised research.

Finally, rankings are often a powerful tool for communication purposes. It is often difficult to compare too many regions at the same time or to remember the specific areas which are more or less secure, but for example, the ranking of the fear and crime (Figure 6.4) shows that the three regions with the highest amount of crime are also the three regions with the highest fear. Rankings often carry a simple message, particularly in the top and bottom places.

6.4.2 Ranking the fear of crime and victimisation rates

Quantitatively speaking, although the perception of security by itself reveals a pattern of beliefs, the absolute measure is much more valuable when compared to other regions or over different time periods since the significance of the isolated number cannot be easily determined. The regional perception of security differentiates between two regions based on their perceptions. Ranking the perception of security gives a valuable insight on the feelings of the whole population and it allows us comparing many regions simultaneously. This approach is not constrained to Mexico or to the geographic level of counties and thus other regions could also be ranked by their perception of security.

One of the results which emerge from quantifying the fear of crime is that there is only a slight relationship with different types of crime suffered

in each region. Data shows that fear of crime is much more complex than assuming that a reduction in the levels of crime on a region will improve its perception, or that the region which has the lowest levels of crime should also be the region with the lowest levels of fear. This complex pattern will be explored in the following chapters.

6.4.3 Individual fear of crime is dynamic

A person's perception of security is affected by past circumstances which, in terms of policy design, shows that events, such as a kidnap, are covered extensively in the media, for example, might have an immediate negative impact and this influence may continue for some time so that improvements in the perception of security tend to be quite slow. Therefore, even if a region is suddenly successful at reducing its victimisation rates, it might take a long time for the perception to improve. However, a decline in the perception of security might happen rapidly, as seen in the fear of crime in a county in Mexico (Coatzacoalcos, Veracruz) found in the 30th place in 2013 ranking of the most insecure counties in Mexico, to the 7th place of in 2014, which clearly indicates a new concern in that county.

The perception of security has a high ranking metric with robbery of a person, meaning that counties in which robbery of a person is more frequent tend to be perceived as being less secure. Robbery of a person is the second most frequent type of crime in Mexico and it is a type of crime in which the victim and the criminal have some form of contact, at least for a few seconds, and perhaps not surprisingly, this type of crime then has the highest ranking correlation. The other two types of crime with a high ranking metric are missing person and kidnap, but these two types of crime are relevant due to their high social impact rather than their frequency. In Mexico during 2014, for every 119 robberies of a person, there was a single kidnap. The fact that partial car theft and vandalism have a low ranking correlation with the perception of security, taking into account that they are the first and third most frequent type of crime respectively, shows that low impact crime

has, in fact, a small impact on the perception of security, especially if the population has more relevant (although perhaps much less frequent) crimes to worry about.

6.4.4 Policy implications of quantifying the fear of crime

From the point of view of policy design, it is relevant to highlight that the perception of security might be quantitatively measured as the average perception of security from the population of a region, such as a county. This measure varies from one place to the other and changes over time due to multiple factors.

Results here highlight that efforts invested in reducing the levels of lower impact crimes (such as vandalism or partial car theft) might not actually improve the perception of security, even when they are the most frequent types of crime. However, a policy-oriented to reduce the levels of robbery a person might have much better results in terms of its effect on the perception of security since it has a strong impact on the perception of security accompanied with a relatively high frequency.

Chapter 7

Individual fear of crime

The previous chapter analysed the fear of crime and its relationship with the crime suffered at a regional level. Results showed that there is only a slight relation between different types of crime and fear and that, in general, a region with a higher number of crimes is not necessarily the region with the highest fear.

Rather than analysing the regions as the units of observation of the perception of security, this chapter presents a model of the fear from an individual perspective and detects the emergence of fear as a collective behaviour. A mathematical model is used to mimic different factors which affect the individual fear of crime, and it helps to detect the dynamics of fear and the reasons why crime is not strictly related to its fear. It is based on published research (Prieto Curiel and Bishop, 2017, 2018).

7.1 Individuals and their fear

Many studies have tried to understand and quantify the factors that affect a person's fear of crime (Grogger and Weatherford, 1995). The reasons why a person fears crime in one region more than another may be based on their past experiences in that region, perceived street disorders (Lewis and Maxfield, 1980) demographic factors (Kershaw and Tseloni, 2005) and because of the specific people within the local community (Tseloni, 2000),

among many other factors, but is this fear merely in direct relation to the actual crime?

Fear of crime is the result, indeed, of suffering a crime (Tseloni, 2007; Hale et al., 1994), but the majority of the population does not suffer any crime (Prieto Curiel and Bishop, 2016b) and therefore, fear of crime is the result of a much more complex social dynamics, which involves crime, to a certain extent, but other factors, which might depend on the specific person (age or gender, for instance), on the region (if it is a dark street) and perhaps the media coverage of crime (Chadee and Ditton, 2005), having the wrong facts about crime and how fear is shared with others. Whilst past victimisation increases the probability that a person actually fears crime (Hale et al., 1994), being the victim of a crime does not entirely explain the generalised fear of crime (Skogan and Maxfield, 1981) and regions are frequently perceived as being insecure even when they suffer a relatively low crime rate (Prieto Curiel and Bishop, 2016b).

7.1.1 Modelling individual fear of crime as an opinion

Most studies about the fear of crime and the perception of security are static observations of the current situation in a particular region (Jackson and Gray, 2010), country (Kershaw and Tseloni, 2005) or group of countries, by analysing the results of victimisation surveys, based on different types of questions about the fear of the individual (Hale, 1996). Elsewhere, detecting those individuals who actually suffered a crime, has allowed the impact of direct victimisation to be measured (Skogan, 1987). However, little is known about how the collective perception of insecurity emerges, how it changes over time, what is the impact of a crime on the perception of the victims and most importantly, what is the impact of crime on the perception of security of the many nonvictims. As it was noted in previous chapters, crime is a rare event and so most of the individuals do not suffer crime, but indeed have fear of crime. Thus, there is a miss-match between crime and its fear (Skogan, 1987) and, even when crime rates have dropped considerably in many

countries in recent years (Pease and Ignatans, 2016; Farrell et al., 2011), the fear of crime has not experienced the same drop.

From the point of view of the perception itself, one approach is to consider it as an opinion. A variety of conceptual models already exist which analyse potential opinion dynamics: how the interactions between people lead to the emergence of a global consensus (Castellano et al., 2009), what is the opinion volatility (Kacperski and Holyst, 1999), what is the role of the social network on the dynamics of opinions (Düring et al., 2009), what is the impact of extreme opinions (Deffuant et al., 2000), or how does the opinion of a leadership affect its dynamics (Düring and Wolfram, 2015). Mathematically, a wide variety of models have been used in the analysis of opinion dynamics, which go from techniques used in epidemiology (Bettencourt et al., 2006), kinetic models to determine distribution of opinions over time (Toscani, 2006), models based on mean field theory, where the impact of all the individuals is simplified into a single averaged effect (Kacperski and Holyst, 1999; Banisch, 2014; Banisch and Lima, 2015), and simulating agents (Deffuant et al., 2000; Hegselmann and Krause, 2002). These models have been applied in a variety of settings, such as the behaviour of voters (Düring et al., 2009), the implementation of a specific tool by a scientific community (Bettencourt et al., 2006), political segregation in the United States (Düring and Wolfram, 2015) and for modelling the spread of misinformation and fake news on the internet (Del Vicario et al., 2016).

The main interest generated by opinion models is how an idea is shared among individuals, how they reach a consensus or, under certain circumstances, how polarisation or fragmentation of opinions is observed and how does the persuasiveness, assertiveness and supportiveness of different individuals change the dynamics. These models placed emphasis on the interaction between individuals and the spread of their ideas, but external factors, which might affect their opinion strongly, are usually ignored or modelled as random noise, sometimes referred to as a process of “self-thinking”.

Connecting these two fields, the analysis of the fear of crime with the study of opinion dynamics, is not straightforward since crime cannot be ignored or modelled as random noise. Thus, it is vital to understand the external factors which affect the perception of security and hence determine its dynamics.

A model is proposed to quantify the dynamics of the perception of security and then simulate the dynamics under a variety of scenarios to mimic different circumstances that are observed in terms of crime and its fear.

7.2 Mathematical model of the fear of crime

Let us suppose that the perception of security of a fixed region, k say (such as a city or a county) is given by a number s_k, between 0 and 1, where 1 means the perception is that the region is the most insecure and 0 is when the place is the most secure. We use a continuous approach for the perception of security to quantify different levels in which a person might fear crime, and a simple way to interpret s_k is that it represents the probability that the person considers the region to be insecure and $(1-s_k)$ is the probability that considers it to be secure, so a larger value of s_k means that it is more likely that he or she considers that particular region to be insecure. This means that if $s_k > s_j$ then the person k considers the same place to be more insecure than the person j.

A common practice when surveying individuals about their perception of security, is to consider binary variables, so individuals can either chose between the answers of “secure” or “insecure”, as in the Mexican Victimisation Survey (from INEGI, 2014), or to provide a fear scale, as in the Crime Survey for England & Wales (Office for National Statistics, 2016). The perception of security of the whole population $k = 1,2,\ldots,n$ or any subgroup, might be summarised by the mean perception of security, S, which is also the expected value of the Poisson Binomial distribution when asking each individual a binary question about their fear or considering a specific thresh-

old for the fear scale.

The perception of security of a person may change over time and to represent this, write $s_k^{(t)}$ and $s_k^{(t+1)}$ for the perception of security at two consecutive time intervals, for instance, from one week to the next. The model considers discrete units of time, defining a discrete dynamical system (Galor, 2007; Sandefur, 1993). Usually, the length of this step will be given naturally by data. In crime studies, a commonly used time step is weekly periods, although corroboration with data may only be available at yearly intervals.

Three reasons why the perception of security of a person might change from one time step to the next are considered: memory loss, suffering a crime and due to the opinion of others. It is assumed that the three reasons (memory loss, suffering a crime and exchanging opinion with others) are observed in discrete units of time which represent intervals of one week. Thus, the perception of security of k is updated according to the general equation

$$s_k^{(t+1)} = f(s_k^{(t)}, \psi_k^{(t)}, I_k^{(t)}, \underline{s}^{(t)}), \tag{7.1}$$

where f is a function which represents the dependence of the updated perception of security of k at the next time step on its current perception of security $s_k^{(t)}$, its memory $\psi_k^{(t)}$, whether he or she suffered a crime in the corresponding time interval $(t, t+1)$, expressed as $I_k^{(t)}$ and the impact of the perception of others at that time, $\left\{s_1^{(t)}, \ldots, s_n^{(t)}\right\}$, written as a vector $\underline{s}^{(t)}$.

Different scenarios which alter the perception of security of k are analysed, and four different outcomes from each time step are considered. Thus, the role of the function f is to update the perception of security of k according to the four distinct scenarios, schematically depicted in Figure 7.1.

7.2.1 Memory of past perception

If the person k does not suffer a crime between t and $t+1$ and everything else remains fixed, then it is assumed that the person will consider that

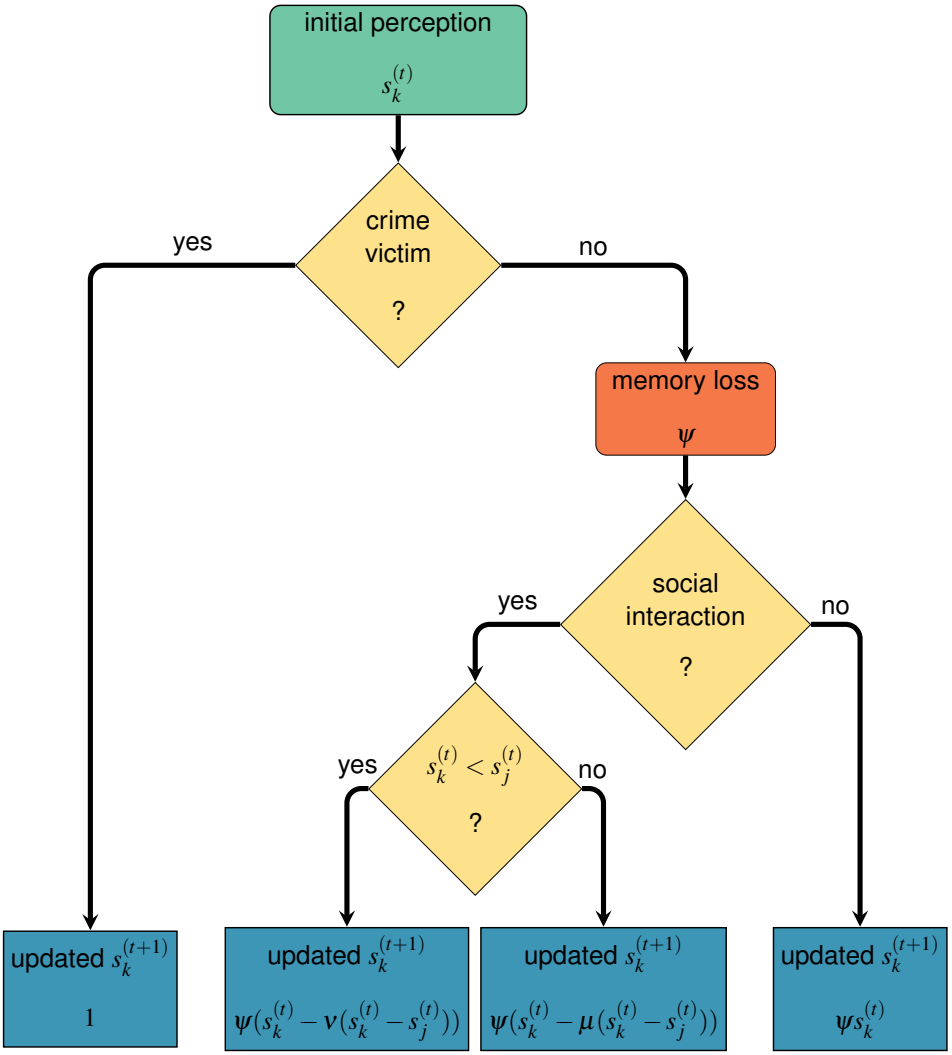

Figure 7.1: Schematic model in which the perception of security of the person k is updated from one week to the next one. There are four possible outcomes depending on whether the person suffered any crime, depending also if the person had an interaction and if that interaction occurred with a more or less fearful person.

security is improving, that is, the memory of past perception is gradually lost. Thus, by isolating the effect of memory from any other factor, it is considered that

$$\text{Impact of memory:} \quad {}_{memory}s_k^{(t+1)} = \psi_k s_k^{(t)} \tag{7.2}$$

with $\psi_k \in (0,1)$, which represents the (constant in time) speed at which the person k has a loss of memory, where ψ_k closer to 1 means that the

perception of security remains the same at the next step, and a value of ψ_k closer to 0 means that the person forgets their past impressions quickly. This expression is referred to as having *exponential decay*, and so, with no other factors, the perception of security will always decay as time goes by.

It is assumed that all individuals have the same memory loss, and that memory loss is the same for all types of crime, with rate ψ, although different speeds at which individuals tend to forget their past perception, based perhaps on the type of crime, could be considered.

7.2.2 Crime

Crime has certain properties which make it hard to analyse and to understand its impact on the perception of security. Firstly, crime is a rare event, and therefore the majority of the population does not actually suffer any crime. Moreover, many of the crimes end up being an attempt (Skogan and Maxfield, 1981), and for some victims, the impact of suffering a crime might decay rapidly over a few weeks or months and might also be of limited consequences (Skogan, 1987). Finally, crime is highly concentrated in certain population groups with some victims tending to suffer more than one crime (Farrell et al., 2005; Grove et al., 2012). Looking, however, at the individual impact of suffering an actual crime, it is assumed that being the victim of a crime causes fear.

If the person k suffers a crime between t and $t+1$, then it is considered that the person will naturally think that the place is insecure. More formally, let $X_k^{(t)}$ be the number of crimes suffered by the person k between t and $t+1$. There is the possibility, albeit small, that a person might suffer more than one crime during one time step, and so $X_k^{(t)}$ could potentially have large values; therefore, a binary variable I_k is used, such that $I_k^{(t)} = 1$ if the person suffered at least one crime between times t and $t+1$ and 0 otherwise. Then if $X_k^{(t)} > 0$, assume that $s_k^{(t+1)} = 1$ regardless of any previous perception.

The impact of crime, then, is expressed as

$$\text{Impact of crime:} \qquad {}_{crime}s_k^{(t+1)} = I_k^{(t)} + (1 - I_k^{(t)})s_k^{(t)}. \tag{7.3}$$

The risk of suffering a crime is not uniformly distributed across the population (Johnson, 2010b; Tseloni and Pease, 2005) and since the objective is to understand the dynamics of the perception of security and how crime affects it, the different distributions of crime across the population need to be considered. Because of the social environment, activities, age, gender and other reasons, crime is more concentrated in some sectors of the population (Hope and Trickett, 2008) and in some specific regions (Freeman, 1996) giving rise to criminal hotspots (Brantingham and Brantingham, 2010). As it was analysed earlier on this thesis, some groups of people may be "immune" to crime and others may suffer chronic victimisation (Hope and Trickett, 2008). A distribution of crime which takes into account this inhomogeneous behaviour is considered (Prieto Curiel et al., 2017a).

Assuming that the number of crimes suffered by a person is independent of others and that suffering a crime does not affect the probability of suffering subsequent crimes, then the number of crimes suffered by the k-th person $X_k^{(t)}$ might be modelled as a Poisson distribution with rate $\lambda_k \geq 0$. These two assumptions, the independence between the crime suffered by individuals and a constant rate of suffering a crime, may be unrealistic but characterise the victimisation profile of the whole population (i.e., to consider the λ_k for each individual), to take into account an immune group (with $\lambda_k = 0$), a chronically victimised group (with λ_k large) and to control the expected number of crimes suffered by the whole population ($\sum_k \lambda_k$). The victimisation profile (Prieto Curiel and Bishop, 2016a) allow simulating the number of crimes that individuals might suffer, so it is possible to compare the perception of security among people who suffer higher or lower amounts of crime.

7.2.3 Opinion dynamics

If the person k interacts with a 'fearful' individual, then it is likely that the perception of k will be changed by the interaction. Fear is contagious and the impact of an interaction might depend on the closeness between the individuals or the strength of their ideas, or its intensity (Latané, 1981) both concepts being assessed appropriately. This situation is usually modelled as two people who have different opinions who reach a state closer to each other (Curtis and Smith, 2008), or closer to a *consensus* opinion after they update their beliefs (Toscani, 2006).

In opinion dynamics, a certain amount of random pairs of individuals might be deemed to have an interaction between them (Deffuant et al., 2000) or all individuals might update their beliefs simultaneously (Hegselmann and Krause, 2002). Regarding the fear of crime, it is considered that not all individuals have an interaction with others each week and so, only a certain proportion of the population, γ, forms pairs of individuals without replacement, and they share and update their opinions.

It is worth considering that the perception of security differs from other types of opinions, like a left or right political leaning, since the impact is not symmetric: a more fearful person might share their own experiences with others, increasing the fear of crime in them, without it reducing his or her own fear in the same magnitude. Thus, there is an opinion-dependent asymmetry (Hegselmann and Krause, 2002), which might be modelled as follows. Let $s_k^{(t)} > s_j^{(t)}$, so the person k considers the region to be more insecure than the person j. Then, isolating this effect:

Impact of social dynamics on k

$$_{interaction}s_k^{(t+1)} = s_k^{(t)} - \mu(s_k^{(t)} - s_j^{(t)}), \tag{7.4}$$

and

Impact of social dynamics on j

$$_{interaction}s_j^{(t+1)} = s_j^{(t)} - \nu(s_j^{(t)} - s_k^{(t)}), \tag{7.5}$$

where $\mu \in (0,1)$ is a parameter which might be considered to be the *resistance* of the perception of insecurity and $\nu \in (0,1)$ is the parameter for the *impact* of the perception of insecurity. Thus, it is assumed that μ is close to zero and ν is close to one, so that the person who fears crime the most retains nearly the same perception at the next time step and this fear has a large impact on the other person.

A simplification to the model could be achieved by assuming that the resistance of the perception of insecurity is negligible (with $\mu = 0$) or by considering the relative influence of the two parties so that ν and μ could be combined into a single parameter, but this simplification has its drawbacks since the parameter μ, although small, includes in the model the impact of any social support given to the victims of crime and to the persons that fears crime the most (Sacco, 1993).

With $\nu > \mu$ there is no conservation of the total perception of insecurity (Toscani, 2006), that is $s_k^{(t+1)} + s_j^{(t+1)} \geq s_k^{(t)} + s_j^{(t)}$ and there is a certain degree of compromise between individuals, so that $|s_k^{(t+1)} - s_j^{(t+1)}| \leq |s_k^{(t)} - s_j^{(t)}|$. In this way, fear of crime is considered to be a contagious process (Gilchrist et al., 1998).

After defining the microscopic level of interactions between individuals, a technique used to model the dynamics of the whole population is to consider the distribution of opinions at a certain time $P(s,t)$ and, by applying methods of kinetic theory of binary interactions, to obtain a Boltzmann-type equation (Toscani, 2006), or to consider a typical individual and analyse their perception using mean field theory (Düring et al., 2009). However, such a system is difficult to study, particularly with factors such as crime or memory, which do not depend on the social dynamics. A commonly used technique for this type of problem is to consider simulated agents (Hegselmann and

Krause, 2002), located perhaps on a lattice (Deffuant et al., 2000) or in a network, where the individuals are represented by the nodes and the edges are the potential interactions between them. In the particular case of opinion dynamics, it is common to consider the effect of a small-world network (Watts and Strogatz, 1998) or a scale-free phenomena (Barabási and Albert, 1999), however, the main interest is not the impact of the topology of the social network and therefore, random pairs of individuals are considered.

In the modelling process, individuals alter their perception of security based on: their memory loss (equation 7.2); the perception of others (equations 7.4 and 7.5); and whether individuals might or might not suffer crime (equation 7.3). This procedure is then repeated at each time step, updating the model in that specific order so that the social interactions occur after a period of memory loss and crime occurs after the social interactions which completely defines the model for the dynamics of the perception of security. The four possible outcomes from each time step are schematically depicted in Figure 7.1.

There are other factors which might play a significant role in the perception of security, for instance, a particular crime that is well reported in the media. This effect could be easily integrated into the model by adding low-frequency shocks which increase the global fear or that of a selected group, such as the readers. For now, we ignore the impact of the media, and we only consider the impact of memory loss, suffering crime and opinion dynamics in the model of the fear of crime.

7.2.4 Social interactions of individuals

People become segregated in many ways and for many reasons, for instance, age, religion, income or even the region of the city in which a group usually inhabits (Schelling, 1971) and so, the interactions between individuals from different groups are often difficult, for example, people from a run-down neighbourhood might have only a few interactions with individuals from a more upmarket community. Two distinct metrics are used for measur-

ing the degree of interactions between individuals. Firstly, their *homophily* η (Newman, 2003), defined as the proportion of times that interactions occur between a pair of individuals from the same group.

Although the homophily measures interactions between individuals from different groups, it is difficult to interpret its value in a more general case since it depends on the respective size of each group ($\bar{q}$). Based on the homophily, a second metric for the degree of interactions is constructed which is easier to interpret. Notice that the probability that two randomly selected individuals belong to the same group, π, is given by

$$\pi = \frac{\sum_{j=1}^{k} n_j^2 - n}{n(n-1)}, \tag{7.6}$$

where n_j is the size of group j.

The degree of *mixing* between groups, ϕ is defined as

$$\phi = \frac{\eta}{\pi}, \tag{7.7}$$

where a value of $\phi > 1$ means that interactions happened more frequently between members of the same group than randomness would suggest and so, a poor level of mixing between different groups exist or individuals have *discouraged mixing*. A value of $\phi < 1$ means that interactions occurred more frequently between individuals from different groups and so there is a high level of mixing between different groups or individuals have *encouraged mixing*, with the extreme case of $\phi = 0$ when individuals never interact with members of their own group. Finally, the case in which $\phi = 1$ means random mixing between different groups (Figure 7.2).

The degree of mixing ϕ gives us a comparable metric for the observed interactions between individuals from different groups.

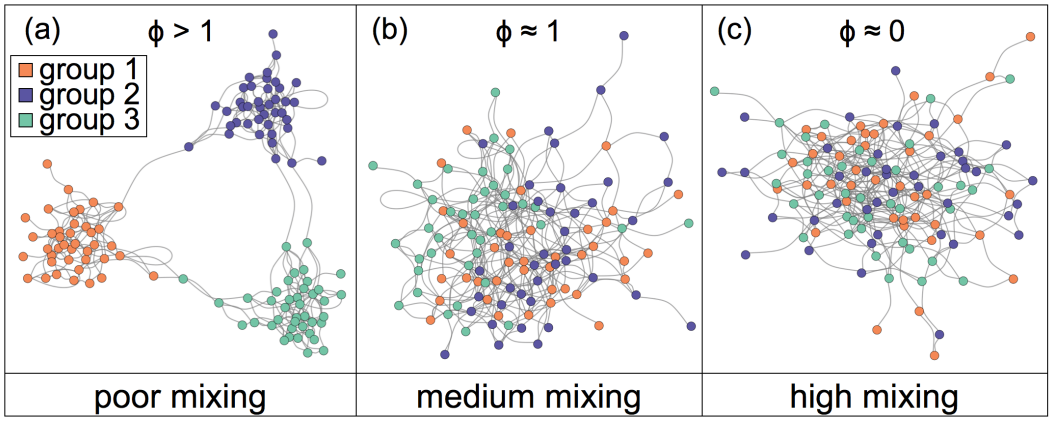

Figure 7.2: Different interactions between individuals from distinct groups. Panel (a) shows individuals who interact mostly with people from their own group (so $\phi > 1$). Panel (b) shows random interactions and panel (so $\phi \approx 1$). (c) shows the case of high mixing, in which individuals interact mostly with people from other groups (so $\phi \approx 0$).

7.3 Numerical simulations

7.3.1 Simulating crime in a population

Without crime and memory, individuals who share their opinions eventually reach a consensus (Deffuant et al., 2000), meaning that all the opinions end up being close to one another. Therefore, what is relevant in this model is the impact of these two elements, crime and memory. Crime is not suffered randomly by the population and therefore it should not be modelled simply as a homogeneous variable or noise. The distribution of crime rates, λ_k, plays a fundamental part in the model. Comparing, for instance, the mean perception of security of two populations who suffer exactly the same amount of crime but with a different distribution (Figure 7.3) reveals that results are highly dependent on how the distribution of the crime suffered is modelled.

To mimic a more realistic distribution of crime, from here onwards it is assumed that the victimisation profile can be described as a mixture model (Prieto Curiel and Bishop, 2016a). The number of crimes suffered by a

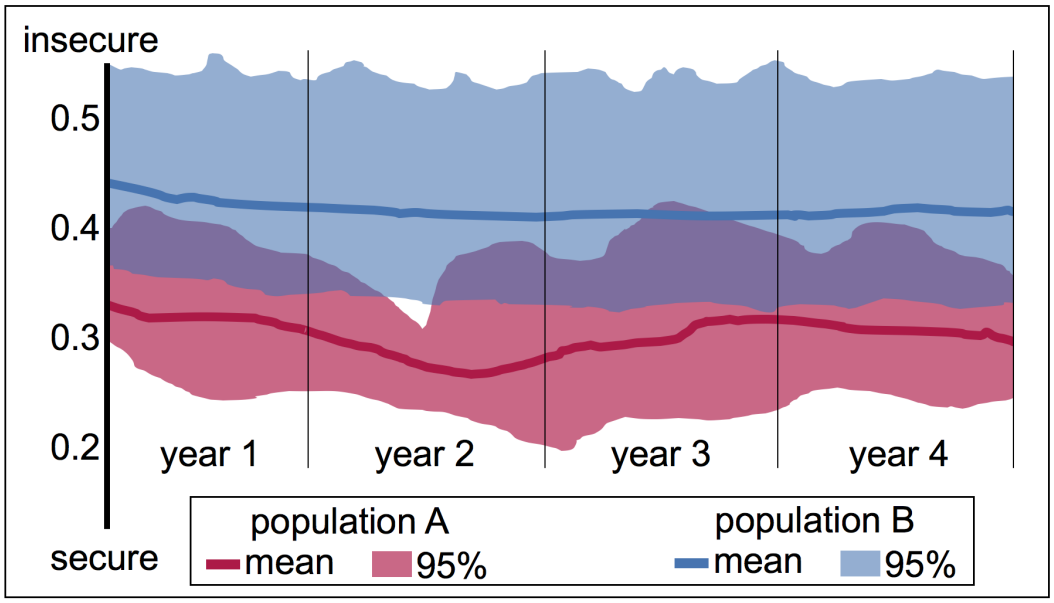

Figure 7.3: Simulated perception of security of two populations who suffer exactly one crime per week among their 10,000 individuals, with $\mu = 0.1$, $\nu = 0.9$, $\psi = 0.5$. Population A is a special case where a fixed individual suffers all the (weekly) crimes and in Population B crime is randomly suffered. On average, Population B has a mean perception of security 0.12 above Population A, even when they both suffer the same amount of crime and have the same dynamics.

random person is given by

$$q_1 \text{Pois}(\lambda_1) + q_2 \text{Pois}(\lambda_2) + \cdots + q_m \text{Pois}(\lambda_m), \tag{7.8}$$

which means that the individual is allocated into one of the m groups (with probability q_j) and then the number of crimes that he or she suffers has a Poisson distribution with the corresponding rate λ_j. Crime rates are reported here on a yearly rate due to their extreme low frequency, but the weekly rate can easily be computed.

For the numerical simulations, firstly $m = 3$ groups are used, with $\bar{q} = (0.65, 0.3, 0.05)$ and a yearly rate $\bar{\lambda} = (0, 0.05, 1.7)$ meaning that it is assumed that 65% of the population suffers no crime (Group 1), with $\lambda_1 = 0$, then 30% of the population suffers crime at a low (yearly) rate $\lambda_2 = 0.05$ (Group 2) and 5% of the population suffers crime at a higher yearly rate

of $\lambda_3 = 1.7$ (Group 3). Under this victimisation profile, the population expects to suffer 10 crimes for every 100 people each year, and 65% of the population is immune to crime. With $m = 3$ groups, with respective sizes $\bar{q} = (0.65, 0.3, 0.05)$, the probability that two randomly selected individuals belong to the same group is $\pi = 0.369 \pm 0.003$ and so, a value of the homophily $\eta = 0.631 \pm 0.003$ occurs when interactions happen randomly; higher values mean that individuals have *preferred* interactions with people from their own group (so that they interact primarily with people who suffer a similar crime rate) and lower values mean *discouraged* interactions with people from their own group (so interactions occur between people who suffer different crime rates).

A simpler model could be obtained by considering only victims and non-victims (that is, only two groups) but crime is not a simple process and evidence shows that we frequently observe different degrees in which crime is suffered, ranging from people who are statistically immune to crime (Hope and Trickett, 2008) to people who experience a small amount of crime and finally, a small population group who suffers a much higher rate (Prieto Curiel and Bishop, 2016a) so that a more realistic victimisation profile is obtained with more than two groups.

7.3.2 Impact of the opinion dynamics

The perception of security and the dynamics of a population with $n = 10,000$ individuals are simulated, who update their perception of security each week, who have memory loss (equation 7.2), might suffer a crime (equation 7.3) and might alter their opinion based on the perception of others (equations 7.4 and 7.5). From the simulations, the mean perception $S^{(t)}$ of each group is reported.

During each step, 10% of the population (1,000 individuals) are randomly selected to interact with another 10% of the population, so that during each step 1,000 distinct pairs with no replacement are made with the individuals sharing their perspective with each other. Individuals who are

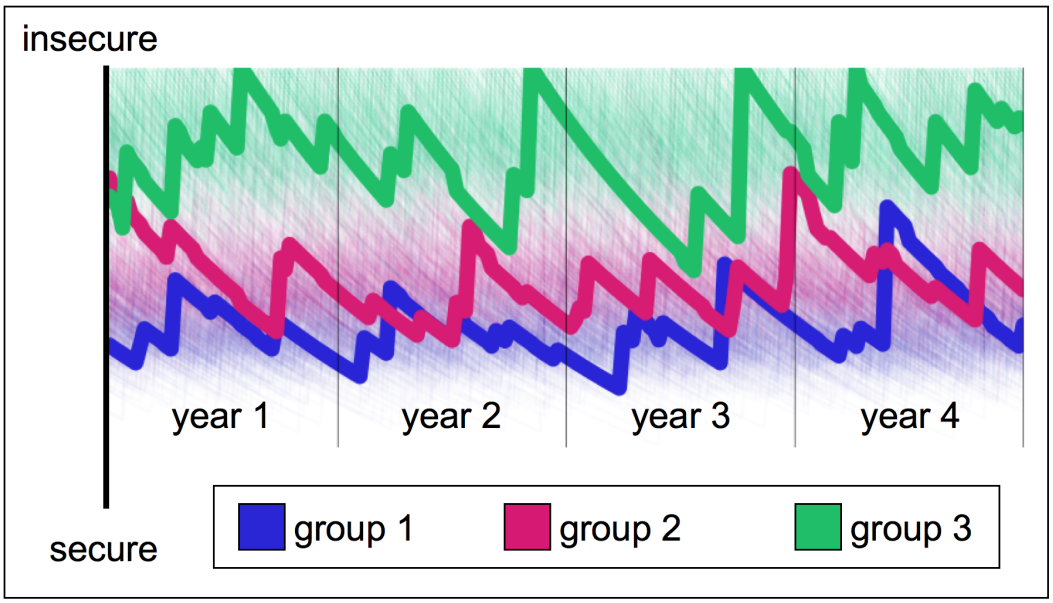

Figure 7.4: Simulated perception of security with $\mu = 0.1$, $\nu = 0.9$, $\psi = 0.5$, preferred interactions (homophily $\eta = 0.981$ so the degree of mixing $\phi = 2.658$) and crime rates $\bar{\lambda} = (0, 0.05, 1.7)$. It shows 500 individuals from each group and three representative members highlighted. The highlighted individual from Group 3 suffered four crimes (when $s_k = 1$); the other jumps are due to the social interactions; the slow decay is the memory loss.

not picked to interact with others simply update their perception of security according to the memory and crime rules. The level of interactions, thus, is $\gamma = 0.2$, meaning that 20% of the individuals have an interaction during each time step.

The simulated individuals begin with a random perception of security and the algorithm is executed for six simulated years and the first two years are discarded to reduce the impact of the initial random perception.

The three groups considered have a different distribution of their perception of security (Figure 7.4). Even the group immune to crime (Group 1) has a mean perception of security of 0.47, so nearly half of their population fears crime.

The effects of the parameters of the model are, perhaps, as expected: a value of ψ closer to 1 means that the population has more memory and therefore, the perception of insecurity remains for a longer period. A higher

value of the impact of insecurity ν increases the overall perception of insecurity, and similarly, but with the opposite effect, with the resistance of insecurity μ. Also when $\nu > \mu$, also increases the mean perception of insecurity. Most of the individuals do not suffer a crime each year (since the global crime rate is such that for every 100 individuals there are 10 crimes) which means that most of the fear of crime actually comes from social interactions. Thus, the impact of γ, which considers the frequency of social interactions, is such that more social interactions (higher value of γ) tend to increase the global fear of crime and, in some cases, only if the individual suffers extreme high crime rates, a higher level of social interactions will reduce the average fear from their group (but not the average from the whole population). Less social interactions are always related to a lower fear of crime.

7.3.3 Impact of suffering more, or less, crime

There are some surprising results from the model. Firstly, since the amount of crime suffered by the population might change, the impact of these fluctuations is measured by considering a *factor* $\kappa > 0$, and simulating a population which suffers a yearly rate $\kappa\bar{\lambda} = (0, 0.05\kappa, 1.7\kappa)$, so that the crime, even when it increases or decreases, maintains the same distribution, or the same victimisation profile (Figure 7.5). Thus, with $\kappa > 1$ crime increases and with $\kappa < 1$ it decreases.

If crime doubles in frequency (with $\kappa = 2$) or similarly, if crime drops to a half (with $\kappa = 1/2$), the mean perception of security of the whole population (from each one of the three groups) undergoes only a slight variation and in a seemingly linear manner (Figure 7.5). Roughly, an increase in κ of the crime rates increases the perception of insecurity by $0.035(\kappa - 1)$. A drastic (nonlinear) change, though, is observed when $\kappa < 0.2$, which is the threshold after which the group that suffers the highest amount of crime (Group 3) actually experiences less crime and shares this with the other groups.

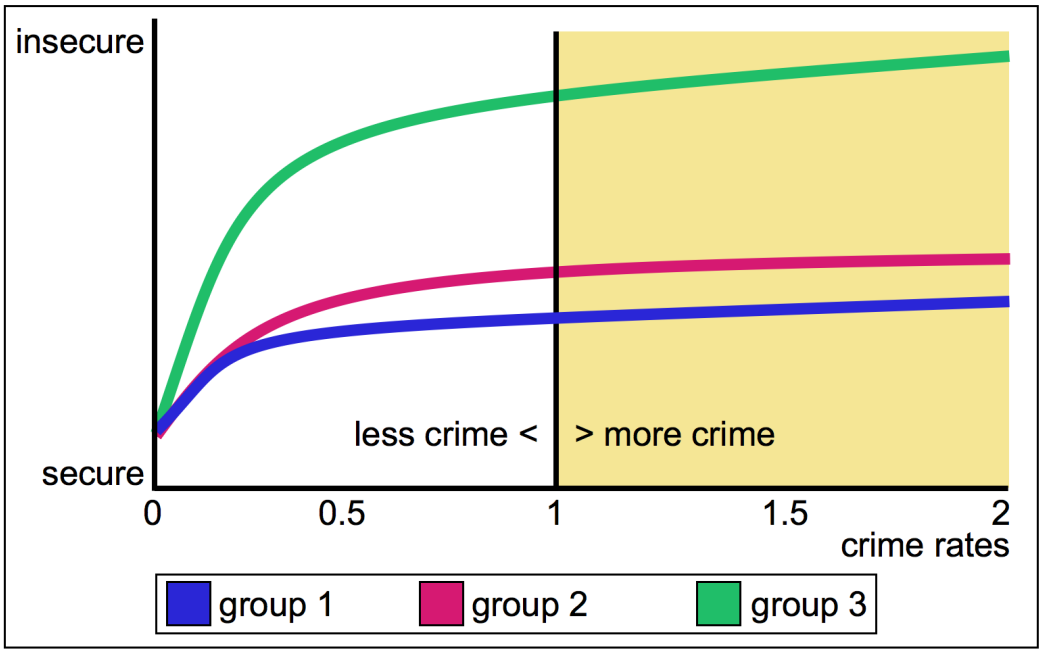

Figure 7.5: Simulated perception of security with $\mu = 0.1$, $\nu = 0.9$, $\psi = 0.5$, preferred interactions ($\eta = 0.981$ so the degree of mixing $\phi = 2.658$) and with yearly crime rates $\kappa\bar{\lambda} = (0, 0.05\kappa, 1.7\kappa)$, with κ on the horizontal axis. Crime increases with $\kappa > 1$ and decreases with $\kappa < 1$.

Thus, to improve the perception of security of a population, crime has to decrease considerably. The impact of a slight reduction in crime on the overall fear of crime is negligible.

7.3.4 Impact of having more encouraged, or discouraged, mixing

Encouraged or discouraged interactions between individuals from different groups play a key role in the dynamics of the perception of security. For people who initially are not victims of crime and do not perceive insecurity, as they interact more with individuals from other groups (that is, a lower mixing ϕ), then the perception that the region is insecure increases.

A value of the mixing $\phi = 1$ occurs when interactions occur randomly, but by modelling preferential or discouraged interactions between members of the same group, allows measuring the impact that encouraged or discouraged interactions has on their mean perception of security (Figure 7.6).

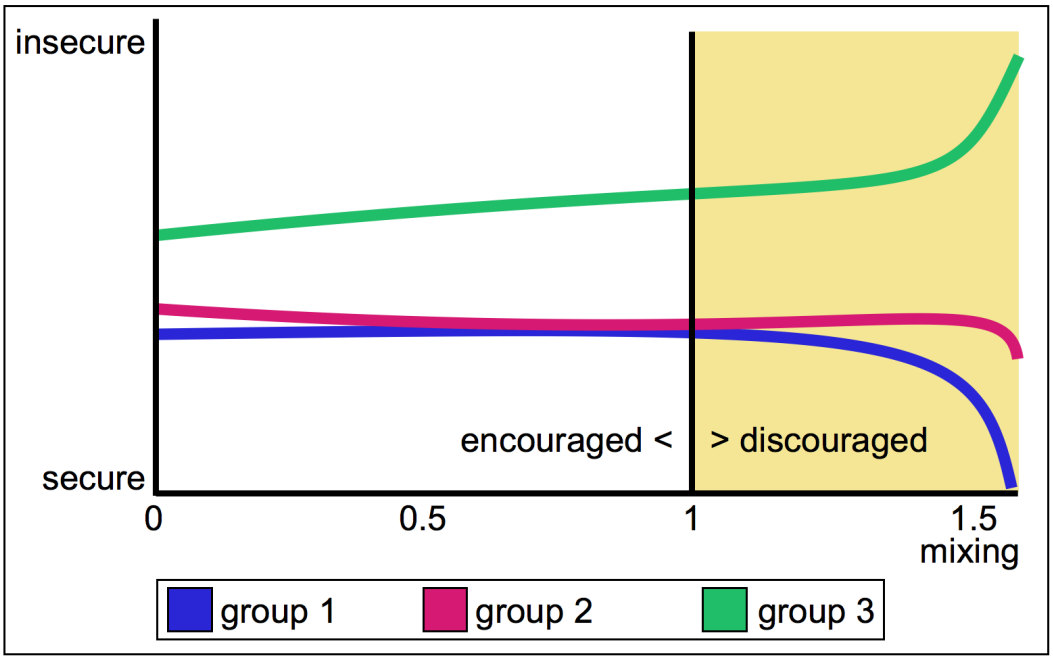

Figure 7.6: Mean perception of security of a population with $\mu = 0.1$, $\nu = 0.9$ and $\psi = 0.5$. Pairwise interactions occur between the three groups with a degree of preference or rejection to interactions with their own population group. The vertical line, where the degree of mixing $\phi = 1$, represents random interactions. With random interactions, the mean perception of Group 1 and Group 2 is nearly the same, but it changes considerably on both extremes of the mixing range.

Under fully mixed population groups (random interactions) there is a fragmentation of the mean perception of security observed among individuals who suffer zero (or close to zero) crime and individuals who suffer higher crime rates (Deffuant et al., 2000; Castellano et al., 2009).

When individuals only interact with members of their own group, each one of the groups can be analysed separately. The results show that the mean perception for Group 3 is $S_3 = 0.858 \pm 0.010$ and for Group 2 is $S_2 = 0.577 \pm 0.014$, which implies that in a population in which only 5% of their members expect to suffer one crime annually (Group 2), the perception of insecurity is already quite high (nearly 60%). Thus, even with the low frequency of crime, we observe a high level in the perception of insecurity. For the members of Group 1, who have no reason to perceive insecurity since their members do not suffer any crime and they do not obtain a perception

of insecurity from others, their mean perception decreases from their initial values to zero. However, as soon as the members of Group 1 interact with individuals of Groups 2 and 3, their perception of insecurity rapidly increases. For instance, with an homophily of 0.99 (which means that only 1% of the interactions occur between individuals of different groups) the mean perception of security of Group 1 is already $S_1 = 0.412 \pm 0.005$. Only a few interactions between individuals which belong to a different group is enough to create a fear of crime in 41% of the population of the group which will not, in fact, suffer any crime and will rarely have an interaction with an actual victim. Thus, the perception that a region is secure is quite unstable with respect to the structure of the social interactions.

7.4 Different distributions of crime and the impact of victim displacement

Comparing now the impact of different distributions of crime, populations with various sizes of the groups $(N_1, N_2, \ldots, N_k)$ together with their corresponding rates $(\lambda_1, \lambda_2, \ldots, \lambda_k)$ are taken into account. It is assumed that a group exists which does not suffer crime, referred to as the *immune* population group (Sparks, 1981; Hope and Trickett, 2008), so $\lambda_1 = 0$, where the size of this group (N_1) may vary from a small number of individuals (even zero in the limit) to a large group containing almost all individuals within the population. Also, it is assumed that the (expected) number of crimes suffered by the population is fixed, so that $C = \sum N_j \lambda_j$ is the same for every victimisation profile. A population with k groups has $2k$ parameters and three restrictions, giving $2k-3$ free parameters to determine for each of the different profiles.

7.4.1 Simulating different victimisation profiles

Populations with $k = 3$ groups and two random variables, a and b are considered, such that $0 < a < b < 1$ and the size of the group 1, $N_1 = aN$ and the

size of the group 2, $N_2 = (b-a)N$ are assigned, so that the size of group 3, $N_3 = (1-b)N$ is fixed. Although more groups could be taken into account, this would not give a more general result in terms of different profiles or the concentration of crime. For the crime rates, two further random variables α and β are considered, such that $0 < \alpha < \beta$, and the crime rates $\lambda_2 = \kappa\alpha$ and $\lambda_3 = \kappa\beta$, are assigned, where

$$\kappa = \frac{C}{N(\alpha(b-a)+\beta(1-b))} \tag{7.9}$$

is a fixed parameter which ensures that all the populations have the same (expected) number of crimes C. Under these assumptions, group 3 suffers β/α times more crime than group 2 and group 1 suffers no crime. The $RECC$ (equation 3.2) is used to measure the concentration of crime.

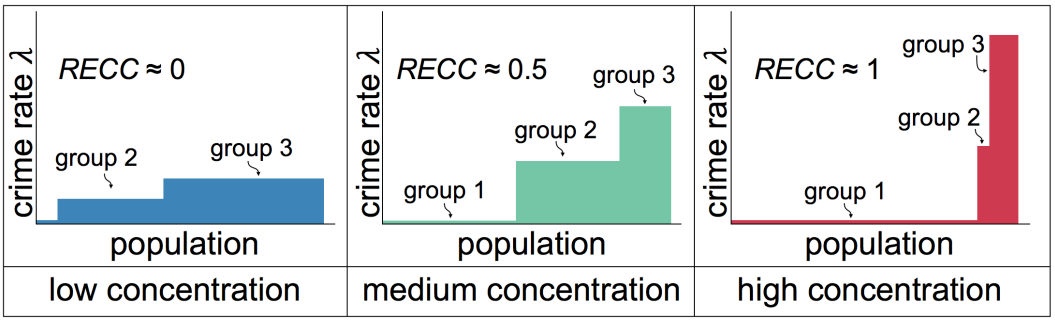

Figure 7.7: Different degrees of concentration of crime. The left panel shows a *low concentration of crime*, observed when all individuals experience a similar crime rate λ_i and therefore, the $RECC$ is close to zero. The panel on the right shows a *high concentration*, observed when a small population group suffers a high crime rate and therefore, the $RECC$ is close to one.

By taking different values of a, b, α and β, different scenarios under which crime could be distributed among the population are considered. The case of a high concentration of crime ($RECC$ close to one) is obtained either as the result of a large size for the immune population (large a) or a small size of the most victimised group ($(1-b)$ small) with a large crime rate (large β). The case of a small concentration of crime ($RECC$ close to zero) is obtained either as a large size of the most victimised group ($(1-b)$ large)

and other scenarios (Figure 7.7).

7.4.2 Simulating different crime dynamics

Simulating 1,000 different victimisation profiles and then, for each, establishing the subsequent dynamics, allows determining the impact that a higher or lower concentration of crime has on the mean fear of crime within a population.

By considering random interactions between individuals from different groups, results show the dynamical behaviour falls into one of two different phases. For low concentrations of crime ($RECC < 0.5$) the impact of different degrees of concentration of crime is negligible. There is a phase transition at high degrees of concentration of crime after which, a slight increase in the concentration has a considerable decrease in the general fear of crime.

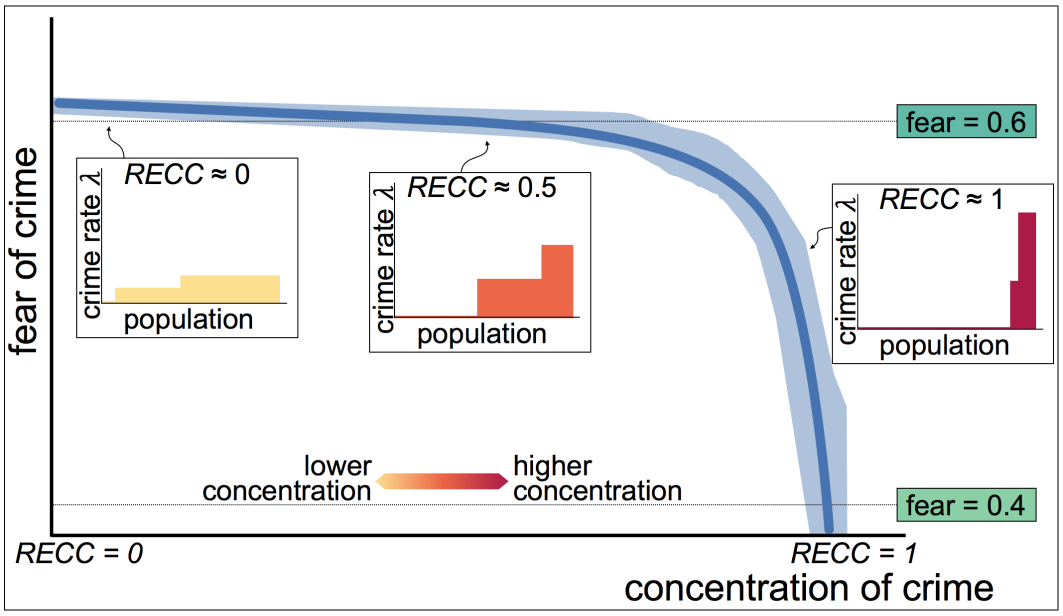

Figure 7.8: Impact of the concentration of crime on the fear of crime. Interactions between individuals from different groups are slightly less frequent than in a random model (with an average $\phi = 0.8$). The mean fear of crime for different victimisation profiles (solid blue line) and the 95% (in light blue) as the degree of concentration of crime varies. For a small or medium level for the concentration of crime, the impact of different victimisation profiles and therefore, the impact of the concentration of crime, is negligible.

There is a phase transition (observed for values of the $RECC$ between 0.85 and 0.9) for which the impact of a higher or lower concentration of crime becomes highly relevant. For high levels of concentration of crime, an even higher concentration reduces the mean fear of crime of the population (Figure 7.8).

7.4.3 Interactions between different groups

The degree of mixing between individuals from different groups changes the impact of the concentration of crime (Figure 7.9). The victimisation profile affects the mean fear of crime of the population, but the impact changes according to the level of interactions between groups.

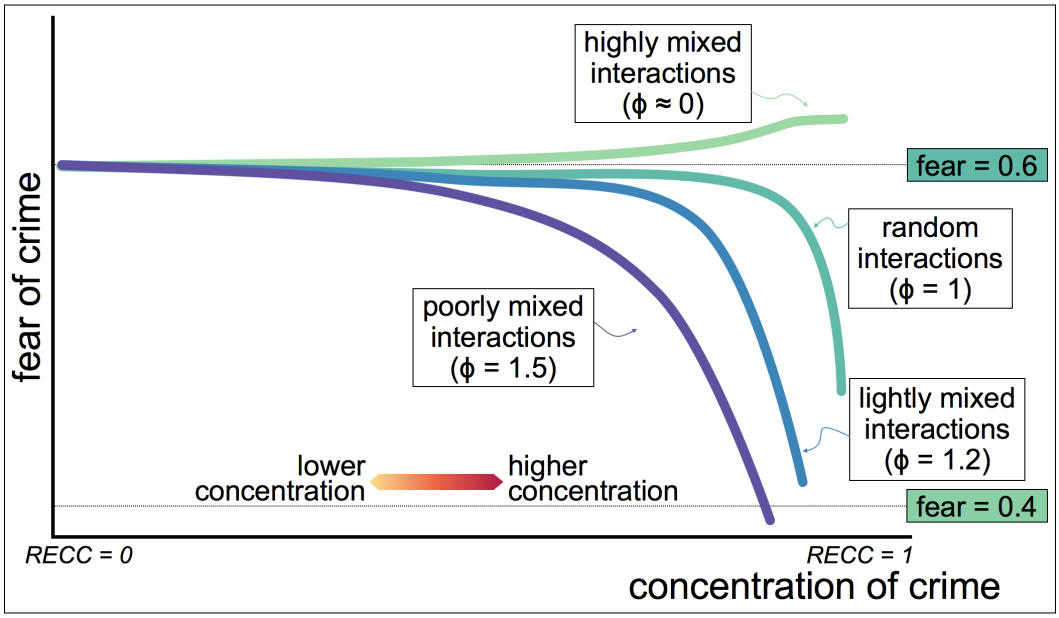

Figure 7.9: Impact of the concentration of crime on the fear of crime according to the degree of mixing between individuals from different groups.

When interactions between different groups are not frequent (corresponding to a higher value of ϕ) then the concentration of crime has a significant impact (Figure 7.9). With random interactions between individuals from different groups ($\phi = 1$) the concentration of crime still has an impact but only at very high concentrations of crime, and this impact is due to the fact that suffering an actual crime is less significant (that is, there is a smaller change in the mean fear of the population) if the person is highly

fearful. When groups are highly mixed ($\phi \approx 0$), which in turn means that interactions between individuals from the same group are less frequent, then people who suffer crime are more likely to interact with individuals who never suffer crime, increasing the mean fear of crime as a result.

7.5 Remarks

The findings provide a theoretical explanation as to why, when viewing crime and its fear, complex behaviour sometimes emerges. For instance, people who are immune to suffering crime still fear it; a higher concentration of crime reduces the generalised fear; the perception that a location is secure might rapidly change. From a global perspective, it is possible to obtain quantitative and qualitative results for the fear of crime and its dynamics. But from an individual perspective, it is not possible to say much regarding why a specific person has a fear of crime. If a person has a fear of crime of 0.7, with no additional information, it is unknown whether that number is the result of a past crime or a recent interaction with a fearful person. Furthermore, the group that the individual belongs to is also unknown, and so it is also unknown if she or he is more fearful than others in that group or perhaps, or whether their fear of 0.7 is as would be expected. The model here indicates the emergence of a social behaviour but not specific aspects of its individuals.

7.5.1 A useful simplification of a complex reality

It is not necessarily the case that more complicated models do a better job, and often, simple models are helpful due to their clarity and tractability (San Miguel et al., 2012). In terms of the fear of crime experienced by individuals, the model is a simplification of a much more complex set of social circumstances. However, it captures the three relevant and perhaps most obvious elements which have a clear impact on the fear of crime: suffering a crime, sharing experiences with others and having a decay of memory of the previous opinions and experiences.

In general, when individuals are more likely to interact with members of their own group, with a higher concentration of crime the fear of crime can drop significantly. People become segregated in many ways and for many reasons (Schelling, 1971) and so, the interactions between individuals from different groups are often difficult, for example, people from a run-down neighbourhood might have only a few interactions with individuals from a more upmarket community. Results shown here indicate that, in general, for poorly mixed groups, which experience a limited amount of interactions with individuals from other groups, a higher concentration of crime also means a reduced fear of crime.

The quantitative part of the model has, admittedly, some limitations in terms of a real-world observation to support the findings and any model validation could be achieved only by using qualitative methods (Pepys, 2016). However, although a simplification of a complex reality, it does help answer questions for which it was built (Johnson, 2000), particularly, the emergence of fear of crime in a population. There are also other ways in which the model could be configured, for instance, it is possible that the population could be separated into more than the three groups assumed here or the population could also be arranged into groups distinct by age, occupation or the neighbourhood in which they live. In this work, the yearly crime rate for all individuals is assumed to be constant and at the same rate for all members of each group, but this is clearly an oversimplification of the way in which people suffer crime. It is also assumed that individuals have a constant probability of having some form of social interaction with others and that they share their fear of crime and we assumed the same (asymmetric) impact for all interactions, but again, society and social interactions are far more complicated than this.

In terms of the quantitative measures, the simulations currently have little value for typical real-world applications and cannot be used to predict trends or forecast fear of crime as they stand. However, the qualitative as-

pects of the model explain the mismatch between crime and its fear and provide insights into why people who never suffer crime might still fear it. Additionally, the model shows that by changing the distribution of crime, but not the total number of crimes, the mean fear of crime can drastically change. The main contribution here is that the impact of the concentration of crime and its impact on the generalised fear of crime can be analysed.

7.5.2 Controlling fear of crime

Although it is not necessarily a palatable policy, results indicate that one way to improve the perception of security of a place is to increase the concentration of crime, that is, having a small population group which suffers the majority of the crimes. Results of the simulations show that two populations might suffer the same number of crimes, might have the same type of interactions and dynamics, but one of the groups might have a much lower fear of crime only as a result of a higher concentration of crime.

Another implication of the findings is that this research helps to understand the impact of a common phenomenon observed in different crime strategies, namely the displacement of crime. Unfortunately, some strategies oriented to prevent crime result in some degree of crime displacement (either the perpetrator chooses different locations, types of crime, *modus operandi*, or victims). That is, due to a policy oriented to reduce crime, different individuals become the victims of crime (Bowers and Johnson, 2003). Although crime displacement (and in particular victim displacement) is difficult to quantify, this work shows the potential result that a policy oriented to reduce crime might have. Whilst a policy could effectively reduce the number of crimes suffered by the whole population, if some degree of victim displacement is observed, then the chances are that the same policy also creates, at the same time, a more fearful population.

7.5.3 Policy implications

Beyond the fear of crime, similar ideas could be applied elsewhere to help

understand public opinions in different situations. For instance, towards international migration, whereby migrants moving to a limited number of cities might experience a different level of acceptance than the same amount of migrants who are more evenly distributed over all the cities of a country; or in respect to the use of firearms, where a significant event, such as a school shooting, might rapidly change the popularity of a gun policy before slowly decaying back to initial levels. In these and other cases, the model may be able to shed light on the way in which attitudes become heightened and social and cultural aspects which affect the opinion dynamics.

Chapter 8

Social media expressions of crime and fear

Previous chapters have analysed the fear of crime, firstly from a regional perspective and then from an individual perspective and how it emerges as a complex issue, not directly related to suffering a crime.

This chapter considers expressions of crime and fear of crime in social media and it correlates these expressions with quantitative measurements of crime and fear, showing that most of the expressions of crime posted in social media are related to fear of crime and not directly to crime itself.

8.1 Fear of crime and social media

To be newsworthy, social events must capture the attention of the viewer/reader and so they have to be either rare, timely, unexpected or have some special significance (Chermak and Gruenewald, 2006). However, the majority of crimes do not have these attributes. Property crimes, such as a mobile phone or a wallet being stolen, are not that rare when a whole city is considered. Also, most crimes do not have serious consequences or are merely attempted crime (Skogan, 1987). As a result, less than 1 out of 400 crimes actually appears in the traditional news (Chadee and Ditton, 2005). In fact, crimes are typically not newsworthy, except for crimes with violence

or with a sexual component, which therefore constitute a much larger part of the news with respect to non-violent crimes (Ditton and Duffy, 1983), even though non-violent crimes are much more frequent. For instance, considering the ten most popular printed daily newspapers published in the UK for a period of four weeks in 1989, it was found that nearly 65% of the space that was devoted to crime, was related to personal violence, whereas official statistics reported that only 6% of crimes involved violence (Dickinson, 1993). Similarly, taking 25 editions of newspapers in cities of the United States, it was found that nearly 30% of crime stories were murders, where in fact only 0.02% of the crimes are murders (Liska and Baccaglini, 1990). Traditional media thus gives a distorted version of the crimes within a city, but the audience is not necessarily aware of this issue and people perhaps consider that this form of media accurately captures the crime in their own city (Hollis et al., 2017).

The use of social media has completely revolutionised the way in which information is now shared and consumed. Social media has given its users the ability to share content and opinions without having to depend on traditional and centralised news media outlets, thus potentially obtaining a more *democratic* distribution of opinions, offering users the ability to reach a large proportion of the population (Kwak et al., 2010). Although most of what is shared in social media are not news, nor posts related to crime or public issues, it has nonetheless become, for some, one of the main sources of political information and news (Gil de Ziga et al., 2012).

A question that naturally arises is whether social media is different from traditional media regarding the information that is shared related to crime. Traditional media, for instance, typically cover major disasters in more depth than social media (Olteanu et al., 2015), and so, in terms of crime, this begs the question: does the information posted on social media provide a more accurate version of the criminal reality of a city than traditional media? Victims, indirect victims, and witnesses might be inclined to share their

experiences after suffering a crime, regardless of whether that crime was just an attempt or relatively trivial in its consequences and so not "newsworthy". If crimes are accurately reflected by Twitter posts, for instance, then it would provide a powerful tool for detecting crime trends and patterns (Kadar and Pletikosa, 2018) and even the locations of crime hotspots (Gerber, 2014; Chen et al., 2015). Leaving aside the potential readability issues (Temnikova et al., 2015) and fake news (Mendoza et al., 2010; Del Vicario et al., 2016), social media could provide insight into the analysis of crime patterns beyond merely estimating the density of people in space/time for a more accurate personal risk estimation (Malleson and Andresen, 2015).

Data collected from social media is a valuable input to analyse the flow of information, opinions, and sentiments and by detecting who shares what and how frequently. Millions (or perhaps even billions) of posts have been used to detect social media activism (Xu et al., 2014), to assist emergency responders (Avvenuti et al., 2016), to analyse the spread of a disease (Lampos and Cristianini, 2012), to detect the role of different users in the network (Martínez Teutle, 2010) and their behaviour (Cresci et al., 2016, 2017), to predict the movements of tourists (Muntean et al., 2015), to detect road traffic (D'Andrea et al., 2015), exposure to cross-ideological contents (Himelboim et al., 2013b), access to political information (Himelboim et al., 2013a) and political participation (Ausserhofer and Maireder, 2013), perception on social phenomena such as migration flows (Coletto et al., 2017), and even to detect the popularity of different types of food (Amato et al., 2017). Detecting the use of different words such as "food" or "wedding" allows the construction of a real-time measure of happiness or *hedonometer* (Dodds et al., 2011) which showed weekly and daily cycles of happiness.

There are many questions to be answered: is it possible to use the billions of social media posts in terms of the analysis of crime and its fear? Does social media provide a more accurate description of the social reality than other media? Is every crime equally likely to be posted on social media

and to be shared by the readers so that it reaches a large audience, or is there a specific type of crime that tends to be more commonly discussed by the users? Is the strong bias observed in the newspapers for reporting violence and sex-related crimes also observed in social media? Are there expressions of fear of crime in user's posts?

Social media has the potential to be used for measuring fear of crime and perceptions of security, that is, the perceived risk of suffering a crime. There is often a mismatch between fear of crime and the actual crime suffered in a city (Skogan, 1987; Prieto Curiel and Bishop, 2016b), or the risk experienced by individuals so that people often fear crime even if they are immune to suffering any (Prieto Curiel and Bishop, 2017). Users of social media might express their concerns and fears of crime more frequently in a more dangerous city, explaining, perhaps, how we arrive at our perception of security (Kounadi et al., 2015). Traditional ways of measuring the fear of crime frequently depend on victimisation surveys (Hale, 1996; Ferraro and Grange, 1987; Carro et al., 2010) which have a considerable time delay between the moment the study is conducted and when the data is available for analysis, but with social media, it would be possible to obtain an almost immediate reflection of the fear of crime in a city. Social media could, therefore, be a powerful tool for measuring the fear of crime and perceptions of security but only if it is, in fact, related to actual fear.

There are some potential issues to consider when it comes to using social media to understand the crime suffered by a population. Firstly, not everything posted on social media is true (Mendoza et al., 2010); secondly, posts on social media might be challenging to understand due to the use of abbreviations, typos, the use of hashtags, lack of connecting words and more (Temnikova et al., 2015). Also, although social media posts offer a fast distribution of information, it was found that there is frequently a delay of possible several months in the case of crime posts in social media (Kounadi et al., 2015), so that, in fact, posts are not necessarily a reflection of the

current crime and security situation of a city.

Using social media to understand crime patterns has been suggested before. For instance, the total number of tweets posted from different locations helps identify regions in a city where a person has a higher risk of suffering a crime, taking into account criminal hotspots and, at the same time, the population density (Malleson and Andresen, 2015). Also, identifying relevant social media topics at different locations has been suggested as a source of information that could help in predicting crime (Gerber, 2014). But, if posts on social media suffer the same bias towards violence and sexual related crimes as traditional media, is it even worth using social media for predicting crime and policing? If social media posts are a biased picture of reality, with a considerable delay between the time of occurrence of the event and the time of the post, is the social media of any use for resource allocation, emergency attention, or policy design? Is social media an actual reflection of fears of crime and perceptions of insecurity?

8.2 Social media expressions of crime

The objective here is to detect expressions of crime and fear in social media and compare them with the actual crime suffered. Data from Twitter users in 18 Spanish-speaking countries in Latin America were collected with this choice of target group selected for several reasons. Firstly, there are roughly 400 million people who have a shared Spanish language and history, with Spanish being only second to Chinese as the most used (first) language. This means that there will be a large amount of data to collect. Secondly, the data spans a number of countries allowing sufficient breadth so that it is possible to compare one country with another where there might be different social norms and crime rates (as opposed to a choice of Chinese or Hindi, for instance, which may have large numbers of speakers but would not give the variation of countries). Additionally, with so many countries, there is a reasonable chance that the data will not be dominated by local trends (for

instance, presidential elections or other political events which could have a substancial impact on local expressions on social media). Finally, according to the World Bank[1], Latin America provides observations from populations with a higher level of internet users (more than 60% of the population in countries such as Venezuela, Mexico, Chile or Argentina, among others), as opposed to the French-speaking nations (less than 15% in countries such as Togo, Benin, Niger, Mali or Burkina Faso, among others).

Data was also collected, where possible, at a metropolitan level, each containing more than one million people, allowing data from 64 cities to be analysed. Smaller cities were not considered since expressions of crime in social media from smaller cities are scarce.

8.2.1 Social media posts

All the data used in this study have been collected from Twitter, by means of a streaming crawler. The data collection process took place from May 22 to July 30 2017 (inclusive), lasting 70 days in total. During this time-span, Twitters Streaming API[2] was exploited to access the global stream of tweets. In order to retain only data relevant for our analysis, the streaming crawler was configured so as to apply a geographic filter on incoming data. Specifically, all geolocated tweets shared from within 18 Spanish-speaking countries from Latin America were collected over the given 70-days time window. Furthermore, retweets have not been considered in the data collection. This resulted in a dataset of 32,513,684 distinct tweets, 27% of them shared from Mexico, 23% from Argentina, 12% from Colombia, and the rest from smaller countries.

Whenever available, more detailed geographic (i.e., city level) information about the collected tweets was also extracted. Indeed, some users allow their location to be available and so, for a subset of all tweets, a city level location can be identified, which allows collected tweets to be assigned to

[1] Data available at `https://bit.ly/2BGjdB8`

[2] `https://bit.ly/2KSsvQp`

different metropolitan areas. In detail, collected tweets were mapped to the 64 largest metropolitan areas from the 18 countries considered. As a result, we obtained 2,678,783 tweets (8.2% of the total) with a city level geographic resolution.

8.2.2 Classification of crime-related tweets

An extensive list of words was created consisting of 392 words related to crime, or fear of crime. Thus, if a tweet contains a word related to crime, the post is considered to be *crime-related*. In the list of relevant words, 274 are in Spanish and the rest of the words are in English since some of the tweets are published using other languages. Words related to crime are, for example, “murder”, “stolen” or “weapon”.

Each word related to crime was also assigned to different categories, such as “violence-related”, “property-crime-related”, “organised-crime-related”, “sexual-crime-related”, “murder-related” and “gun-related”. Some words are not exclusive to a single category and so, for instance, the word “massacre” is included in the violence, murder, and gun-related topics.

If a tweet contains any of the words related to crime, then the tweet is considered to be crime-related. In addition, if the tweet contains a word from a specific category, then the tweet is also deemed to be relevant for that category. Since words might be assigned to more than one category and since a tweet might contain two or more words related to crime, a crime-related tweet might be assigned to more than one category.

Every tweet is considered equally to determine the expressions of crime and fear in social media, although it is important to note that some tweets might reach a large audience while others do not (Mendoza et al., 2010), and in general, accounts owned by organisations are more influential than those by individual users in respect to the number of followers (Xu et al., 2014).

8.2.3 Non criminal crime-related tweets

Some tweets might contain phrases like “I would *kill* for a holiday right now” or similar expressions that are not actually related to crime or the perception of insecurity. To determine the frequency of this type of mismatch expressions that are not associated with crime, 3,000 tweets were individually read and manually analysed.

Although for some tweets it is not easy to determine whether or not the post was related to an actual crime, of this 3,000 manually analysed crime-related tweets posts, roughly 66% were found to be related to a crime, or an expression of fear of crime, or a demand for justice. It is also possible that many tweets are related to crime but they do not include any of the 392 pre-selected words, however, by having such an extensive list of items related to crime or fear, this type of error is considered to be kept to the minimum without having to manually read every tweet and determine whether or not the post is related to crime or fear.

No temporal or geographical pattern was identified in the mismatch from the crime-related tweets and since it is very time-consuming to manually read the collected tweets, following the test sample it is assumed that roughly only two-thirds of the crime-related tweets are actually related to crime (the other third is just an expression of something else), and this is assumed to be uniform across all tweets considered to be crime-related. Since it is not possible to classify crime-related tweets between those which are actually related to crime and those which are not, all crime-related tweets are kept and considered further.

The majority of the crime-related tweets were identified as being posted by three different types of account: a news agency, such as a newspaper or a reporter; an institution such as the Police Department and finally, people who have outspoken via social media harbouring a general complaint about the levels of crime, fear of crime or social justice -in other words, *activists* for criminal problems in their city or country. Crime-related tweets from other

accounts tend to be much less frequent. In most of the cases, when a specific crime was mentioned, the tweet also contained a link to the information source, and the majority referred to a traditional media website, that is, the tweet was not posted by the victim or victims of the crime or witnesses. Similar to what has been encountered in the political debate (Himelboim et al., 2013a), traditional media has a strong impact on social media in terms of what is posted related to crime.

8.2.4 Crime-related tweets

Since Twitter might have a lower or higher penetration in different countries or even cities, the number of crime-related tweets is reported for every 1,000 tweets posted, thus, a lower or higher penetration should not impact on the metric.

Based on the approach outlined, of the 32.5 million tweets collected, 501,057 are crime-related. Thus, 15.41 tweets out of 1,000 tweets posted in the major Spanish-speaking countries in Latin America are considered to be related to crime.

During the 70 days over which data was collected, there are 317.5 tweets posted every minute from the 18 countries considered and from these tweets, 5 are crime-related.

8.2.5 Crime-related categories

From the 15.41 out of 1,000 crime-related tweets, 6.51 out of 1,000 tweets are violence-related, with this being the most frequent crime-related category. Categories might have a considerable overlap or might even be fully contained, for instance, murder-related tweets are fully contained in the category of violence-related tweets.

The most frequent crime-related tweet is linked to violence, which is 3.7 times more frequent than property-crime tweets. Also, murder-related tweets are 2.3 times more frequent than property-crime tweets.

Every minute, there are 5 crime-related tweets posted within the 18

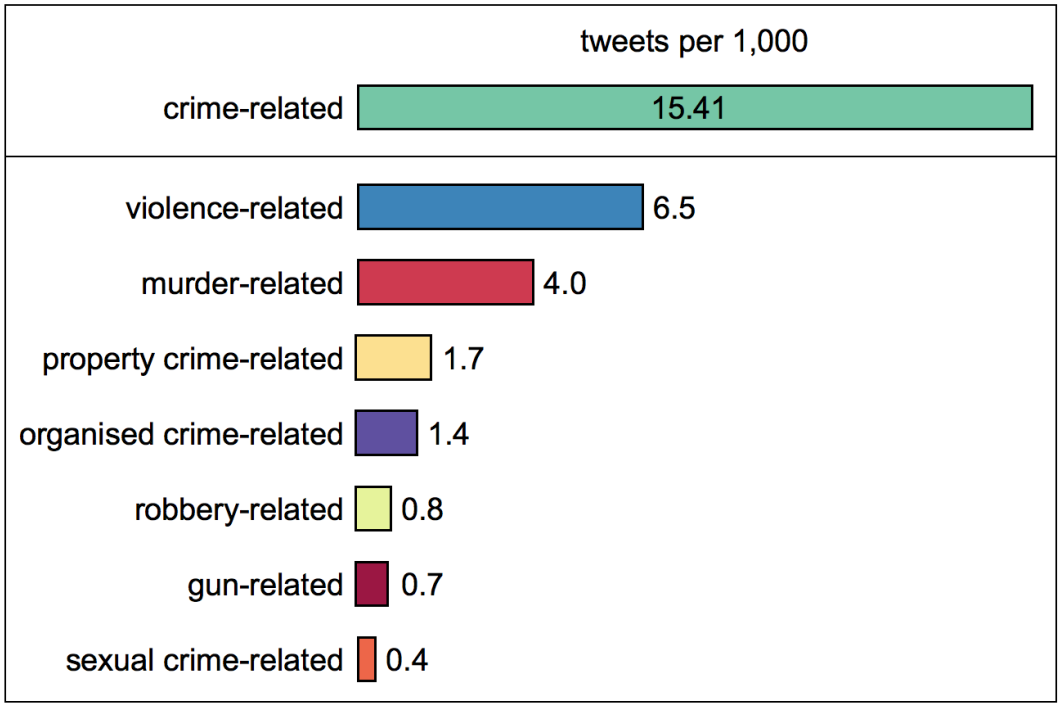

Figure 8.1: Crime-related tweets per 1,000 in Latin American Countries. The most frequent crime-related tweet is related to violence forming more than 40% of the collected tweets.

countries in Latin America and every minute, there are 2 violence-related tweets.

8.3 Social media posts against the observed reality at a country level

There is an essential difference between the number of crime-related tweets per 1,000 observed in the different Latin American countries. In Venezuela, the county with the highest proportion of crime-related tweets, 38.1 tweets out of 1,000 are crime-related, whereas, in Nicaragua, Paraguay, Costa Rica and Bolivia, less than 10 tweets per 1,000 are crime-related.

8.3.1 Crime at a national level

Unfortunately, the majority of crimes are not reported to the police and any cross-national comparison strongly depends on the definitions used for different types of crime. Therefore, *intentional homicides* are used for com-

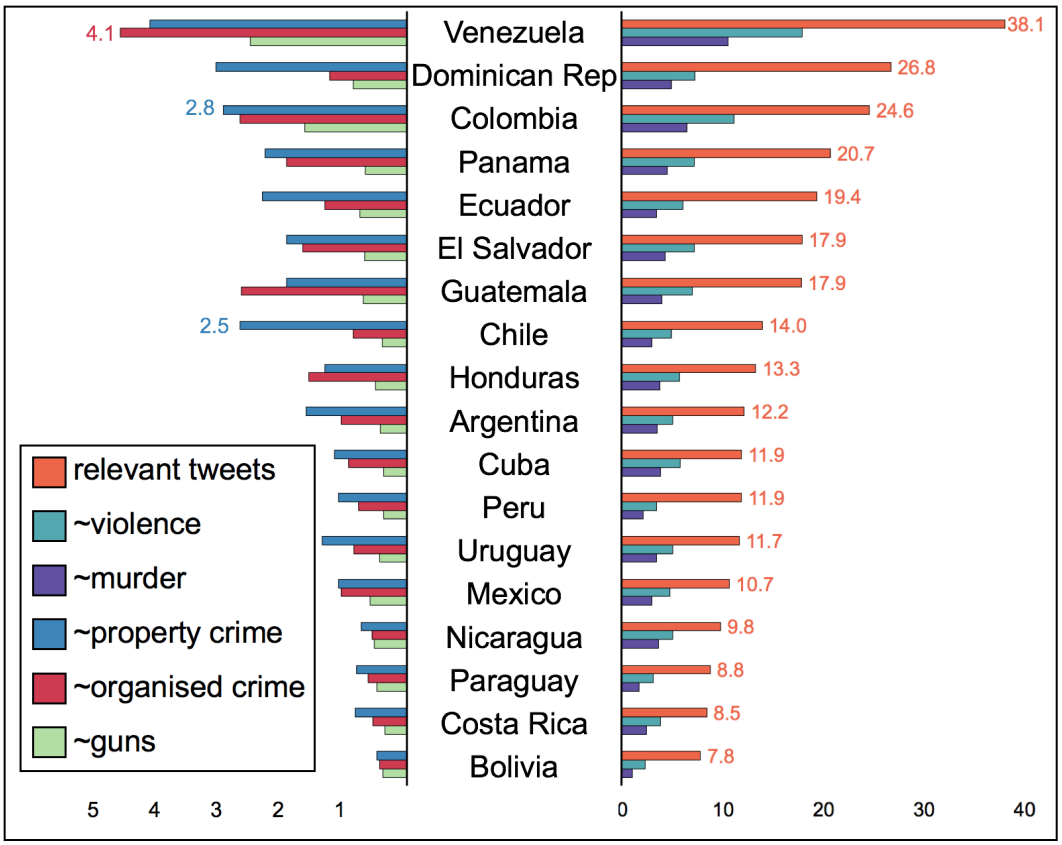

Figure 8.2: Crime-related tweets per 1,000 in Latin American Countries. In Venezuela, 38.1 tweets from every 1,000 published are crime-related, a much higher number than the 15.4 observed in the 18 countries observed.

paring between different countries since they are perhaps the most reliable way to compare crime between different countries. The most up to date data for the number of intentional homicides, as well as the number per 100,000 inhabitants, is published by the United Nations Office on Drugs and Crime[3] per country. The number of intentional homicides for the year 2015 is available for most of the countries analysed, although in some cases, it reflects data from previous years. In any case, the data has a considerable delay with respect to the social media posts being analysed and should be considered only as a proxy to determine a general level of insecurity observed at a country level.

The Pearson's correlation comparing the number of crime-related

[3]Data downloaded on the 1st October 2017 from `https://data.unodc.org/`

tweets, violence-related tweets and murder-related tweets out of 1,000 tweets posted, against the number of murders, considering the 18 countries shows a positive relationship between them so that countries with a higher number of murders are also expected to have more crime-related tweets. Results show that countries with a higher number of murders are more likely to have crime, violence, and murder-related tweets (Figure 8.3). Also, a linear regression, using the number of crime-related tweets as dependent variables and as independent variables the crime rates and the fear of crime was computed, with the full results on Figure B.3. In other words, every murder in the country increases the number of crime-related, murder-related, and violence-related tweets.

	crime-related tweets per 1,000	violence-related tweets per 1,000	murder-related tweets per 1,000	property-crime-related tweets per 1,000
murders	0.449	0.614	0.582	0.358
murders per 100,000	0.424	0.520	0.507	0.268
fear of crime index	0.561	0.427	0.367	0.454
% of people with strong fear	0.750	0.652	0.619	0.638

Figure 8.3: Pearson's correlation taking into account the 18 Spanish-speaking countries in Latin America as observations.

Broadly speaking, one murder in Latin America is associated with 8.4 murder-related tweets, 13.7 violence-related tweets, and 32.4 crime-related tweets. However, some countries feature a higher level of "activism" expressed in social media. In Uruguay and Chile (two of the countries with the highest GDP per capita of the region), a murder is associated with more than 60 murder-related tweets but in Bolivia, Cuba, Honduras, Guatemala, and El Salvador (in general, countries with a low GDP per capita), a murder

is associated with less than 3 murder-related tweets.

The population size is significant when considering the amount of crime in a country. Mexico has the largest population in the Spanish-speaking Latin region and it has 2.5 times the inhabitants of Colombia and 2.8 times the inhabitants of Argentina, the second and third largest countries respectively. Mexico also has 19 times the population of Nicaragua and 36 times the population of Uruguay and so just by considering the size of the country, Uruguay or Nicaragua should have fewer murders, in total, than Mexico. Therefore, comparing the number of murders per 100,000 people takes into account the impact of the size of the country. Results show that countries with a higher murder rate are also more likely to observe more tweets related to crime, violence, and murder, and also, countries with a higher number of murders have more tweets related to crime, violence, and murder (Figure 8.3). This is, perhaps, as expected, since it is likely that every murder is covered by national news media outlets who then post the event on their own social media, so that countries with more murders are expected to have more tweets related to crime and similarly for the murder rate.

8.3.2 Fear at a national level

Fear of crime is difficult to measure and has many interpretations (Ferraro and Grange, 1987). It depends on the source of the information and unfortunately, there is no standard way to assess fear. For instance, questions such as “How safe do you feel walking alone after dark?” are frequently used in surveys with many interpretations and different answers (Farrall et al., 1997; Hale, 1996), which makes cross-national comparisons complicated. There are other ways of measuring fear of crime, but they are linked to the way the data is asked and then coded during the survey.

The Latin American Public Opinion Project LAPOP (Latin American Public Opinion Project , LAPOP) carries out a public opinion survey in Latin America and covers 34 countries, including the 18 countries in which Tweets were collected. Among many valuable topics, LAPOP covers crime and fear

of crime and includes the question “Speaking of your neighbourhood and taking into account the probability of being the victim of a robbery of a person, do you feel (a) very secure; (b) somehow secure; (c) somehow insecure; and (d) very insecure. Two different metrics are considered; firstly, the percentage of people from the country who feel very insecure and secondly, an index is constructed by assigning a $+4$ to the answer “very insecure”, a $+3$ to “somehow insecure”, a $+2$ to “somehow secure” and $+1$ to “very secure” and then compute the average value of the answers in a *fear of crime index*. A similar technique constructing an index from the ordinal answers (ordered answers from the least fear to the most fear) has been used (Tseloni, 2007; Pantazis, 2000) on the Crime Survey for England and Wales (Office for National Statistics, 2016).

In the case of the data available for the analysis of fear of crime, as well as for the data for the number of murders, there is a time delay with respect to the social media posts which is, for practical reasons, unavoidable. Although the time window for both data sets (the social media and the surveys or the crime data) does not match, it provides a good starting point to compare whether crimes with violence or a sexual element are more likely to be mentioned than others, and to detect whether fear encourages more activism in social media. The crime rates in a country and their fear of crime might vary from one year to the next one, but usually, it is only a slow change, so that past crime and past fear of crime patterns are sufficiently accurate to describe present patterns (Prieto Curiel and Bishop, 2016b).

Similar to the analysis for intentional murders and the number of crime-related tweets, the Pearson’s correlation (and a linear regression) between the fear of crime and crime-related tweets helps explain whether fearful countries display more expressions of crime on social media. At a national level, there is a positive and significant relationship between the percentage of people who have a strong fear and the number of crime-related and violence-related tweets per 1,000. The results obtained in the models using

either the percentage of people who feel very insecure or the fear of crime index, reveals exactly the same pattern for both metrics since, at least for the case of the LAPOP data, they are highly correlated. However, the variable *percentage of people who feel very insecure* seems to be a better way of explaining posts in social media as opposed to the constructed "fear of crime index" as it has a higher correlation.

Countries with a higher number of murders and a higher murder rate have more crime-related, murder-related, and violence-related tweets per 1,000 tweets. However, fear of crime is a more relevant variable as to why a country has more crime-related tweets than the actual number of murders or the murder rate. Although crime and fear of crime are related, this relationship is not perfect (Prieto Curiel and Bishop, 2016b), and the results show that activism in social media is explained more by the level of fear of crime than the actual crime (Figure 8.4).

Although the causal mechanism of whether more fear of crime creates more activism on social media, or whether a higher awareness of crime on social media creates more fear is not clear, results, however, show that social media can be utilised to measure, to a certain degree, the fear of crime in the society at a national level.

There are significant outliers (e.g., El Salvador and Honduras) that have a much less number of tweets related to crime or violence than their number of murders would explain. These two countries also have a relatively small percentage of the population with a strong fear and so it seems that the fact that El Salvador and Honduras have less crime-related tweets than expected could be explained by the fact that they also have less fear than might be predicted based on their murders or their crime rates.

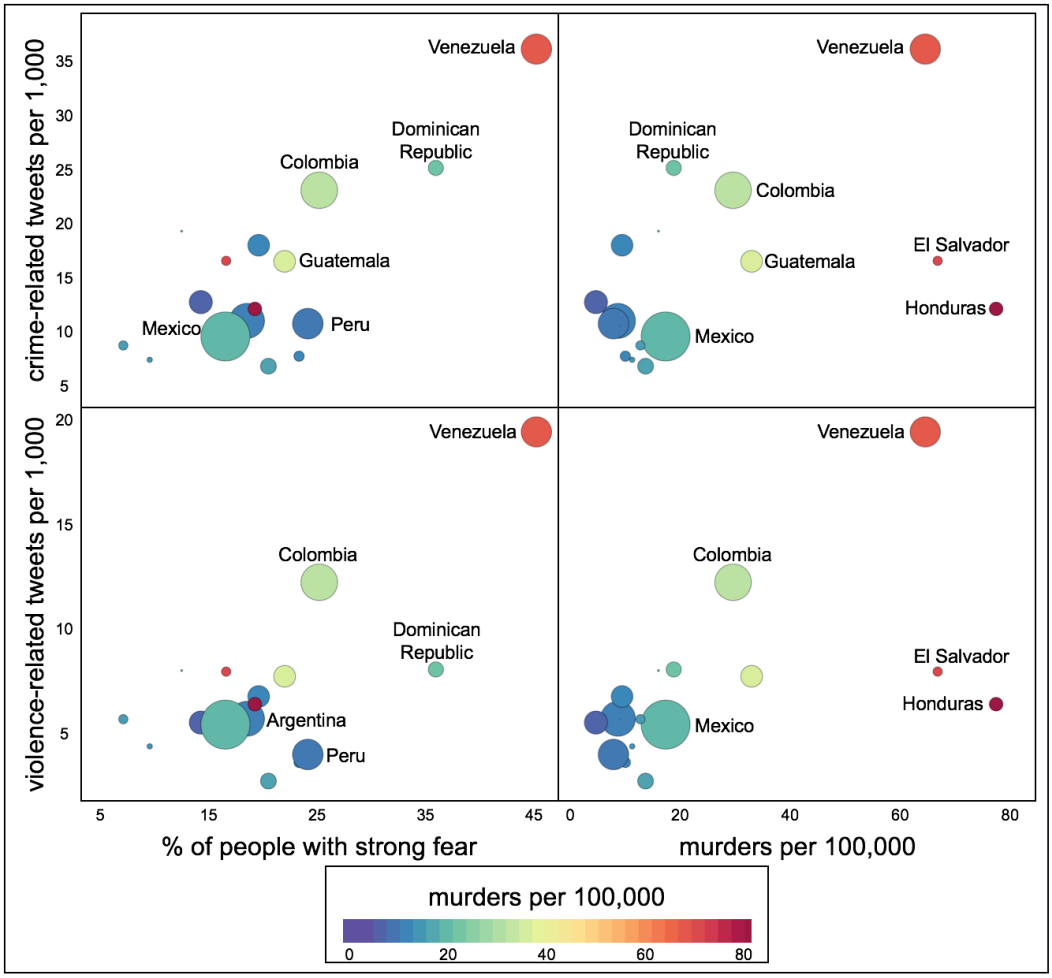

Figure 8.4: Tweets related to crime and violence per 1,000 against fear and murder in Latin American Countries. The size of the disc represents the size of the country, and the colours represent the murder rate.

8.4 Social media posts against the observed reality at city level

Comparing the actual crimes suffered in each metropolitan area to the ones expressed in social media at a city level is complicated with respect to both variables. Unfortunately, a comparison between cities from different countries is not possible due to the varying definitions of crime, ways of measuring crime, and significant, but varying, issues with unreported crime. The LAPOP survey used at a national level is not representative at a metropolitan level and therefore, it cannot be used either. For the social media posts,

a more local dimension, such as cities, is complicated since only a small number of users publish their location and it was found that the accounts which post their location might be local newspapers which tweet about crime much more frequently than a general user would which creates a strong bias in some cities.

The natural expectation is that cities with higher crime rates also have a higher number of tweets related to crime.

The number of users who share their location, so that they can be assigned to a specific metropolitan area, is considerably smaller. From the 32 million tweets collected, only 2.68 million are assigned to a city (8.3%). In addition, only 19,912 tweets are crime-related tweets and share their location, which represent 7.4 per 1,000 of the tweets collected at a city level. Activism drops to less than half, from the 15.4 out of 1,000 crime-related tweets at a national level, to 7.4 out of 1,000 when users actually share their location on social media.

The number of relevant tweets at a city level drops considerably, compared to the national level. There are cases, for instance, Cochabamba (Bolivia) and Arequipa (Peru), where the number of relevant tweets detected for a period of 70 days is only 2 and 5, respectively. In 31 cities, less than 70 geolocated crime-related tweets were collected which means that in nearly half of the 65 largest cities in Spanish-speaking countries in Latin America, less than one relevant tweet is published each day.

Taking into account that a tweet is considered to be relevant if it contains at least one of the 392 words related to crime, such as “gun” or “murder”, having less than 4 relevant tweets each day, as occurred in 51 of the 65 cities (78% of them) and in cities such as Medellín (Colombia) Guatemala (Guatemala), or Monterrey (Mexico), which have more than 4 million inhabitants in their metropolitan area, shows that at a city level results do not necessarily represent their criminal situation or their fear of crime.

8.4.1 Social media against crime and fear at city level

Despite the aforementioned issues, here, a focus on Mexico is useful since data is available to compare between the 23 metropolitan areas with at least 750,000 inhabitants. A victimisation survey from Mexico Encuesta Nacional de Victimización y Percepción sobre Seguridad Pública (from INEGI, 2016) provides estimates for the crime rates suffered by the population of each city, divided into different types of crime, and provides a metric of the population who have fear of crime in their city and therefore, it allows quantifying the level of crime and fear of crime in each of the 23 cities using the same definitions. Considering other countries is possible but they give a smaller number of city-observations (7 in Colombia; 6 in Argentina and 5 in Venezuela).

For the specific case of La Laguna, a metropolitan area in the North part of Mexico with 1.2 million inhabitants, it was detected a media consortium (including three different newspapers) which frequently shared crime-related tweets on their own accounts (in total more than 95% of the crime-related tweets from the city) and which also shared their location. Since a similar situation does not happen in any other city and newspapers do not often share their location on every tweet, La Laguna was dropped from as an observation from the statistical analysis to avoid having strong outliers (Figure 8.5).

The victimisation survey in Mexico considered the total number of crimes suffered by the population in 2016 and so the time intervals between the two data sources, again, do not match. However, the objective here is to detect differences between the expressions of crime in social media and the actual crime suffered by the population or their fear of crime and so, even when the time interval for both data sets do not match, the survey is used as a proxy to their crime and fear.

8.4.2 Crime at a city level

There is practically no relationship between the number of crime-related tweets per 1,000 to either crime, property crime, perceived fear at a local

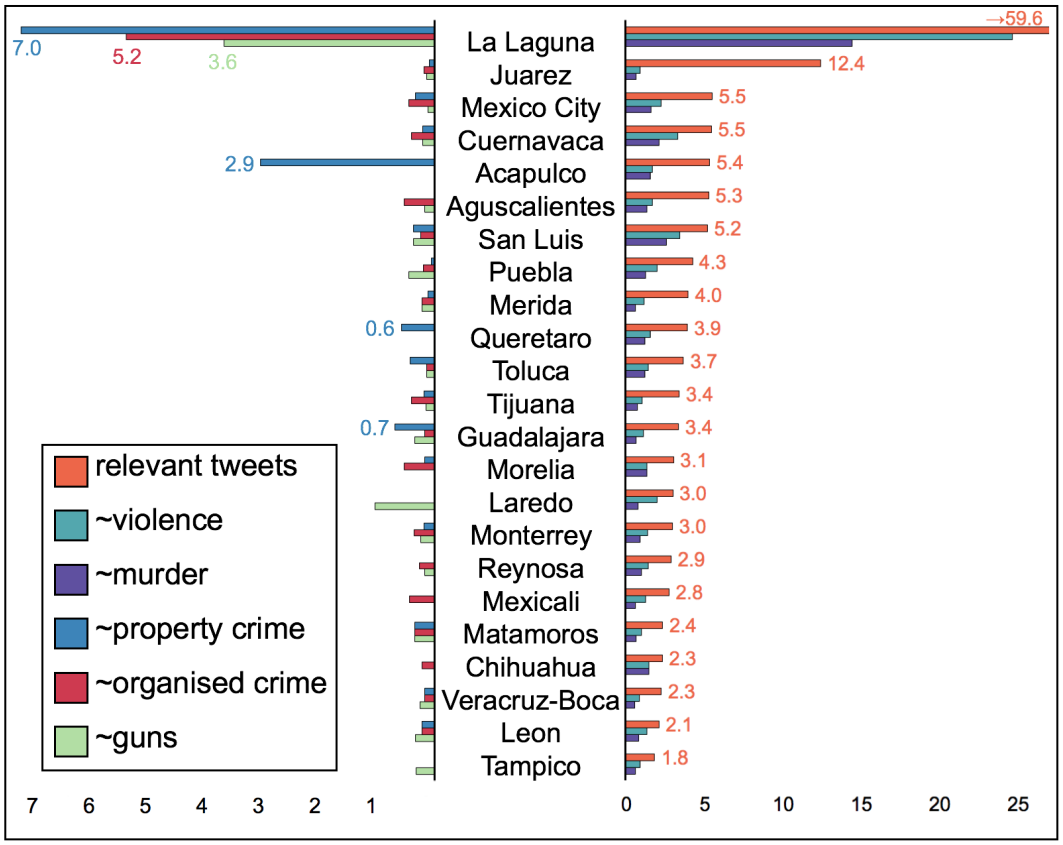

Figure 8.5: Crime-related tweets per 1,000 in Mexico. The number of tweets related to crime, violence, murder, property crime, organised crime and guns per 1,000 tweets in the largest 23 metropolitan areas in Mexico.

level, at county level or at province level (Figure 8.7). The expressions of crime in social media and why a city has more crime-related tweets per 1,000 than other cities is not due to its murder rates, violence or any combination. The Pearson's correlation between the number of crime-related, murder-related or property-crime-related tweets per 1,000 against the number of crimes or the fear of the population are reported (Figure 8.6) and also, a linear regression, using the number of crime-related tweets as dependent variables and as independent variables the crime rates and the fear of crime was computed (the full results on Figure B.4). The number of property-crime-related tweets per 1,000 is only loosely related to the hard crime rate observed in each city (where "hard crime" includes murder, kidnap and missing person).

	crime-related tweets per 1,000	violence-related tweets per 1,000	murder-related tweets per 1,000	property-crime-related tweets per 1,000
murders	0.264	0.173	0.139	0.138
murders per 100,000	0.375	-0.223	0.121	0.776
hard crimes	0.115	0.120	0.148	0.101
hard crimes per 100,000	0.102	0.195	0.147	0.668
robbery of a person	0.238	0.259	0.285	0.323
robbery of a person per 100,000	0.199	0.305	0.307	0.259
any crime per 100,000	0.180	0.101	0.135	0.170
fear in the county	-0.157	0.135	0.143	0.320

Figure 8.6: Pearson's correlation taking into account cities in Mexico as observations.

8.4.3 Fear at a city level

In the Mexican victimisation survey (from INEGI, 2016), fear of crime is based on the question "*In terms of crime, do you consider your region to be secure or insecure?*" with answers 'Yes' and 'No', so fear of crime at city level is considered to be the proportion of people in the city who consider their region to be insecure (Prieto Curiel and Bishop, 2016b).

As opposed to the national level, with a fairly clear relationship between the number of social media posts related to crime and the actual crime rate

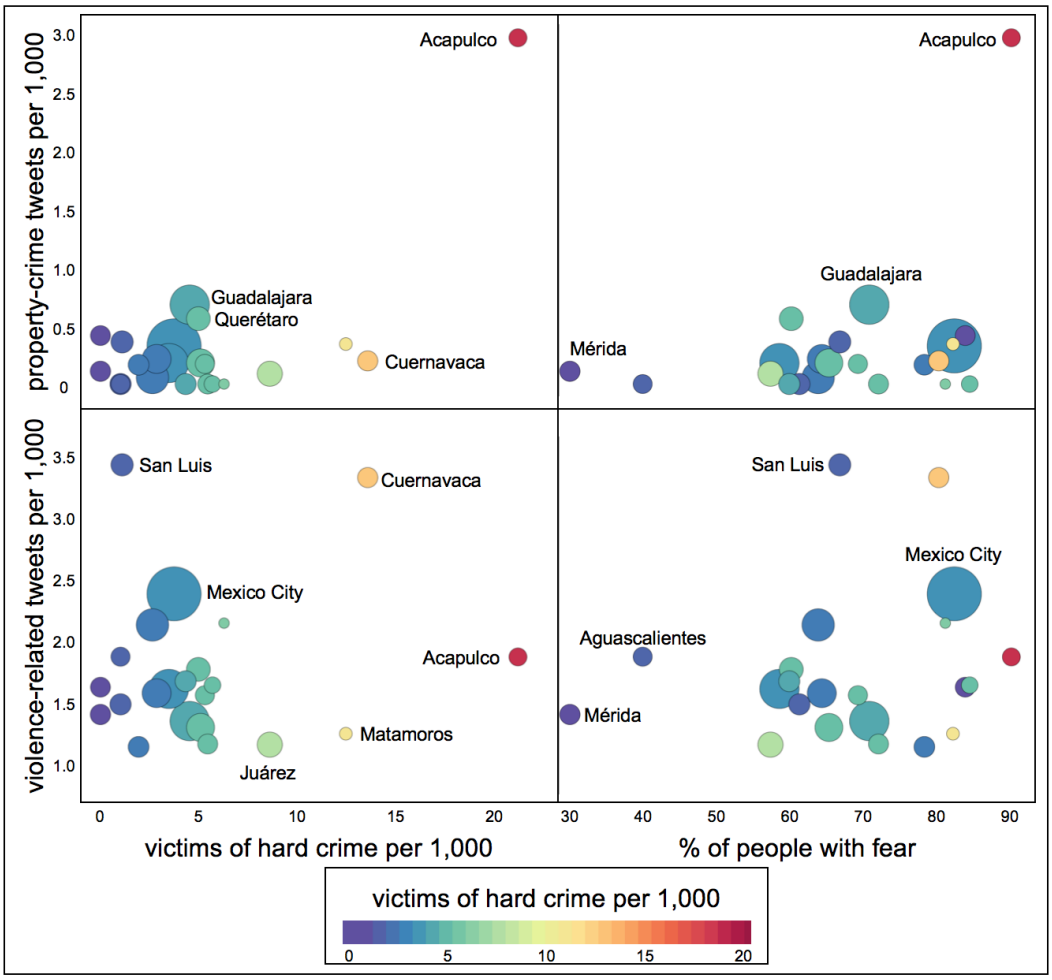

Figure 8.7: The size of the disc represents the size of the country and the colours represent the murder rate.

or fear of the country, the relation between crime-related tweets and actual crime or fear of crime is less pronounced at a city level. There are two explanations for this observed decrease in significance between social media posts related to crime and actual crime when a smaller geographical unit is explored. The first explanation has to do with the crime-related topics. It was found that many of the general complaints and expressions about insecurity and activism on social media are not specific to a city but to the country as a whole. Thus, many of the expressions related to murder, violence and with a sexual component do not necessarily match the city in which the tweet is posted.

The second explanation has to do with the tweets. From the 32 million

tweets collected which are assigned to a unique country in Latin America, only 8.2% of them can be assigned to a particular city, which means only 850,000 in the 23 cities considered in Mexico. From these 850,000 tweets, only 6 out of 1,000 tweets are related to crime (in total 5,097 tweets from Mexico are crime-related and can be assigned to one of the 23 metropolitan areas analysed), which means that for some cities (Tampico and Matamoros, for instance) the number of tweets and therefore, crime-related tweets or violence-related tweets, is close to zero. In 15 cities in Mexico, less than 5 property-crime-related tweets were detected during the 70 days of the data collection. Thus, at a city level, it is no longer a big data type of problem with millions of tweets, but it is only a few users tweeting about their crime or fear. The vast majority of the users do not share their location and so, in the city of Puebla, for instance, with its metropolitan area of 2.7 million inhabitants and collecting tweets for a period of 70 days, only two geo-located property-crime-related tweets were detected.

Unfortunately, at a city level, social media posts offer little information about the crime suffered or the fear of crime. Hence, forecasting crime, using tweets to detect hotspot patterns and for policing, measuring fear of crime, activism or public opinions seems almost impossible at a city level using tweets.

8.5 Social media compared against reality

Results show that social media posts at a city level are not an adequate approximation to reality but, at a national level, they are. Considering only the tweets from Mexico, it is possible to compare the number of posts related to different types of crime with what is actually suffered in the country.

According to the Mexican victimisation survey ENVIPE (from INEGI, 2016) for every murder there are roughly 34 crimes with a sexual component (including rape, rape attempts, harassment, exhibitionism); 917 property crimes (including violent and non-violent crimes in which property is

stolen from the victim, such as car theft, robbery of a person, burglary and others) and 1,391 crimes including all types of crime, but this is far from what is portrayed in social media.

Although there is not a match for the time studied between the posts on social media and the victimisation survey, results show that in Mexico:

- there are 0.0144 tweets related to crime for every crime suffered, regardless of whether or not the crime was violent or with a sexual component;
- there are 0.0021 tweets related to property crime for every property crime suffered in the country;
- there are 0.0141 tweets related to sexual crimes for every sexual crime suffered in the country; and
- there are 5.675 tweets related to murder for every murder suffered.

Assuming that tweets are a direct response to a specific type of crime, results show that a crime with a sexual component is tweeted in a sexual-crime-related post 6.61 times more frequently than property crimes are posted in property-crime-related tweets. Murders are tweeted 401 times more frequently on a murder-related tweet than sexual crimes on a sexual-crime-related tweet.

In the tweets collected, 28.3% of the crime-related tweets are related to murder although murder accounts for only 0.072% of the crimes suffered in Mexico. This mimics almost to perfection the results found in the 25 editions of leading newspapers in each of the 26 US major cities, in which it was found that nearly 30% of the crime stories were murder, but it represents only 0.02% of crimes (Liska and Baccaglini, 1990).

8.5.1 Temporal expressions of crime in social media

An aspect of the expressions of crime in social media is the time at which posts are published, particularly the day of the week. Users post more

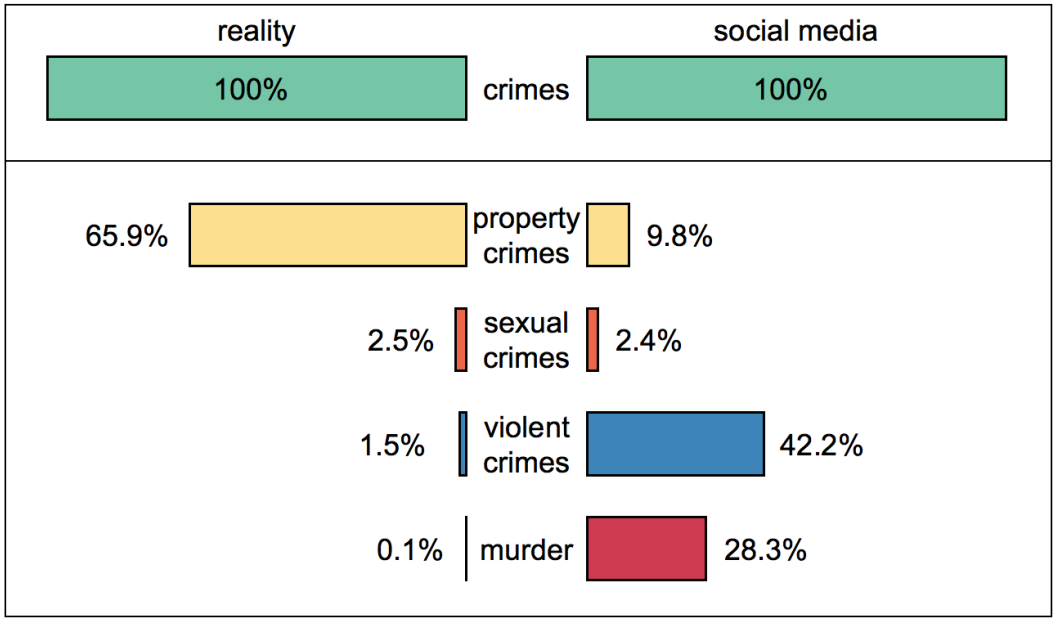

Figure 8.8: Observed frequency of different types of crime against crime-related tweets. Comparing different types of crime suffered in Mexico against the frequency at which they are tweeted shows that social media has a strong bias to sexual and murder-related crimes.

tweets each day from Monday to Friday, more specifically on Wednesdays and Thursdays than during the weekends. The average number of tweets per day drops by 13% during the weekend in the countries studied.

From Monday to Friday, in the 18 countries considered, there are 15.84 crime-related tweets per 1,000 tweets, but this decreases to 14.14 during Saturday and Sunday. Similar to the weekly pattern observed in the tweets related to happiness (Dodds et al., 2011), crime-related tweets decrease by 12% during the weekend. The same trend is observed throughout the 18 countries; in Mexico, for instance, crime-related tweets per 1,000 decrease 19.5% during the weekends and in Peru, they decrease 20.6%.

Similarly, reductions occur for the violence-related (from 6.6 tweets per 1,000 during the weekdays to 6.1 during the weekends) and murder-related tweets (from 4.1 tweets per 1,000 during the weekdays to 3.8 during the weekend), meaning that tweets are, in general, less likely to be related to murder, crime or violence during the weekends than during the rest of the week.

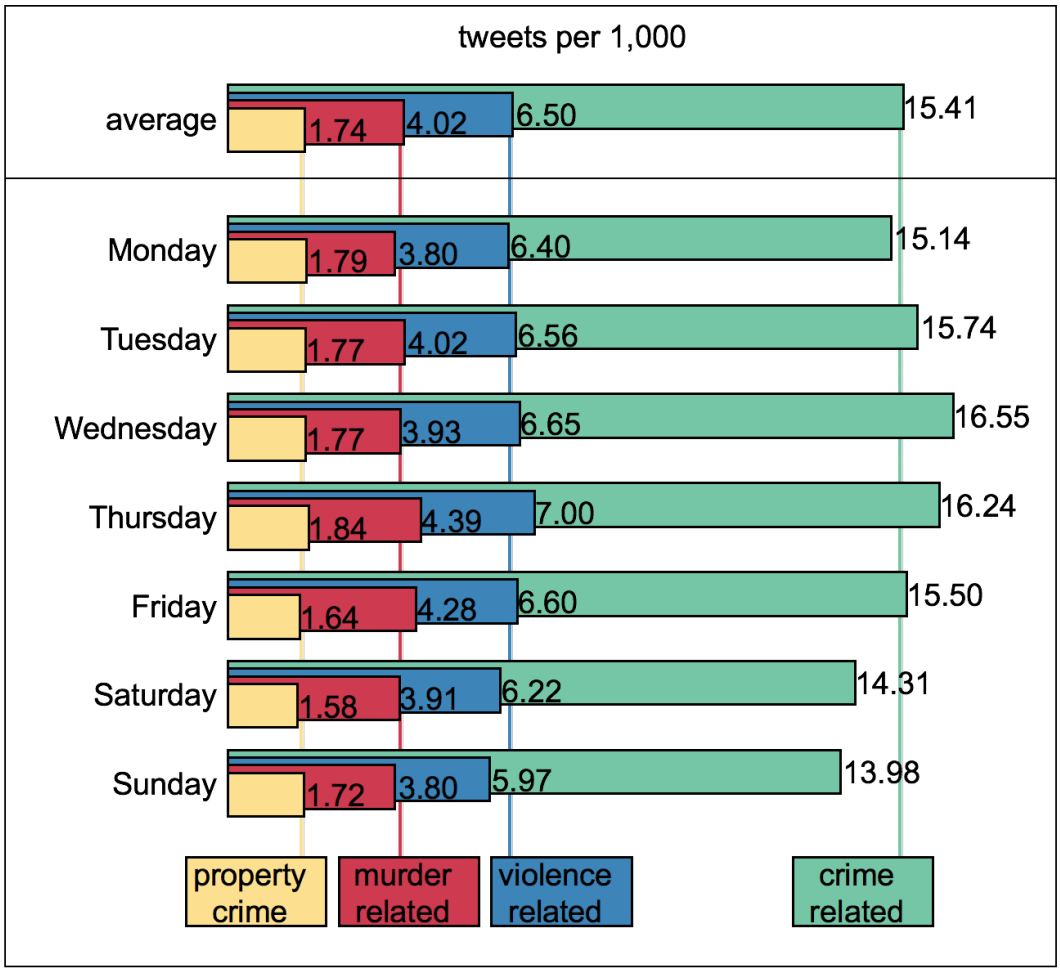

Figure 8.9: Weekly pattern of expressions of crime in social media. During the days of the week, the posts related to fear of crime, crime and violence change according to the day of the week.

In another study, using social media, a robust weekly cycle on the happiness was detected (Dodds et al., 2011) and in the case of crime, results here show that social media activism also has a weekly cycle and that this activism decreases during the weekend.

8.6 Remarks

Collecting tweets over a period from users in the larger countries in Latin America resulted in a large dataset of tweets. Using social media to measure expressions of crime, fear and activism and by collecting 32.5 million tweets from the largest 60 cities in Latin America, a comparison between

what it is observed in terms of crime and what is posted on social media, was made, also showing that social media is more closely related to fear of crime than it is to crime itself.

8.6.1 Classification of the tweets

Considering tweets to be related to crime simply because they include a crime-related word has both advantages and disadvantages. Firstly, since the list of words is long (so that almost every word which has any connotation or relation to crime is included) except perhaps for the case of typos or unknown terms and abbreviations, the method should pick up all relevant posts. However, not all of the crime-related tweets are related to crime, but could simply use the same of the language. Manually reading 3,000 crime-related tweets allowed measuring the size of this type of "error" and nearly one-third of the crime-related tweets are clearly not related to crime. However, even with a manual inspection of the tweets, it is very complicated to determine, in many, cases whether a tweet is related to a crime or not or tu understand the context under which the user posted a tweet (for instance, a post with a URL link required further investigation to detect that the user was unhappy about the result of a football match). Perhaps, as it has been suggested before in the case of the measure of happiness in social media (Dodds et al., 2011), requiring two words instead of only one could potentially improve the results.

8.6.2 Expressions of fear of crime at country and at city level

Most of the tweets collected (on average 984 per 1,000) were not related to crime. The crime-related tweets rate is different in each country: in Venezuela, 38 of their tweets per 1,000 are related to crime, but in Bolivia, less than 8 tweets per 1,000 are related to crime.

The analysis of the tweets show that countries with a higher number of murders, or a higher murder rate, have more crime-related tweets per

1,000, but this was not significant at a city level for the data collected, and so murders increase the engagement with social media (termed activism) at a national level but not necessarily at a city level. Thus, a person from Colombia, for instance, is more likely to post a murder-related tweet due to a murder in their own country than in Peru, say. Events nearer to home are more relevant than those which are more distant (Tobler, 1970).

The closer the person lives to a murder, the more likely that they will tweet about it (Kounadi et al., 2015). However, in most Twitter data, it is not possible to differentiate between tweets from different cities. Comparing the number of tweets collected at a national level with the ones who are assigned to a specific city, the frequency of crime-related tweets decreases considerably. In Mexico, for instance, there are 10.68 tweets related to crime per 1,000 at a national level, but it drops to 5.98 per 1,000 when the user shares their location and so the publication can be assigned to a particular city. In Colombia, the number of crime-related tweets per 1,000 drops from 24.64 at a national level to 5.13 per 1,000 at a city level and in Venezuela, this drop is from 38.14 drops to 21.85. In every country, there is a considerable decrease of tweets related to crime per 1,000 when posts are analysed at a city level which means that users who share their location are less likely to post crime-related tweets.

8.6.3 Expressions of violence in social media

The most frequent crime-related tweet is associated with violence, and in particular, with murder. Very similar to what occurs with the traditional media, violence is the most frequent reason why a person tweets a crime-related post, although it is not the most frequent type of crime suffered in the region. Traditional media shows a strong bias towards violent crime and the evidence here is that social media displays a similar bias towards violent crime, possibly even more pronounced, with tweets not reflecting the overall crime levels. For instance, in Mexico, for every murder, there are 917 property crimes but yet the same ratio for tweets on social media is that

for every tweet related to murder there are 0.34 tweets related to property crimes, so that a murder is 2,653 times more likely to be associated to a Twitter publication than a property crime is.

8.6.4 Activism in social media

Twitter is incredibly useful for spreading information or personal views, but it is not perfectly representative as it does not provide a random sample of the population since not everyone uses the service (for example, less than 10% of the population in the case of Mexico and likely to suffer from an age bias). Even among those who do use Twitter, tweets are not a random sample of opinions from its users since some accounts post dozens of tweets per day, whilst others post with a much lower frequency (Kwak et al., 2010). It is noted that those people with extreme opinions are often opposed to a change of mind (Toscani, 2006) and tend to be more expressive of their feelings so that users with the highest fear of crime, or the ones with the highest assertiveness in the case of activists (Düring and Wolfram, 2015; Kacperski and Holyst, 1999), might be more likely to post crime-related tweets, which in turn, skews the information collected.

Turning to the actual content of tweets, again people with extreme opinions tend to be more expressive in their posts, which in terms of crime, means that users with the highest fear of crime, or activists who are promoting a particular view, are perhaps more likely to post tweets which in this study get flagged as being related to crime. The collection of tweets thus do not express a generalised feeling from all the users and should not be treated as such.

Social media posts are thus not a true reflection of the crime in a country or a city, but they do provide evidence of trends affecting those who are actively involved with, in this case, Twitter.

Without performing a sentiment analysis on the individual tweets (Pak and Paroubek, 2010), only considering them as “relevant” or related to a specific type of crime simply by the inclusion of specific words, little more

than 1.5% of the tweets were considered to be relevant. The percentage of crime-related tweets (or, per 1,000 as numbers get considerably small) has been shown to be correlated with the number of murders of the country, the murder rate and, to a stronger degree, with the fear of crime in the country. This means that, in the first place, a large part of the observed crime-related tweets are users with an active opinion of the insecurity in their country, termed “activists.

The level of activism in a country, or engagement in public affairs, can be measured by the number of tweets that a murder in the country “causes” (although, it is not possible to ensure that the murders are the only cause for the murder-related tweets). It was found that in Uruguay and Chile, each murder is associated with more than 60 murder-related tweets, but it drops to less than three tweets in some other countries. A higher level of activism in social media, measured as the number of murder-related tweets which are the result of a murder, was found in countries with a higher GDP per capita and, in general, a higher standard of living.

Taking into account that the vast majority of the relevant tweets are related to murder and security or justice claims, a decrease during the weekends was found in the number of tweets per 1,000 related to crime, murder or violence seems to indicate a decrease in the “activism” in social media.

8.6.5 Population size matters

As expected, there are more crime-related tweets in countries with a higher murder rate. However, countries with a higher number of murders (not the murder rate) also have more crime-related tweets, meaning that size matters. Consider, for instance, a country with the population size of Mexico (approximately 130 million people) and a country with the size of Uruguay (approximately 3.3 million people), then even if both countries suffered the same murder rate, indicating the same risk for their population, Mexico would experience 39 times more murders simply due to the difference in population size. However, if the large country has more murders, only due

to its size, this would also mean that it would have more crime-related tweets and would have more social media activism.

8.6.6 Publications and fear of crime

At a national level, the highest observed correlation between the number of crime-related tweets is not with the actual crime suffered by the population, the number of murders or the murder rate, but it is with the fear of crime. From a crime science perspective, the content of the tweets and not just the location, as has been used before (Malleson and Andresen, 2015), might have a negligible value unless the focus was on the well-being of a community which had a perception of insecurity.

Social media can hence still provide a valuable source of information. Traditional ways of measuring fear of crime strongly rely on victimisation surveys, but there is often a delay of months or even years between when the data is collected and when the information and the results of the surveys are published. Besides, surveys might not be comparable between different countries, since different questions may be asked and with different interpretations. However, by measuring the number of crime-related tweets posted, it is possible to obtain a proxy to the fear of crime of the population of a country. By systematically storing and classifying tweets related to crime, a fairly good proxy to determine the fear of crime can be constructed, that is both timely (with almost no time delay between the time when they are posted in social media) and also economically efficient (at a negligible cost compared to the costly surveys). This social media expression of fear of crime could be valuable to detect people's reaction to a highly reported crime, for instance, or to detect the speed at which memory of past events is lost (Prieto Curiel and Bishop, 2017).

Chapter 9

Discussion and conclusions

Within this thesis, the emergence of certain social patterns has been analysed using a mathematical framework, from migration, victimisation, fear of crime and road accidents. In all cases, the patterns which emerge are detected by looking at collective individual behaviours, meaning that the reasons why a road accident occur, why a specific person experiences fear of crime or why a particular person decided to move, is not considered, but instead, it is by looking at millions of migrants, thousands of crimes or accidents that patterns are observed, revealing a collective behaviour.

In this thesis, mathematical models have been introduced to consider different social challenges. Whether it is city to city migration, crime patterns, fear and activism and road accidents, mathematical models help to construct a system that allows us to understand better the structural patterns which are observed and which emerge as a result of our social behaviour.

9.1 Social modelling

9.1.1 Migration patterns

People migrate from one place to another for various reasons. Their decision is made individually but when a large number of people is analysed, a pattern is detected in which the size of the city in which a person lives plays a role in their decision and choice. Using data from the US census,

it was determined that people from small cities are more likely to migrate than people from large cities and they are more likely to select a destination about the same size than their original city of residence. However, this pattern changes for cities with between 1 and 5 million inhabitants. The scaling properties of migration between cities of different sizes as well as some patterns of international migration were analysed within a mathematical framework.

9.1.2 Crime patterns

In terms of the crime suffered by an individual, it is worth noting that a person might suffer more than one crime during the time interval being considered, so that a binary model, dividing people into victims and non-victims does not give the whole picture. Also, there is a random element to the number of crimes suffered and so, even when an individual has a high/low risk of suffering a crime, they might instead experience a small/high number of crimes simply by chance. Assuming a Poisson distribution for the number of crimes suffered by an individual has some disadvantages, for instance, the number of crimes suffered by some individuals might be correlated, but the distribution takes into account a random component and so, for example, an individual with a rate $\lambda = 1$, which means that they expect to suffer one crime per year, has, in fact, a 0.368 probability of suffering no crimes and a 0.264 probability of suffering more than one crime.

Assuming a Poisson distribution for the number of crimes suffered by a person also allows considering distinct rates for different individuals or groups of individuals. The methodology presented provides a way to consider different distributions of crime in a population and its relation to the fear of crime while keeping the expected number of crimes suffered by the population as a whole, constant. Thus, it provides a methodology to simulate different distributions of crime, which consider an immune population group and the fact that crime is, in general, rare and highly concentrated.

9.1.3 Fear of crime patterns

A mathematical model to simulate crime and measure its degree of concentration, was then used in a model based on a model of opinion dynamics to explain why fear of crime emerges as a complex phenomenon in society, why fear is observed in regions and population groups which suffer no crime and why a decrease of crime might not reduce the levels of fear.

Frequently, fear of crime is assumed to be only as a direct consequence of suffering crime and thus, it is thought that by lowering the levels of crime, fear will consequently also be reduced. However, this might not be the case. Fear of crime is a problem in its own right, with costly and long-lasting consequences to the social life of a city and therefore, understanding its causes and the reasons why it emerges as a social phenomenon plays a key role in the correct design of policies.

9.1.4 Road accidents patterns

Unfortunately, road accidents are one of the most frequent causes around the world. Extending the methodology developed for simulating crime in a population, the distribution and the concentration of road accidents can be described, using data from the central area of London and on the main motorways which link with Mexico City to benchmark parameters. Obtaining the distribution of road accidents and a measure of their degree of concentration allows designing better policies and to make evidence-based decisions: to detect whether accidents occurred due to environmental factors (and plan a road intervention, for instance) or related to the driver (and design a speed-reduction or alcohol consumption campaign).

9.2 Mathematics and social challenges

The research presented in this thesis has been motivated by a variety of social challenges and therefore, certain challenges with respect to the social aspect were encountered.

9.2.1 Interdisciplinary research

Perhaps the biggest challenge faced to produce this thesis was the interdisciplinary type of research needed. In the case of human migration, for instance, a mathematician is perhaps more interested in the driving forces and the emergent patterns of migration, modelling flows with a model, challenge it through data and compare against other state or the art models. However, migration is a sensitive topic, which has political connotations and the fact that migrants are humans, not birds, who in some cases are undergoing severe circumstances, cannot be overlooked. Migration requires the power of mathematics, but at the same time, statisticians, economists, demographers, lawyers, human rights specialists and social scientists. A single discipline alone has little to contribute to the significant challenges of the world.

Similarly, the analysis of crime or fear of crime patterns, requires crime scientists, psychologists, sociologists, urban planners, policymakers to work on the interface between these and many more disciplines.

Interdisciplinary collaborations pose several challenges. Communicating ideas and basic concepts between different fields of study is the first challenge, but moving ideas forward with modelling techniques and drawing conclusions from mathematical models is even more challenging when different disciplines meet.

Beyond the challenging aspect of interdisciplinary research, other fields of study outside mathematics and physical sciences provide mathematicians with challenges. The same way in which fluid mechanics becomes a relevant aspect of mathematical research, social sciences encourage areas of research in which mathematicians have something to contribute. The study of social networks and big data techniques, spatial analysis, finance, sports modelling, opinion dynamics, political support and polarisation, for instance, are areas of research which begin outside physical sciences and mathematics but have encouraged advancements on the mathe-

matical spectrum.

9.2.2 Continued research

Research is a continuous process. Particularly, for the research conducted to produce this thesis, *'answering' one question 'created' two more* and so there are still many challenges with respect to every piece of research presented here.

9.2.2.1 In terms of migration

The scaling and the gravity-scaling models of migration were tested using data from the US census. Testing the same model in different regions of the world (for instance, in China, with more than 20 cities with a population larger than 5 million inhabitants) would allow internal migration to be tested. Differences might not reveal errors in the model but fundamentally different social behaviour between the two cultures. Data from the European Union could perhaps be used for measuring international migration between cities, answering very topical questions of the day.

By considering, perhaps, many years of migration data, is it possible to model, for an individual, subsequent migrations with a Poisson distribution? Then, is migration a rare event, with low frequency and a high level of concentration? Is it possible to construct the *migration profile* of a population? Is the concentration of the migration rates, similar across different countries and stable for many years, or is it possible to detect changes in the migration patterns?

Also, ignoring population growth, is the model of migration between cities capable of revealing a stable distribution for the number of inhabitants within a city? And, if so, is that distribution simply a scaled version in different regions of the world, following, for example, Zipf's law? If within the next century the world will reach its stable population, will we also reach a stable distribution of the population within our cities? The thesis opens the way to study these questions and more.

9.2.2.2 In terms of crime and road accidents

The model used for constructing the victimisation profile of the population was based on data from the Netherlands (burglaries) and from Mexico (robberies) and it is based on the number of crimes suffered by the population (either homes or individuals). Beyond comparing other types of crime and other regions of the world, with different geographical units, there are other challenges left open. What about the *criminality profile*, that is, the number of crimes committed by each person? If one city or country suffers less crime than the other, is it because they have a smaller number of criminals or because a similar portion of criminals are less active? Is it possible to use the same technique used for the spatial concentration of road accidents for the spatial distribution of crimes? Is it possible to use the spatial victimisation profile for simulating crimes or burglaries and detect outliers?

9.2.2.3 In terms of the fear of crime

Fear of crime was measured at a regional level and was then tracked through social media. Furthermore, a model for the fear of crime, based on opinion dynamics, was constructed. This revealed certain patterns in the way fear of crime emerges in a population and is shared among individuals, with reasons going beyond directly suffering a crime. There are, however, many more challenges to tackle. Social interactions and links between different individuals (either on social media or on real life) tend to have clusters, hubs and have small-world properties. What is the impact of the network topology on the dynamics of fear? What would be different in the dynamics if an Erdös-Rényi, a Watts-Strogatz or a Barabási-Albert model is considered? Is the connectivity between nodes working as a friction or an accelerating factor for how fear is shared? Also, is it possible to construct a model based on a mean-field approach, rather than simulating agents? These thesis works is a start in the process to answer these questions.

9.2.2.4 In terms of the applications of the models

A model for the fear of crime which considers the actual crime suffered by individuals and its distribution was constructed. Is it possible to extend the model and apply it to the fear of terrorism? or to use a similar technique to model attitudes towards the immigrants? Or is it possible to detect internet safety precautions and and its relationship with scams suffered by the users? Again, this thesis works is a start in the process to answer these questions.

9.2.3 Having an impact

Mathematical modelling of social systems and conducting interdisciplinary research produces results which are interest beyond the scholarly researcher. In terms of human mobility and migration, results show there is a high frequency of internal migration. A person in the US moves, from one county to another, several times in their lifetime and this may be in other parts of the world. Whether a person moves looking for a better job, the love of their life or a place for retirement, projections show that internal and international migration will become more frequent and therefore, to understand the scaling of migration plays a fundamental role on the design of mobility policies.

In terms of crime, the construction of the victimisation profile considers the random aspect of suffering a crime, thus, is a useful tool for simulating crime and determining whether a policy intervention has resulted on crime displacement or deterrence.

Fear of crime has a direct impact in the life of a person who experiences crime and therefore, having the correct tools for measuring fear and understanding how it is shared among individuals allows policies to be designed in terms of fear and not just directly in terms of crime.

9.3 Conclusions

There are perhaps subtle relationships between different, seemingly discon-

nected social problems. Crime and fear of crime have caused the displacement of millions of people from one city to another or, in some cases, to another country. The World Bank forecasts that by 2050 more than 140 million individuals from certain developing parts of the world (including Latin America, Sub-Saharan Africa and South Africa) will be forced to migrate within their own countries to escape the impacts of climate change. Rural to urban and intra-urban migration is one of the reasons why informal settlements from around the world are growing, today housing more than one billion people globally. The emergence of irregular settlements increases the risk and vulnerability of the population and reduces their resilience and so geological events, such as earthquakes or storms, become disasters which can practically wipe towns off the map. The spread of misinformation causes, among other social issues, echo chambers that are difficult to change, themselves causing prejudice and segregation, and, to a certain extent, misinformation is also one of the reasons why fear or crime is shared and why opinions are becoming more polarised. The analysis of the complex connections between different social issues amplifies our vision of society.

Mapping and transforming social challenges into mathematical models gives the ability to use quantitative tools, simulate and consider 'what would happen if' scenarios, which can be then transformed into tools for policy-making.

In the case of city to city migration, a different pattern is observed for individuals who live in small or large cities. Whilst migration is the main driver for city changes, it potentially accelerates the disparities between small and large cities.

Crime concentrates, but a considerably large part of this concentration is the result of randomness, which is often ignored. Thus, policies oriented to prevent and reduce crime, such as increased local policing, needs to consider the uncertainty and instability of social patterns. A similar phenomenon

occurs when looking at the spatial pattern of road accidents, where an intervention to reduce accidents on a specific road junction might have a negligible effect if some of the road accidents at that junction happen due to random elements. Understanding the concentration and stability of social patterns help designing better policies.

Fear of crime emerges, but it is not necessarily the result of direct victimisation. Although a certain degree of fear of crime encourages individual healthy precautions, understanding why fear of crime emerges needs further study. In a similar manner to misinformation, which is shared as fake news and still believed even when discredited, fear of crime might not be reduced even if crime rates are substantially reduced, showing that fear of crime needs to be analysed, quantified and understood with multidisciplinary techniques to gain a better understanding of this social problem.

Better policies, improving our social well-being and tackling the significant challenges of the 21st century requires the tools and insights from different disciplines and sciences and a multidisciplinary approach, but one in which mathematics is at the core since it provides quantitative and verifiable results.

Mathematics would thus be seen not only as the language of nature, but also the tool for solving our social problems.

Appendix A

Publications

The following papers have been published:

- Prieto Curiel, R. and Bishop, S. (2016a). A measure of the concentration of rare events. Scientific Reports, 6(1):16. (Prieto Curiel and Bishop, 2016a).

- Prieto Curiel, R. and Bishop, S. R. (2016b). A metric of the difference between perception of security and victimisation rates. Crime Science, 5(1) (Prieto Curiel and Bishop, 2016b).

- Prieto Curiel, R., Collignon Delmar, S., and Bishop, S. R. (2017). Measuring the distribution of crime and its concentration. Journal of Quantitative Criminology, pages 129. (Prieto Curiel et al., 2017a).

- Prieto Curiel, R., Heinrigs, P., Heo, I. (2017). Cities and Spatial Interactions in West Africa. West African Papers (OECD) (Prieto Curiel et al., 2017b)

- Prieto Curiel, R., and Bishop, S. R. (2017). Modelling the fear of crime. Proceedings of the Royal Society A (Prieto Curiel and Bishop, 2017).

- Farnham, A. and others (2017). Early career researchers want Open Science. Genome Biology 18 (1), 221 (Farnham et al., 2017)

- Prieto Curiel, R., and Bishop, S. R. (2018). Fear of crime: the impact of different distributions of victimisation. Palgrave Communications 4(1):46-54 (Prieto Curiel and Bishop, 2018)

- Prieto Curiel, R., Pappalardo, L. Gabrielli, L. and Bishop, S. R. (2018). Gravity and scaling laws of city to city migration. Plos One (Prieto Curiel et al., 2018b)

- Prieto Curiel, R., Gonzlez Ramrez, H. and Bishop, S. R. (2018). A novel rare event approach to measure the randomness and concentration of road accidents. Plos One (Prieto Curiel et al., 2018a)

The following papers have been submitted:

- Crime, fear of crime and activism in social media

  Rafael Prieto Curiel, Stefano Cresci, Cristina Muntean and Steven R. Bishop.

- The dimensions of external and internal migration in West Africa

  Rafael Prieto Curiel, Luca Pappalardo and Lorenzo Gabrielli.

- Temporal and spatial analysis of the media spotlight

  Rafael Prieto Curiel, Carmen Cabrera Arnau, Mara Torres Pinedo, Humberto González Ramírez and Steven R. Bishop.

Appendix B

Tables of coefficients

B.1 Human migration

Outflow	$\hat{\beta}$	$s.e.$	$\hat{\log\alpha}$	$s.e.$	Adj. R^2
Leaving the city	0.8829	0.0147	-1.7863	0.1867	0.9033
The countryside	0.6846	0.0273	-0.7214	0.3464	0.6199
City with **less** than					
200,000 inhabitants	0.8060	0.0263	-2.8555	0.3325	0.7101
300,000 inhabitants	0.8199	0.0232	-2.6039	0.2949	0.7636
400,000 inhabitants	0.8171	0.0222	-2.285	0.2811	0.7794
500,000 inhabitants	0.8224	0.0206	-2.1614	0.2607	0.8061
1,000,000 inhabitants	0.8363	0.0175	-1.9163	0.2224	0.8554
3,000,000 inhabitants	0.8609	0.0159	-1.8352	0.2021	0.8863
5,000,000 inhabitants	0.8689	0.0157	-1.8189	0.1999	0.8877
City with **more** than					
1,000,000 inhabitants	0.9570	0.0211	-3.4759	0.2674	0.8426
3,000,000 inhabitants	1.0073	0.0307	-4.8421	0.3891	0.7369
5,000,000 inhabitants	1.0499	0.0337	-5.8437	0.4271	0.7163
8,000,000 inhabitants	1.1688	0.0506	-8.5108	0.6408	0.5814
10,000,000 inhabitants	1.2984	0.0619	-10.6921	0.7855	0.5327

Table B.1: Coefficients obtained for the outflow of internal migration in the US. The estimation of the parameters uses a logarithmic transformation on both sides of equation 2.1 and so the results for $\hat{\alpha}$ are expressed in terms of its natural logarithm.

Inflow	$\hat{\beta}$	*s.e.*	$\hat{\log\alpha}$	*s.e.*	Adj. R^2
The countryside	0.5971	0.0342	0.4299	0.4331	0.4421
City with **less** than					
200,000 inhabitants	0.7997	0.0299	-2.7992	0.3799	0.6492
300,000 inhabitants	0.8036	0.0277	-2.4173	0.3507	0.6869
400,000 inhabitants	0.8110	0.0262	-2.2685	0.3315	0.7144
500,000 inhabitants	0.8159	0.0245	-2.1231	0.3103	0.7429
1,000,000 inhabitants	0.8331	0.0211	-1.9238	0.2678	0.8018
3,000,000 inhabitants	0.8437	0.0203	-1.6504	0.2570	0.8184
5,000,000 inhabitants	0.8521	0.0202	-1.6248	0.2555	0.8230
City with **more** than					
1,000,000 inhabitants	0.9193	0.0251	-2.8957	0.3177	0.7777
3,000,000 inhabitants	0.9776	0.0346	-4.2905	0.4390	0.6744
5,000,000 inhabitants	1.0184	0.0382	-5.2401	0.4851	0.6480
8,000,000 inhabitants	1.1180	0.0460	-7.4146	0.5834	0.6053
10,000,000 inhabitants	1.2539	0.0574	-9.6531	0.7272	0.5538
International	1.1884	0.0339	-7.9153	0.4305	0.761
Africa	1.5794	0.0728	-16.5087	0.9230	0.5500
Asia	1.2207	0.0519	-9.3143	0.6588	0.5891
Americas (outside US)	1.2808	0.0424	-10.5175	0.5374	0.7036
Europe	1.2264	0.0515	-10.2535	0.6530	0.5956
Oceania	1.3312	0.0747	-14.5814	0.9465	0.4520

Table B.2: Coefficients obtained for the inflow of migration to the cities and the countryside in the US.

B.2 Victimisation profile in Mexico

The victimisation profile for all years and all states is presented here.

states in Mexico 2015		group 1	2	3	4	5	6	7
Aguascalientes	q_j	0.84	0.16	0.00				
	λ_i	0.00	0.23	2.40				
Baja California	q_j	0.64	0.26	0.10	0.01			
	λ_i	0.00	0.04	0.47	2.65			
Baja California Sur	q_j	0.87	0.13					
	λ_i	0.00	0.15					
Campeche	q_j	0.96	0.04					
	λ_i	0.00	0.85					
Coahuila	q_j	0.91	0.09	0.00				
	λ_i	0.00	0.42	1.75				
Colima	q_j	0.62	0.35	0.02				
	λ_i	0.00	0.01	0.43				
Chiapas	q_j	0.90	0.01	0.09				
	λ_i	0.00	0.32	0.36				
Chihuahua	q_j	0.70	0.30	0.01				
	λ_i	0.00	0.11	1.27				
Ciudad de Mexico	q_j	0.35	0.07	0.56	0.03			
	λ_i	0.00	0.04	0.31	1.36			
Durango	q_j	0.89	0.11	0.00				
	λ_i	0.00	0.39	5.18				
Guanajuato	q_j	0.38	0.61	0.01				
	λ_i	0.00	0.11	2.86				
Guerrero	q_j	0.80	0.10	0.10	0.01			
	λ_i	0.00	0.25	0.26	2.09			
Hidalgo	q_j	0.03	0.95	0.02				
	λ_i	0.00	0.03	0.83				
Jalisco	q_j	0.11	0.82	0.07	0.00			
	λ_i	0.00	0.04	0.49	2.58			
Estado de Mexico	q_j	0.45	0.38	0.17	0.00			
	λ_i	0.00	0.30	1.55	7.83			
Michoacan	q_j	0.10	0.88	0.03				
	λ_i	0.00	0.02	0.92				
Morelos	q_j	0.71	0.03	0.26	0.00			
	λ_i	0.00	0.38	0.40	3.53			
Nayarit	q_j	0.20	0.78	0.02				
	λ_i	0.00	0.01	0.76				
Nuevo Leon	q_j	0.82	0.18					
	λ_i	0.00	0.44					
Oaxaca	q_j	0.35	0.62	0.03				
	λ_i	0.00	0.06	1.23				
Puebla	q_j	0.01	0.94	0.05				
	λ_i	0.00	0.04	0.97				
Queretaro	q_j	0.90	0.10	0.00				
	λ_i	0.00	0.40	3.87				
Quintana Roo	q_j	0.12	0.83	0.05	0.00			
	λ_i	0.00	0.03	0.81	2.24			
San Luis Potosi	q_j	0.00	0.98	0.02				
	λ_i	0.00	0.03	1.70				
Sinaloa	q_j	0.85	0.15	0.00				
	λ_i	0.00	0.33	7.73				
Sonora	q_j	0.90	0.10	0.00				
	λ_i	0.00	0.30	2.32				
Tabasco	q_j	0.02	0.97	0.02	0.00			
	λ_i	0.00	0.06	1.21	3.48			
Tamaulipas	q_j	0.77	0.23	0.00				
	λ_i	0.00	0.20	4.82				
Tlaxcala	q_j	0.80	0.20					
	λ_i	0.00	0.39					
Veracruz	q_j	0.74	0.26	0.00				
	λ_i	0.00	0.12	2.53				
Yucatan	q_j	0.92	0.08	0.00				
	λ_i	0.00	0.33	1.68				
Zacatecas	q_j	0.94	0.06					
	λ_i	0.00	0.55					

immune population size
immune rate $\lambda_1 = 0$
chronic population size
chronic rate $\lambda_j > 2$

Figure B.1: Victimisation profile (individual crime rates $\underline{\lambda}$ and group sizes $\underline{q}$) for the 32 states in Mexico in 2015.

states in Mexico 2016		group 1	2	3	4	5	6	7
Aguascalientes	q_i	0.81	0.19					
	λ_i	0.00	0.22					
Baja California	q_i	0.52	0.39	0.09				
	λ_i	0.00	0.02	0.28				
Baja California Sur	q_i	0.65	0.32	0.03	0.00			
	λ_i	0.00	0.02	0.45	2.65			
Campeche	q_i	0.87	0.02	0.11				
	λ_i	0.00	0.25	0.28				
Coahuila	q_i	0.92	0.01	0.07	0.00			
	λ_i	0.00	0.39	0.40	3.73			
Colima	q_i	0.73	0.27	0.00				
	λ_i	0.00	0.06	2.61				
Chiapas	q_i	0.01	0.98	0.01				
	λ_i	0.00	0.03	1.67				
Chihuahua	q_i	0.87	0.04	0.08	0.00			
	λ_i	0.00	0.27	0.28	2.50			
Ciudad de Mexico	q_i	0.25	0.73	0.01				
	λ_i	0.00	0.27	1.68				
Durango	q_i	0.11	0.84	0.05				
	λ_i	0.00	0.02	0.53				
Guanajuato	q_i	0.10	0.78	0.13				
	λ_i	0.00	0.03	0.44				
Guerrero	q_i	0.87	0.12	0.02				
	λ_i	0.00	0.48	0.90				
Hidalgo	q_i	0.83	0.17	0.00				
	λ_i	0.00	0.16	1.96				
Jalisco	q_i	0.02	0.93	0.05				
	λ_i	0.00	0.04	1.06				
Estado de Mexico	q_i	0.00	0.97	0.04				
	λ_i	0.00	0.23	2.10				
Michoacan	q_i	0.79	0.21	0.00				
	λ_i	0.00	0.14	1.77				
Morelos	q_i	0.63	0.03	0.31	0.03			
	λ_i	0.00	0.23	0.25	1.12			
Nayarit	q_i	0.01	0.99	0.00				
	λ_i	0.00	0.02	2.13				
Nuevo Leon	q_i	0.06	0.90	0.04				
	λ_i	0.00	0.03	0.53				
Oaxaca	q_i	0.26	0.66	0.09				
	λ_i	0.00	0.02	0.60				
Puebla	q_i	0.47	0.52	0.01				
	λ_i	0.00	0.11	1.16				
Queretaro	q_i	0.46	0.41	0.13				
	λ_i	0.00	0.03	0.26				
Quintana Roo	q_i	0.10	0.80	0.10				
	λ_i	0.00	0.03	0.46				
San Luis Potosi	q_i	0.03	0.96	0.01				
	λ_i	0.00	0.02	1.11				
Sinaloa	q_i	0.83	0.17	0.00				
	λ_i	0.00	0.18	6.80				
Sonora	q_i	0.87	0.13	0.00				
	λ_i	0.00	0.31	6.79				
Tabasco	q_i	0.04	0.91	0.06				
	λ_i	0.00	0.05	0.45				
Tamaulipas	q_i	0.15	0.84	0.02				
	λ_i	0.00	0.02	0.65				
Tlaxcala	q_i	0.85	0.15	0.00				
	λ_i	0.00	0.36	3.50				
Veracruz	q_i	0.19	0.77	0.04				
	λ_i	0.00	0.02	0.44				
Yucatan	q_i	0.91	0.01	0.08				
	λ_i	0.00	0.26	0.30				
Zacatecas	q_i	0.77	0.13	0.11				
	λ_i	0.00	0.04	0.08				

immune population size
immune rate $\lambda_1 = 0$
chronic population size
chronic rate $\lambda_j > 2$

Figure B.2: Victimisation profile (individual crime rates $\underline{\lambda}$ and group sizes $\underline{q}$) for the 32 states in Mexico in 2016.

B.3 Fear of crime in social media

	crime-related tweets per 1,000			violence-related tweets per 1,000			murder-related tweets per 1,000			property-crime-related tweets		
	coefficient	std. error	p-value	coefficient	std. error	p-value	coefficient	std. error	p-value	coefficient	std. error	p-value
murders	intercept 1.31E+01	2.04E+00	8.35E-06	intercept 4.68E+00	8.13E-01	2.95E-05	intercept 2.98E+00	4.92E-01	1.67E-05	intercept 1.40E+00	2.53E+00	4.57E-05
	slope 6.18E-04	2.69E-04	3.53E-02	slope 3.59E-04	1.07E-04	3.90E-03	slope 2.02E-04	6.40E-05	6.67E-03	slope 6.22E-05	3.33E-05	8.01E-02
	adj. R-squared	0.2013		adj. R-squared	0.3773		adj. R-squared	0.3386		adj. R-squared	0.1279	
murders per 100,000	intercept 1.22E+01	2.37E+00	9.22E-05	intercept 4.36E+00	1.01E+00	5.00E-04	intercept 2.78E+00	5.97E-01	3.00E-04	intercept 1.36E+00	2.99E-01	3.34E-04
	slope 1.69E-01	7.75E-02	4.50E-02	slope 8.90E-02	3.29E-02	1.57E-02	slope 5.13E-02	1.96E-02	1.85E-02	slope 1.49E-02	9.78E-03	1.47E-01
	adj. R-squared	0.1798		adj. R-squared	0.2703		adj. R-squared	0.2567		adj. R-squared	0.072	
fear of crime index	intercept -1.19E+01	9.83E+00	2.46E-01	intercept -3.96E+00	4.88E+00	4.29E-01	intercept -1.55E+00	2.95E+00	6.07E-01	intercept -1.09E+00	1.25E+00	3.97E-01
	slope 1.96E+01	6.79E+00	1.13E-02	slope 7.20E+00	3.37E+00	4.94E-02	slope 3.81E+00	2.04E+00	8.12E-02	slope 1.96E+00	8.62E-01	3.82E-02
	adj. R-squared	0.3144		adj. R-squared	0.1825		adj. R-squared	0.1349		adj. R-squared	0.2065	
% of people with strong fear	intercept 2.07E+00	3.29E+00	5.39E-01	intercept 6.75E-01	1.71E+00	6.99E-01	intercept 7.12E-01	1.04E+00	5.05E-01	intercept 2.71E-01	4.52E-01	5.59E-01
	slope 7.91E+01	1.70E+01	3.17E-04	slope 3.17E+01	8.86E+00	2.70E-03	slope 1.79E+01	5.40E+00	4.70E-03	slope 8.10E+00	2.34E+00	3.48E-03
	adj. R-squared	0.5626		adj. R-squared	0.4248		adj. R-squared	0.3832		adj. R-squared	0.4072	

Figure B.3: Table of coefficients of the linear regression using the number of crime-related tweets per 1,000 collected on each country as dependent variable and crime rates, number of crimes and fear of crime as independent variables. The dependent variable is the number of crime-related, violence-related, murder-related and property-crime-related tweets per 1,000 and the independent variables are, for each model, the number of murders, the murder rate, the fear of crime index and the % of the population with a strong fear. The significant coefficients for the slope (that is, with a p-value smaller than 0.05) are highlighted in red.

	crime-related tweets per 1,000			violence-related tweets per 1,000			murder-related tweets per 1,000			property-crime-related tweets		
	coefficient	std. error	p-value	coefficient	std. error	p-value	coefficient	std. error	p-value	coefficient	std. error	p-value
murders	intercept 3.60E+00	5.17E-01	9.37E-07	intercept 1.54E+00	1.69E-01	1.36E-08	intercept 1.08E+00	1.31E-01	7.58E-08	intercept 2.33E-01	1.47E-01	1.27E-01
	slope 2.58E-04	1.61E-04	1.24E-01	slope 3.28E-05	5.23E-05	5.38E-01	slope 3.16E-05	4.07E-05	4.47E-01	slope 5.40E-05	4.55E-05	2.50E-01
	adj. R-squared	0.0699		adj. R-squared	-0.0298		adj. R-squared	-0.01927		adj. R-squared	0.01891	
murders per 100,000	intercept 3.42E+00	5.15E-01	1.81E-06	intercept 1.59E+00	1.76E-01	1.67E-08	intercept 1.06E+00	1.35E-01	1.49E-07	intercept 2.58E-02	9.67E-02	7.92E-01
	slope 4.86E+02	2.31E+02	4.79E-02	slope -1.05E+00	7.89E+01	9.89E-01	slope 4.05E+01	6.04E+01	4.14E-01	slope 2.48E+02	4.33E+01	1.32E-05
	adj. R-squared	0.1409		adj. R-squared	-0.0499		adj. R-squared	-0.01469		adj. R-squared	0.6023	
hard crimes	intercept 3.63E+00	5.69E-01	3.16E-06	intercept 1.51E+00	1.79E-01	4.89E-08	intercept 1.07E+00	1.40E-01	2.44E-07	intercept 2.36E-01	1.59E-01	1.53E-01
	slope 5.95E-05	5.26E-05	2.71E-01	slope 1.38E-05	1.65E-05	4.13E-01	slope 9.59E-06	1.29E-05	4.66E-01	slope 1.30E-05	1.47E-05	3.85E-01
	adj. R-squared	0.01316		adj. R-squared	-0.01449		adj. R-squared	-0.02181		adj. R-squared	-0.01012	
hard crimes per 100,000	intercept 3.43E+00	6.91E-01	7.39E-05	intercept 1.52E+00	2.19E-01	1.01E-06	intercept 1.03E+00	1.70E-01	6.03E-06	intercept -1.26E-01	1.43E-01	3.89E-01
	slope 1.03E+02	9.29E+01	2.82E-01	slope 1.42E+01	2.95E+01	6.36E-01	slope 1.70E+01	2.28E+01	4.64E-01	slope 8.11E+01	1.92E+01	4.10E-04
	adj. R-squared	0.01037		adj. R-squared	-0.03802		adj. R-squared	-0.02155		adj. R-squared	0.446	
robbery of a person	intercept 3.06E+00	7.74E-01	7.93E-04	intercept 1.29E+00	2.38E-01	2.66E-05	intercept 8.73E-01	1.84E-01	1.27E-04	intercept 4.63E-03	2.08E-01	9.82E-01
	slope 1.07E+01	7.13E+00	1.48E-01	slope 3.48E+00	2.19E+00	1.29E-01	slope 2.87E+00	1.70E+00	1.06E-01	slope 3.56E+00	1.91E+00	7.80E-02
	adj. R-squared	0.05667		adj. R-squared	0.06722		adj. R-squared	0.08138		adj. R-squared	0.1045	
robbery of a person per 100,000	intercept 3.08E+00	8.14E-01	1.17E-03	intercept 1.24E+00	2.45E-01	6.12E-05	intercept 8.43E-01	1.91E-01	2.63E-04	intercept 2.74E-02	2.21E-01	9.03E-01
	slope 1.30E+01	9.55E+00	1.87E-01	slope 5.10E+00	2.87E+00	9.12E-02	slope 4.00E+00	2.24E+00	8.92E-02	slope 4.11E+00	2.60E+00	1.29E-01
	adj. R-squared	0.03946		adj. R-squared	0.09285		adj. R-squared	0.09449		adj. R-squared	0.06692	
any crime per 100,000	intercept 1.92E+00	1.66E+00	2.61E-01	intercept 1.15E+00	5.24E-01	4.05E-02	intercept 6.67E-01	4.03E-01	1.13E-01	intercept 2.63E-02	4.71E-01	9.56E-01
	slope 6.39E+00	4.90E+00	2.07E-01	slope 1.37E+00	1.55E+00	3.86E-01	slope 1.40E+00	1.19E+00	2.52E-01	slope 8.89E-01	1.39E+00	5.30E-01
	adj. R-squared	0.03243		adj. R-squared	-0.01026		adj. R-squared	0.01831		adj. R-squared	-0.02901	
fear in the county	intercept 5.57E+00	2.30E+00	2.48E-02	intercept 1.05E+00	7.09E-01	1.55E-01	intercept 7.10E-01	5.53E-01	2.14E-01	intercept -7.55E-01	5.93E-01	2.18E-01
	slope -2.34E+00	3.32E+00	4.89E-01	slope 8.08E-01	1.02E+00	4.39E-01	slope 6.10E-01	7.99E-01	4.54E-01	slope 1.58E+00	8.56E-01	8.00E-02
	adj. R-squared	-0.02453		adj. R-squared	-0.01831		adj. R-squared	-0.02031		adj. R-squared	0.1027	

Figure B.4: Results of the linear models at a city level in Mexico taking the 23 cities from the country as observations. The dependent variable is the number of crime-related, violence-related, murder-related and property-crime-related tweets per 1,000 and the independent variables are, for each model, the number of murders, the murder rate, the number of hard crimes (which include murder, missing person and kidnap) and the hard crime rates, the number of robbery of a person, the robbery of a person rate, any crime rate (that is, all the crimes) and the fear experienced in their county. The significant coefficients for the slope (that is, with a p-value smaller than 0.05) are highlighted in red.

References

Albuja, S. (2014). Criminal violence and displacement in Mexico. *Forced Migration Review*, 1(45):28.

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. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1333–1336. ACM.

Anderson, D. A. (1999). The aggregate burden of crime. *The Journal of Law and Economics*, 42(2):611–642.

Anderson, J. E. (2010). The gravity model. Technical report, National Bureau of Economic Research.

Anderson, T. K. (2007). Comparison of spatial methods for measuring road accident hotspots: a case study of London. *Journal of Maps*, 3(1):55–63.

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

Andresen, M. A., Linning, S. J., and Malleson, N. (2016). Crime at places and spatial concentrations: Exploring the spatial stability of property crime

in Vancouver BC, 2003–2013. *Journal of Quantitative Criminology*, pages 1–21.

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.

Austin, M., Furr, A., and Spine, M. (2002). The effects of neighborhood conditions on perceptions of safety. *Journal of Criminal Justice*, 30(1):417–427.

Avvenuti, M., Cresci, S., Marchetti, A., Meletti, C., and Tesconi, M. (2016). Predictability or early warning: using social media in modern emergency response. *IEEE Internet Computing*, 20(6):4–6.

Baddeley, A. (2010). *Analysing spatial point patterns in R*. CSIRO.

Banisch, S. (2014). From microscopic heterogeneity to macroscopic complexity in the contrarian voter model. *Advances in Complex Systems*, 17(05):1450025.

Banisch, S. and Lima, R. (2015). Markov chain aggregation for simple agent-based models on symmetric networks: the voter model. *Advances in Complex Systems*, 18(05):1550011.

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

Batty, M. (2007). *Cities and complexity: understanding cities with cellular automata, agent-based models, and fractals*. The MIT press.

Batty, M. (2013). *The new science of cities*. The MIT press.

Baudains, P., Braithwaite, A., and Johnson, S. D. (2013). Target choice during extreme events: A discrete spatial choice model of the 2011 London riots. *Criminology*, 51(2):251–285.

Becker, G. S. (1968). *Crime and punishment: An economic approach*, pages 13–68. Palgrave Macmillan UK, London.

Benach, J., Muntaner, C., Delclos, C., Menéndez, M., and Ronquillo, C. (2011). Migration and "low-skilled" workers in destination countries. *PLoS Med*, 8(6):e1001043.

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

Bernasco, W. and Steenbeek, W. (2016). More places than crimes: Implications for evaluating the law of crime concentration at place. *Journal of Quantitative Criminology*, pages 1–17.

Bettencourt, L. M., Cintrn-Arias, A., Kaiser, D. I., and Castillo-Chávez, C. (2006). The power of a good idea: quantitative modeling of the spread of ideas from epidemiological models. *Physica A: Statistical Mechanics and its Applications*, 364(1):513 – 536.

Bettencourt, L. M., 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., 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):e13541.

Bíl, M., Andrášik, R., and Janoška, Z. (2013). Identification of hazardous road locations of traffic accidents by means of kernel density estimation and cluster significance evaluation. *Accident Analysis & Prevention*, 55:265–273.

Birch, C. P., Oom, S. P., and Beecham, J. A. (2007). Rectangular and hexagonal grids used for observation, experiment and simulation in ecology. *Ecological Modelling*, 206(3):347–359.

Böhning, D. (1998). Zero-inflated Poisson models and ca man: a tutorial collection of evidence. *Biometrical Journal*, 40(7):833–843.

Böhning, D., Dietz, E., and Schlattmann, P. (1998). Recent developments in computer-assisted analysis of mixtures. *Biometrics*, pages 525–536.

Böhning, D., Schlattmann, P., and Lindsay, B. (1992). Computer-assisted analysis of mixtures (CA MAN): statistical algorithms. *Biometrics*, pages 283–303.

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, 99(suppl 3):7280–7287.

Borooah, V. K. and Carcach, C. A. (1997). Crime and fear, evidence from Australia. *British Journal of Criminology*, 37(4):635–657.

Bowers, K., Johnson, S., and Pease, K. (2005). Victimisation and revictimisation risk, housing type and area: A study of interactions. *Crime Prevention and Community Safety: An International Jounal*, 7(1):7–17.

Bowers, K. J. and Johnson, S. D. (2003). Measuring the geographical displacement and diffusion of benefit effects of crime prevention activity. *Journal of Quantitative Criminology*, 19(3):275–301.

Box, S., Hale, C., and Andrews, G. (1988). Explaining fear of crime. *The British Journal of Criminology*, 28(3):340–356.

Brame, R., Nagin, D. S., and Wasserman, L. (2006). Exploring some analytical characteristics of finite mixture models. *Journal of Quantitative Criminology*, 22(1):31–59.

Brands, J., Schwanen, T., and Van Aalst, I. (2015). Fear of crime and affective ambiguities in the night-time economy. *Urban Studies*, 52(3):439–455.

Brantingham, P. and Brantingham, P. (2010). Criminality of place. *European Journal on Criminal Policy and Research*, 3(3):5–26.

Brockmann, D., Hufnagel, L., and Geisel, T. (2006). The scaling laws of human travel. *Nature*, 439(7075):462–465.

Brunton-Smith, I. and Sturgis, P. (2011). Do neighborhoods generate fear of crime? an empirical test using the British Crime Survey. *Criminology*, 49(2):331–369.

Bureau of Justice Statistics (2016). National Crime Victimization Survey (NCVS). `http://bjs.ojp.usdoj.gov/index.cfm?ty=dcdetail&iid=245`. Accessed on May 2016.

Burger, M., Van Oort, F., and Linders, G.-J. (2009). On the specification of the gravity model of trade: zeros, excess zeros and zero-inflated estimation. *Spatial Economic Analysis*, 4(2):167–190.

Bushway, S. D. and Tahamont, S. (2016). Modeling long-term criminal careers what happened to the variability? *Journal of Research in Crime and Delinquency*, 53(3):372–391.

Cantor, D. J. (2014). The new wave: forced displacement caused by organized crime in Central America and Mexico. *Refugee Survey Quarterly*.

Carr, D. B., Olsen, A. R., and White, D. (1992). Hexagon mosaic maps for display of univariate and bivariate geographical data. *Cartography and Geographic Information Systems*, 19(4):228–236.

Carro, D., Valera, S., and Vidal, T. (2010). Perceived insecurity in the public space: personal, social and environmental variables. *Quality and Quantity*, 44(2):303–314.

Castellano, C., Fortunato, S., and Loreto, V. (2009). Statistical physics of social dynamics. *Reviews of Modern Physics*, 81(2):591.

Chadee, D. and Ditton, J. (2005). Fear of crime and the media: assessing the lack of relationship. *Crime, Media, Culture*, 1(3):322–332.

Chen, S. X. and Liu, J. S. (1997). Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Statistica Sinica*, 7(1):875–892.

Chen, X., Cho, Y., and Jang, S. Y. (2015). Crime prediction using Twitter sentiment and weather. In *Systems and Information Engineering Design Symposium (SIEDS), 2015*, pages 63–68. IEEE.

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

Coletto, M., Esuli, A., Lucchese, C., Muntean, C. I., Nardini, F. M., Perego, R., and Renso, C. (2017). Perception of social phenomena through the multidimensional analysis of online social networks. *Online Social Networks and Media*, 1(Supplement C):14 – 32.

Constant, A. F. and Zimmermann, K. F. (2012). The dynamics of repeat migration: A Markov chain analysis. *International Migration Review*, 46(2):362–388.

Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., and Tesconi, M. (2016). DNA-inspired online behavioral modeling and its application to spambot detection. *IEEE Intelligent Systems*, 31(5):58–64.

Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., and Tesconi, M. (2017). Social fingerprinting: detection of spambot groups through DNA-inspired behavioral modeling. *IEEE Transactions on Dependable and Secure Computing*.

Cullen, J. B. and Levitt, S. D. (1999). Crime, urban flight, and the consequences for cities. *Review of Economics and Statistics*, 81(2):159–169.

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

D'Andrea, E., Ducange, P., Lazzerini, B., and Marcelloni, F. (2015). Real-time detection of traffic from Twitter stream analysis. *IEEE Transactions on Intelligent Transportation Systems*, 16(4):2269–2283.

Davies, T. P., Fry, H. M., Wilson, A. G., and Bishop, S. R. (2013). A mathematical model of the London riots and their policing. *Scientific Reports*, 3:1303.

Deffuant, G., Neau, D., Amblard, F., and Weisbuch, G. (2000). Mixing beliefs among interacting agents. *Advances in Complex Systems*, 3(4):87–98.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., and Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559.

Dickinson, P. W. J. (1993). Fear of crime: read all about it? the relationship between newspaper crime reporting and fear of crime. *British Journal of Criminology*, 33(1):33–56.

Diggle, P. (2014). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. CRC Press, Florida, US., third edition edition.

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.

Dorfman, R. (1979). A formula for the Gini coefficient. *The Review of Economics and Statistics*, pages 146–149.

D'Orsogna, M. R. and Perc, M. (2015). Statistical physics of crime: A review. *Physics of Life Reviews*, 12:1–21.

Düring, B., Markowich, P., Pietschmann, J.-F., and Wolfram, M.-T. (2009). Boltzmann and Fokker–Planck equations modelling opinion formation in the presence of strong leaders. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 465(2112):3687–3708.

Düring, B. and Wolfram, M.-T. (2015). Opinion dynamics: inhomogeneous Boltzmann-type equations modelling opinion leadership and political segregation. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 471(2182).

Dustmann, C., Glitz, A., and Frattini, T. (2008). The labour market impact of immigration. *Oxford Review of Economic Policy*, 24(3):477.

Eck, J. E., Lee, Y. J., O, S. H., and Martinez, N. N. (2017). Compared to what? estimating the relative concentration of crime at places using systematic and other reviews. *Crime Science*, 6(1):8.

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.

Farnham, A., Kurz, C., et al. (2017). Early career researchers want open science. *Genome Biology*, 18(1):221.

Farrall, S., Bannister, J., Ditton, J., and Gilchrist, E. (1997). Questioning the measurement of the 'fear of crime': findings from a major methodological study. *British Journal of Criminology*, 37(4):658–679.

Farrall, S., Bannister, J., Ditton, J., and Gilchrist, E. (2000). Social psychology and the fear of crime. *British Journal of Criminology*, 40(3):399–413.

Farrell, G. (2015). Crime concentration theory. *Crime Prevention & Community Safety*, 17(4):233–248.

Farrell, G. and Pease, K. (1993). Once bitten, twice bitten: Repeat victimization and its implications for crime prevention. *Police Research Group, Crime Prevention Unit*, Paper 46(1):1–32.

Farrell, G., Tilley, N., Tseloni, A., and Mailley, J. (2011). The crime drop and the security hypothesis. *Journal of Research in Crime and Delinquency*, page 0022427810391539.

Farrell, G., Tseloni, A., and Pease, K. (2005). Repeat victimization in the ICVS and the NCVS. *Crime Prevention & Community Safety*, 7(3):7–18.

Farrington, D. P., Jolliffe, D., Loeber, R., Stouthamer-Loeber, M., and Kalb, L. M. (2001). The concentration of offenders in families, and family criminality in the prediction of boys' delinquency. *Journal of Adolescence*, 24(5):579–596.

Ferguson, T. (1952). The young delinquent in his social setting. A Glasgow study. *The Young Delinquent in his Social Setting. A Glasgow Study.*

Ferraro, K. F. and Grange, R. L. (1987). The measurement of fear of crime. *Sociological Inquiry*, 57(1):70–97.

Fox, J. A. and Tracy, P. E. (1988). A measure of skewness in offense distributions. *Journal of Quantitative Criminology*, 4(3):259–274.

Freeman, S. (1996). The spatial concentration of crime. *Journal of Urban Economics*, 40(1):216–231.

Frith, M. J., Johnson, S. D., and Fry, H. M. (2017). Role of the street network in burglars' spatial decision-making*. *Criminology*, pages 1–33.

from INEGI, O. W. (2014). Victimisation survey, 2014. `http://www3.inegi.org.mx/sistemas/microdatos/encuestas.aspx?c=34517&s=est`. Accessed on August 2015.

from INEGI, O. W. (2016). Victimisation survey, 2016. `http://www3.inegi.org.mx/sistemas/microdatos/encuestas.aspx?c=34517&s=est`. Accessed on October 2016.

Galam, S., Gefen, Y., and Shapir, Y. (1982). Sociophysics: A new approach of sociological collective behaviour. i. mean-behaviour description of a strike. *Journal of Mathematical Sociology*, 9(1):1–13.

Galor, O. (2007). *Discrete dynamical systems*. Springer Science & Business Media.

Garcia, A. J., Pindolia, D. K., Lopiano, K. K., and Tatem, A. J. (2015). Modeling internal migration flows in Sub-Saharan Africa using census microdata. *Migration Studies*, 3(1):89–110.

Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*, 61:115–125.

Gil de Ziga, H., Jung, N., and Valenzuela, S. (2012). Social media use for news and individuals' social capital, civic engagement and political participation. *Journal of Computer-Mediated Communication*, 17(3):319–336.

Gilchrist, E., Bannister, J., Ditton, J., and Farrall, S. (1998). Women and the 'fear of crime' challenging the accepted stereotype. *British Journal of Criminology*, 38(2):283–298.

Giulianotti, R., Armstrong, G., Hales, G., and Hobbs, D. (2015). Global sport mega-events and the politics of mobility: the case of the London 2012 Olympics. *The British Journal of Sociology*, 66(1):118–140.

Glaeser, E. L. and Sacerdote, B. (1996). Why is there more crime in cities? Technical report, National Bureau of Economic Research.

Glaeser, E. L., Sacerdote, B., and Scheinkman, J. A. (1995). Crime and social interactions. Technical report, National Bureau of Economic Research.

Global Volcanism Program (2013). Volcanoes of the World, v. 4.5.0. Venzke, E (ed.). Smithsonian Institution. Accessed on June 2016.

Glueck, S. and Glueck, E. (1950). Unraveling juvenile delinquency. *Juvenile Court Judges Journal*, 2:32.

Gnisci, D. (2008). West African mobility and migration policies of OECD countries. *Paris, Organisation for Economic Cooperation and Development, Sahel and West Africa Club*.

Gonzalez, M. C., Hidalgo, C. A., and Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196):779–782.

Gordon, M. B. (2010). A random walk in the literature on criminality: A partial and critical view on some statistical analyses and modelling approaches. *European Journal of Applied Mathematics*, 21(4-5):283–306.

Gottfredson, M. R. and Hindelang, M. J. (1981). Sociological aspects of criminal victimization. *Annual Review of Sociology*, pages 107–128.

Greenland, S. (2004). Interval estimation by simulation as an alternative to and extension of confidence intervals. *International Journal of Epidemiology*, 33(6):1389–1397.

Grogger, J. and Weatherford, S. (1995). Crime, policing and the perception of neighborhood safety. *Political Geography*, 14(6/7):521–541.

Grove, L., Farrell, G., Farrington, D., and Johnson, S. (2012). *Preventing Repeat Victimization: A Systematic Review*. The Swedish National Council for Crime Prevention, Stockholm, Sweden, first edition edition.

Guerette, R. T. and Bowers, K. J. (2009). Assessing the extent of crime displacement and diffusion of benefits: A review of situational crime prevention evaluations. *Criminology*, 47(4):1331–1368.

Hale, C. (1996). Fear of crime: a review of the literature. *International Review of Victimology*, 4(2):79–150.

Hale, C., Pack, P., and Salked, J. (1994). The structural determinants of fear of crime: an analysis using census and crime survey data from England and Wales. *International Review of Victimology*, 3(3):211–233.

Harris, J. R. and Todaro, M. P. (1970). Migration, unemployment and development: a two-sector analysis. *The American Economic Review*, 60(1):126–142.

Heath, L. and Gilbert, K. (1996). Mass media and fear of crime. *American Behavioral Scientist*, 39(4):379–386.

Hegselmann, R. and Krause, U. (2002). Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3).

Helbing, D., Brockmann, D., Chadefaux, T., Donnay, K., Blanke, U., Woolley-Meza, O., Moussaid, M., Johansson, A., Krause, J., Schutte, S., and Perc, M. (2015). Saving human lives: What complexity science and information systems can contribute. *Journal of Statistical Physics*, 158(3):735–781.

Henry, N. W., McGinnis, R., and Tegtmeyer, H. W. (1971). A finite model of mobility. *The Journal of Mathematical Sociology*, 1(1):107–118.

Henry, S., Boyle, P., and Lambin, E. F. (2003). Modelling inter-provincial migration in Burkina Faso, West Africa: the role of socio-demographic and environmental factors. *Applied Geography*, 23(2):115–136.

Hijmans, R. J. (2016). *raster: Geographic Data Analysis and Modeling*. R package version 2.5-8.

Himelboim, I., Hansen, D., and Bowser, A. (2013a). Playing in the same Twitter network: political information seeking in the 2010 US gubernatorial elections. *Information, Communication & Society*, 16(9):1373–1396.

Himelboim, I., McCreery, S., and Smith, M. (2013b). 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.

Hindelang, M. J., Gottfredson, M. R., and Garofalo, J. (1978). *Victims of personal crime: An empirical foundation for a theory of personal victimization*. Ballinger Cambridge, MA, Massachusetts, US.

Hipp, J. R. and Kim, Y.-A. (2016). Measuring crime concentration across cities of varying sizes: Complications based on the spatial and temporal scale employed. *Journal of Quantitative Criminology*, pages 1–38.

Hollis, M. E., Downey, S., del Carmen, A., and Dobbs, R. R. (2017). The relationship between media portrayals and crime: perceptions of fear of crime among citizens. *Crime Prevention and Community Safety*, 19(1):46–60.

Hope, T. and Norris, P. A. (2013). Heterogeneity in the frequency distribution of crime victimization. *Journal of Quantitative Criminology*, 29(4):543–578.

Hope, T. and Trickett, A. (2008). The distribution of crime victimisation in the population. *International Review of Victimology*, 15(1):37–58.

Horwood, L. J. and Fergusson, D. M. (2000). Drink driving and traffic accidents in young people. *Accident Analysis & Prevention*, 32(6):805–814.

Hunt, J. C., Timoshkina, Y., Baudains, P. J., and Bishop, S. R. (2012). System dynamics applied to operations and policy decisions. *European Review*, 20(3):324–342.

Ibáñez, A. M. and Vélez, C. E. (2008). Civil conflict and forced migration: The micro determinants and welfare losses of displacement in Colombia. *World Development*, 36(4):659–676.

Jackson, J. and Gray, E. (2010). Functional fear and public insecurities about crime. *British Journal of Criminology*, 50(1):1–22.

Johnson, J. H. (2000). The can you trust it? problem of simulation science in the design of socio-technical systems. *Complexity*, 6(2):34–40.

Johnson, J. H. (2010a). The future of the social sciences and humanities in the science of complex systems. *Innovation–The European Journal of Social Science Research*, 23(2):115–134.

Johnson, S. (2010b). A brief history of the analysis of crime concentration. *European Journal of Applied Mathematics*, 21(1):349–370.

Johnson, S., Guerette, R., and Bowers, K. (2014). Crime displacement: what we know, what we dont know, and what it means for crime reduction. *Journal of Experimental Criminology*, 10(1):549–571.

Johnson, S., Summers, L., and Pease, K. (2009). Offender as forager? a direct test of the boost account of victimization. *Journal of Quantitative Criminology*, 25(2):181–200.

Johnson, S. D. and Groff, E. R. (2014). Strengthening theoretical testing in criminology using agent-based modeling. *Journal of Research in Crime and Delinquency*, 51(4):509–525.

Jovanis, P. P. and Chang, H.-L. (1986). Modeling the relationship of accidents to miles traveled. *Transportation Research Record*, 1068:42–51.

Jovanis, P. P. and Chang, H.-L. (1989). Disaggregate model of highway accident occurrence using survival theory. *Accident Analysis & Prevention*, 21(5):445–458.

Kacperski, K. and Holyst, J. (1999). Opinion formation model with strong leader and external impact: a mean field approach. *Physica A: Statistical Mechanics and its Applications*, 269(2):511–526.

Kadar, C. and Pletikosa, I. (2018). Mining large-scale human mobility data for long-term crime prediction. *EPJ Data Science*, 7(1):26.

Karemera, D., Oguledo, V. I., and Davis, B. (2000). A gravity model analysis of international migration to North America. *Applied Economics*, 32(13):1745–1755.

Kelley, A. C. and Weiss, L. W. (1969). Markov processes and economic analysis: the case of migration. *Econometrica: Journal of Econometric Society*, pages 280–297.

Kendall, M. G. (1948). *Rank correlation methods*. C. Griffin, London, UK, first edition edition.

Kershaw, C. and Tseloni, A. (2005). Predicting crime rates, fear and disorder based on area information: evidence from the 2000 British Crime Survey. *International Review of Victimology*, 12(3):293–311.

Kleck, G. and Barnes, J. (2014). Do more police lead to more crime deterrence? *Crime & Delinquency*, 60(5):716–738.

Konseiga, A. (2006). Household migration decisions as survival strategy: The case of Burkina Faso. *Journal of African Economies*, 16(2):198–233.

Kounadi, O., Lampoltshammer, T. J., Groff, E., Sitko, I., and Leitner, M. (2015). Exploring Twitter to analyze the publics reaction patterns to recently reported homicides in London. *PloS One*, 10(3):e0121848.

Kumar, R. and Vassilvitskii, S. (2010). Generalized distances between rankings. *International World Wide Web Conference Committee (IW3C2)*, 39(1):17–24.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, pages 591–600. ACM.

Laczko, F., Aghazarm, C., et al. (2009). *Migration, environment and climate change: Assessing the evidence*. International Organization for Migration Geneva, Geneva, Switzerland.

Lambiotte, R., Blondel, V. D., De Kerchove, C., Huens, E., Prieur, C., Smoreda, Z., and Van Dooren, P. (2008). Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387(21):5317–5325.

Lampos, V. and Cristianini, N. (2012). Nowcasting events from the social web with statistical learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4):72.

Lane, J. and Meeker, J. W. (2003). Ethnicity, information sources, and fear of crime. *Deviant Behavior*, 24(1):1–26.

Latané, B. (1981). The psychology of social impact. *American Psychologist*, 36(4):343.

Latin American Public Opinion Project (LAPOP), T. (2017). The Americas Barometer. data from `www.LapopSurveys.org`. Accessed on October 2017.

Laycock, G. and Farrell, G. (2003). Repeat victimization: Lessons for implementing problem-oriented policing. *Crime Prevention Studies*, 15(1):213–237.

Lee, J. J., Helke, J., and Laczko, F. (2015). *World Migration Report 2015*. International Organization for Migration, Geneva, Switzerland.

Lee, Y. J., Eck, J. E., O, S., and Martinez, N. N. (2017). How concentrated is crime at places? a systematic review from 1970 to 2015. *Crime Science*, 6(1):6.

Levin, A., Rosenfeld, R., and Deckard, M. (2016). The law of crime concentration: An application and recommendations for future research. *Journal of Quantitative Criminology*, pages 1–13.

Levy, M. (2010). Scale-free human migration and the geography of social networks. *Physica A: Statistical Mechanics and its Applications*, 389(21):4913–4917.

Levy, M. and Goldenberg, J. (2014). The gravitational law of social interaction. *Physica A: Statistical Mechanics and its Applications*, 393:418–426.

Lewer, J. J. and Van den Berg, H. (2008). A gravity model of immigration. *Economics Letters*, 99(1):164–167.

Lewis, D. A. and Maxfield, M. G. (1980). Fear in the neighborhoods: an investigation of the impact of crime. *Journal of Research in Crime and Delinquency*, 17(2):160–189.

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

Loecher, M. and Ropkins, K. (2015). RgoogleMaps and loa: unleashing R graphics power on map tiles. *Journal of Statistical Software*, 63(4):1–18.

Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, 9(70):209–219.

Maher, M. (1990). A bivariate negative binomial model to explain traffic accident migration. *Accident Analysis & Prevention*, 22(5):487–498.

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.

Maltz, M. D. (1996). From Poisson to the present: Applying operations research to problems of crime and justice. *Journal of Quantitative Criminology*, 12(1):3–61.

Mansury, Y. and Shin, J. (2015). Size, connectivity, and tipping in spatial networks: Theory and empirics. *Computers, Environment and Urban Systems*, 54:428–437.

Marsh, C. and Elliott, J. (2008). *Exploring data: an introduction to data analysis for social scientists*. Polity.

Martinez, N. N., Lee, Y., Eck, J. E., and O, S. (2017). Ravenous wolves revisited: a systematic review of offending concentration. *Crime Science*, 6(1):10.

Martínez Teutle, A. R. (2010). Twitter: network properties analysis. *Electronics, Communications and Computer (CONIELECOMP), 2010 20th International Conference on*, pages 180–186.

McGinnis, R. (1968). A stochastic model of social mobility. *American Sociological Review*, pages 712–722.

McLachlan, G. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

Mendoza, M., Poblete, B., and Castillo, C. (2010). Twitter under crisis: can we trust what we RT? In *Proceedings of the first workshop on social media analytics*, pages 71–79. ACM.

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., and Tita, G. E. (2012). Self-exciting point process modeling of crime. *Journal of the American Statistical Association*.

Mooney, C. Z. (1997). *Monte Carlo Simulation*. Sage Publication, Inc., California, USA, first edition edition.

Morgan, P. (1978). *Delinquent fantasies*. Temple Smith London.

Muntean, C. I., Nardini, F. M., Silvestri, F., and Baraglia, R. (2015). On learning prediction models for tourists paths. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(1):8.

Myers, C. A., Slack, T., and Singelmann, J. (2008). Social vulnerability and migration in the wake of disaster: the case of hurricanes Katrina and Rita. *Population and Environment*, 29(6):271–291.

Myers, G. C., McGinnis, R., and Masnick, G. (1967). The duration of residence approach to a dynamic stochastic model of internal migration: a test of the axiom of cumulative inertia. *Eugenics Quarterly*, 14(2):121–126.

Nagin, D. S. and Land, K. C. (1993). Age, criminal careers, and population heterogeneity: Specification and estimation of a nonparametric, mixed Poisson model. *Criminology*, 31(3):327–362.

Nathan, M. (2014). The wider economic impacts of high-skilled migrants: a survey of the literature for receiving countries. *IZA Journal of Migration*, 3(1):4.

Naude, W. (2008). *Conflict, disasters and no jobs: Reasons for international migration from Sub-Saharan Africa*. Number 85 in 1. UNU-WIDER.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167–256.

Newman, M. E. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5):323–351.

Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., and Mascolo, C. (2012). A tale of many cities: universal patterns in human urban mobility. *PloS One*, 7(5):e37027.

O, S. H., Martinez, N. N., Lee, Y. J., and Eck, J. E. (2017). How concentrated is crime among victims? a systematic review from 1977 to 2014. *Crime Science*, 6(1):9.

of Commerce, U. C. B. D. (2015). Metro Area-to-Metro Area Migration Flows: 2010-2014; American Community Survey (ACS). `https://www.census.gov/data/tables/2014/demo/geographic-mobility/metro-to-metro-migration.html`. Accessed April 2017.

Office for National Statistics (2016). Crime Survey for England and Wales (CSEW). `http://dx.doi.org/10.5255/UKDA-SN-7280-4`. Accessed on May 2016.

Okabe, A., Boots, B., Sugihara, K., and Chiu, S. N. (2009). *Spatial tessellations: concepts and applications of Voronoi diagrams*, volume 501. John Wiley & Sons, Sussex, UK.

Oliveira, M., Bastos-Filho, C., and Menezes, R. (2017). The scaling of crime concentration in cities. *PloS one*, 12(8):e0183110.

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

Osborn, D. R. and Tseloni, A. (1998). The distribution of household property crimes. *Journal of Quantitative Criminology*, 14(3):307–330.

Osgood, D. W. (2000). Poisson-based regression analysis of aggregate crime rates. *Journal of Quantitative Criminology*, 16(1):21–43.

Page, S. (2010). *Diversity and complexity*. Princeton University Press, Princeton, New Jersey.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. *LREc*, 10(2010).

Pan, X., Han, C. S., Dauber, K., and Law, K. H. (2007). A multi-agent based framework for the simulation of human and social behaviors during emergency evacuations. *Ai & Society*, 22(2):113–132.

Pantazis, C. (2000). 'Fear of crime', vulnerability and poverty. Evidence from the British Crime Survey. *British Journal of Criminology*, 40(1):414–436.

Pappalardo, L., Pedreschi, D., Smoreda, Z., and Giannotti, F. (2015a). Using big data to study the link between human mobility and socio-economic development. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 871–878.

Pappalardo, L., Simini, F., Rinzivillo, S., Pedreschi, D., Giannotti, F., and Barabási, A.-L. (2015b). Returners and explorers dichotomy in human mobility. *Nature Communications*, 6:8166 EP –.

Pappalardo, L., Vanhoof, M., Gabrielli, L., Smoreda, Z., Pedreschi, D., and Giannotti, F. (2016). An analytical framework to nowcast well-being using mobile phone data. *International Journal of Data Science and Analytics*, 2(1):75–92.

Park, S. M. and Eck, J. E. (2013). Understanding the random effect on victimization distributions: A statistical analysis of random repeat victimizations. *Victims & Offenders*, 8(4):399–415.

Paul, B. K. (2005). Evidence against disaster-induced migration: the 2004 tornado in north-central Bangladesh. *Disasters*, 29(4):370–385.

Pease, K. (1998). *Repeat victimisation: Taking stock*. Home Office Police Research Group London, United Kingdom.

Pease, K. and Ignatans, D. (2016). The global crime drop and changes in the distribution of victimisation. *Crime Science*, 5(1):1–6.

Pepys, R. C. (2016). *Developing mathematical models of complex social processes: radicalisation and criminality development*. PhD thesis, UCL (University College London).

Peterson, D. W. (1986). Volcanoes: tectonic setting and impact on society. *Active Tectonics, Geophysics Study Committee, National Research Council. National Academy Press, Washington DC*, pages 231–246.

Prasannakumar, V., Vijith, H., Charutha, R., and Geetha, N. (2011). Spatio-temporal clustering of road accidents: Gis based analysis and assessment. *Procedia-Social and Behavioral Sciences*, 21:317–325.

Prieto Curiel, R. and Bishop, S. R. (2016a). A measure of the concentration of rare events. *Scientific Reports*, 6.

Prieto Curiel, R. and Bishop, S. R. (2016b). A metric of the difference between perception of security and victimisation rates. *Crime Science*, 5(1).

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

Prieto Curiel, R. and Bishop, S. R. (2018). Fear of crime: the impact of different distributions of victimisation. *Palgrave Communications*, 4(1):46.

Prieto Curiel, R., Collignon Delmar, S., and Bishop, S. R. (2017a). Measuring the distribution of crime and its concentration. *Journal of Quantitative Criminology*, pages 1–29.

Prieto Curiel, R., González Ramírez, H., and Bishop, S. R. (2018a). A novel rare event approach to measure the randomness and concentration of road accidents. *PloS one*, 13(8):e0201890.

Prieto Curiel, R., Heinrigs, P., and Heo, I. (2017b). Cities and spatial interactions in West Africa. Technical report, OECD Publishing.

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

Pumain, D. and Guerois, M. (2004). Scaling laws in urban systems. In *Santa Fe Institute, Working Papers*, pages 1–26.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

Reuveny, R. (2007). Climate change-induced migration and violent conflict. *Political Geography*, 26(6):656–673.

Rhee, I., Shin, M., Hong, S., Lee, K., Kim, S. J., and Chong, S. (2011). On the Levy-walk nature of human mobility. *IEEE/ACM transactions on networking (TON)*, 19(3):630–643.

Rinzivillo, S., Gabrielli, L., Nanni, M., Pappalardo, L., Pedreschi, D., and Giannotti, F. (2014). The purpose of motion: Learning activities from individual mobility networks. In *2014 International Conference on Data Science and Advanced Analytics (DSAA)*, pages 312–318.

Rosser, G., Davies, T., Bowers, K. J., Johnson, S. D., and Cheng, T. (2016). Predictive crime mapping: Arbitrary grids or street networks? *Journal of Quantitative Criminology*, pages 1–26.

Ruijsbroek, A., Droomers, M., Groenewegen, P. P., Hardyns, W., and Stronks, K. (2015). Social safety, self-rated general health and physical activity: changes in area crime, area safety feelings and the role of social cohesion. *Health & Place*, 31:39–45.

Sacco, V. F. (1993). Social support and the fear of crime. *Canadian Journal of Criminology*, 35:187.

Sagberg, F. (1999). Road accidents caused by drivers falling asleep. *Accident Analysis & Prevention*, 31(6):639–649.

San Miguel, M., Johnson, J. H., Kertesz, J., Kaski, K., Díaz-Guilera, A., MacKay, R. S., Loreto, V., Érdi, P., and Helbing, D. (2012). Challenges in complex systems science. *The European Physical Journal Special Topics*, 214(1):245–271.

Sandefur, J. T. (1993). *Discrete dynamical modeling*. Oxford University Press on Demand.

Savolainen, P. T., Mannering, F. L., Lord, D., and Quddus, M. A. (2011). The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. *Accident Analysis & Prevention*, 43(5):1666–1676.

Schelling, T. C. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology*, 1(2):143–186.

Schläpfer, M., Bettencourt, L. M., 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.

Schlattmann, P. (2005). On bootstrapping the number of components in finite mixtures of Poisson distributions. *Statistics and Computing*, 15(3):179–188.

Schlattmann, P. and Böhning, D. (1993). Mixture models and disease mapping. *Statistics in Medicine*, 12(19-20):1943–1950.

Schlattmann, P., Hoehne, J., and Verba, M. (2015). *CAMAN: finite Mixture Models and Meta-Analysis Tools - Based on C.A.MAN*. R package version 0.73.

Schuurman, N., Cinnamon, J., Crooks, V. A., and Hameed, S. M. (2009). Pedestrian injury and the built environment: an environmental scan of hotspots. *Bio Med Central Public Health*, 9(1):233.

Schwartz, A. (1973). Interpreting the effect of distance on migration. *Journal of Political Economy*, 81(5):1153–1169.

Shieh, G. S. (1998). A weighted Kendall's tau statistic. *Statistics and Probability Letters*, 39(1):17–24.

Short, M., Brantingham, J., Bertozzi, A., and Tita, G. (2010). Dissipation and displacement of hotspots in reaction-diffusion models of crime. *Proceedings of the National Academy of Sciences*, 107(1):3961–3965.

Short, M. B., D'Orsogna, M. R., Brantingham, P. J., and Tita, G. E. (2009). Measuring and modeling repeat and near-repeat burglary effects. *Journal of Quantitative Criminology*, 25(3):325–339.

Short, M. B., D'Orsogna, M. R., Pasour, V. B., Tita, G. E., Brantingham, P. J., Bertozzi, A., and Chayes, L. B. (2008). A statistical model of criminal behavior. *Mathematical Models and Methods in Applied Sciences*, 18(supp01):1249–1267.

Simini, F., González, M. C., Maritan, A., and Barabási, A.-L. (2012). A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100.

Skogan, W. G. (1987). The impact of victimization on fear. *Crime & Delinquency*, 33(1):135–154.

Skogan, W. G. and Maxfield, M. G. (1981). *Coping with crime: individual and neighborhood reactions*. Sage Publications Beverly Hills, CA.

Song, C., Qu, Z., Blumm, N., and Barabási, A.-L. (2010). Limits of predictability in human mobility. *Science*, 327(5968):1018–1021.

Sparks, R. F. (1981). Multiple victimization: Evidence, theory, and future research. *Journal of Criminal Law and Criminology*, 72(2):762–778.

Stark, R. (1987). Deviant places: A theory of the ecology of crime. *Criminology*, 25(4):893–910.

Steenberghen, T., Aerts, K., and Thomas, I. (2010). Spatial clustering of events on a network. *Journal of Transport Geography*, 18(3):411 – 418. Tourism and climate change.

Taylor, J. E., Filipski, M. J., Alloush, M., Gupta, A., Rojas Valdes, R. I., and Gonzalez-Estrada, E. (2016). Economic impact of refugees. *Proceedings of the National Academy of Sciences*, 113(27):7449–7453.

Temnikova, I., Vieweg, S., and Castillo, C. (2015). The case for readability of crisis communications in social media. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1245–1250. ACM.

Thomas, I. (1996). Spatial data aggregation: exploratory analysis of road accidents. *Accident Analysis & Prevention*, 28(2):251–264.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(sup1):234–240.

Todaro, M. P. (1969). A model of labor migration and urban unemployment in less developed countries. *The American Economic Review*, 59(1):138–148.

Toscani, G. (2006). Kinetic models of opinion formation. *Communications in Mathematical Sciences*, 4(3):481–496.

Tseloni, A. (2000). Personal criminal victimization in the United States: fixed and random effects of individual and household characteristics. *Journal of Quantitative Criminology*, 16(4):415–442.

Tseloni, A. (2007). Fear of crime, perceived disorders and property crime: a multivariate analysis at the area level. *Crime Prevention Studies*, 21:163–185.

Tseloni, A., Mailley, J., Farrell, G., and Tilley, N. (2010). Exploring the international decline in crime rates. *European Journal of Criminology*, 7(5):375–394.

Tseloni, A. and Pease, K. (2003). Repeat personal victimization. Boosts or Flags? *British Journal of Criminology*, 43(1):196–212.

Tseloni, A. and Pease, K. (2004). Repeat personal victimization random effects, event dependence and unexplained heterogeneity. *British Journal of Criminology*, 44(6):931–945.

Tseloni, A. and Pease, K. (2005). Population inequality: the case of repeat crime victimization. *International Review of Victimology*, 12(1):75–90.

Tseloni, A., Wittebrood, K., Farrell, G., and Pease, K. (2004). Burglary victimization in England and Wales, the United States and the Netherlands: A cross-national comparative test of routine activities and lifestyle theories. *British Journal of Criminology*, 44(1):66–91.

Wang, X.-W., Han, X.-P., and Wang, B.-H. (2014). Correlations and scaling laws in human mobility. *PloS One*, 9(1):e84954.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of small-worldnetworks. *Nature*, 393(6684):440–442.

Weisburd, D. (2015). The law of crime concentration and the criminology of place. *Criminology*, 53(2):133–157.

Wesolowski, A., Buckee, C. O., Pindolia, D. K., Eagle, N., Smith, D. L., Garcia, A. J., and Tatem, A. J. (2013). The use of census migration data to approximate human movement patterns across temporal scales. *PloS one*, 8(1):e52971.

Westerlund, J. and Wilhelmsson, F. (2011). Estimating the gravity model without gravity using panel data. *Applied Economics*, 43(6):641–649.

WFP (2017). At the root of exodus: Food security, conflict and international migration. World Food Programme.

White, S., Yehle, T., Serrano, H., Oliveira, M., and Menezes, R. (2014). The spatial structure of crime in urban environments. In *International Conference on Social Informatics*, pages 102–111. Springer.

Wilcox Rountree, P. and Land, K. (1996). Perceived risk versus fear of crime: Empirical evidence of conceptually distinct reactions in survey data. *Social Forces*, 74(4):1353–1376.

Wolfgang, M. E. (1983). Delinquency in two birth cohorts. In *Prospective studies of crime and delinquency*, pages 7–16. Springer.

Wolfgang, M. E., Figlio, R. M., and Sellin, T. (1987). *Delinquency in a birth cohort*. University of Chicago Press, Illinois, US.

World Bank; International Monetary Fund (2015). Global monitoring report 2015/2016: Development goals in an era of demographic change. *Global Monitoring Report*, 2016.

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.

Yntema, D. B. (1933). Measures of the inequality in the personal distribution of wealth or income. *Journal of the American Statistical Association*, 28(184):423–433.

Zeileis, A. (2014). *Ineq: measuring inequality, concentration, and poverty*. R package version 0.2-13.