VNIVERSITAT DE VALÈNCIA

FACULTAT DE CIÈNCIES MATEMÀTIQUES

DEPARTAMENT D'ESTADÍSTICA I INVESTIGACIÓ OPERATIVA

PROGRAMA DE DOCTORAT EN ESTADÍSTICA I OPTIMITZACIÓ

TESI DOCTORAL

Spatio-temporal methods for the analysis of crime and traffic safety data

by

Álvaro Briz Redón

Supervised by

Francisco Martínez Ruiz

Francisco Montes Suay

April, 2020

Francisco Martínez Ruiz, Estadístic de l'Ajuntament de València, i Francisco Montes Suay, Professor Emèrit del Departament d'Estadística i Investigació Operativa de la Universitat de València,

CERTIFIQUEN que la present memòria d'investigació, titulada:

"Spatio-temporal methods for the analysis of crime and traffic safety data"

ha estat realitzada sota la seua direcció per Álvaro Briz Redón i constitueix la seua tesi per optar al grau de Doctor per la Universitat de València Estudi General.

I perquè així conste, en compliment amb la normativa vigent, n'autoritzen la presentació davant la Facultat de Ciències Matemàtiques de la Universitat de València perquè en puga ser tramitada la lectura i defensa pública.

València, 1 d'abril de 2020.

Francisco Martínez Ruiz

Francisco Montes Suay

Agradecimientos

En primer lugar, he de dar las gracias, fundamentalmente, a mis dos directores de tesis, Paco Montes y Paco Martínez, por su dedicación y confianza durante estos años. Sin ellos, este proyecto no habría podido salir adelante.

A la Policía Nacional y a la Policía Local de València, gracias por proporcionar los datos que han sido analizados para el desarrollo de esta tesis. En particular, a la Unidad de Atestados y al Centro Integral de Seguridad y Emergencias Sala 092 de València, y a los jefes de la Policía Local de València durante este periodo, José Serrano y José Vicente Herrera.

Gracias a la Oficina de Estadística del Ayuntamiento de València, a su personal actual y a las personas que han formado parte de ella durante los últimos años. Sin su labor, buena parte de los estudios incluidos en esta tesis no habrían podido llevarse a cabo.

En general, gracias a todas las personas que se han preocupado por el desarrollo de esta tesis. En especial, a Daymé y Ekaterina por su ánimo constante.

A Carolina, gracias por todo lo que me aporta desde hace unos meses.

Por último, debo dar las gracias a mis padres y abuelos, por su apoyo incondicional en todo momento.

Contents

List of Figures v

List of Tables xi

1 Introduction 1
- 1.1 Linear networks 3
- 1.2 Applications 3
  - 1.2.1 Traffic safety analysis 3
  - 1.2.2 Criminology 6

2 Modelling traffic accident counts at the road segment level: Accounting for road intersections 8
- 2.1 Introduction 8
  - 2.1.1 Review of models and methods 9
  - 2.1.2 Traffic safety at road intersections 10
- 2.2 Data 12
  - 2.2.1 Accident information 12
  - 2.2.2 Network structure 12
  - 2.2.3 Network-related covariates 13
- 2.3 Methodology 14
  - 2.3.1 Software 14
  - 2.3.2 Definition of intersection zones 14
  - 2.3.3 Concept of neighbourhoods 16
  - 2.3.4 Road segment neighbourhood geometry 17
  - 2.3.5 Accident count modelling 18
  - 2.3.6 Model checks 20
  - 2.3.7 Kernel density estimation 22
  - 2.3.8 Coldspot/hotspot detection 23
- 2.4 Results and discussion 23
- 2.5 Conclusions 27

3 Identification of differential risk hotspots along a linear network 31
- 3.1 Introduction 31
- 3.2 Data 32
  - 3.2.1 Network structure 32
  - 3.2.2 Accident dataset 33
- 3.3 Methodology 35
  - 3.3.1 Software 35

3.3.2 Estimating a relative probability for each type of accident . . 35
3.3.3 Detecting differential risk hotspots 37
3.3.4 Assessing hotspot significance 39
3.3.5 Ranking hotspots according to further criteria 40
3.4 Results and discussion 41
3.4.1 Example of application: motorcycle accidents 41
3.4.2 General application of the methodology 42
3.5 Conclusions 46

4 Adjusting the Knox test for the analysis of the near-repeat phenomenon 50
4.1 Introduction 50
4.1.1 Factors influencing crime risk: the case of burglaries 52
4.1.2 The statistical assessment of near-repeats 54
4.2 Data 56
4.2.1 Burglaries dataset 56
4.2.2 Road network structure 56
4.3 Methods 57
4.3.1 Software 57
4.3.2 Time information uncertainty 57
4.3.3 The Knox test 58
4.3.4 Adjusting the Knox test 60
4.3.5 Metropolis-Hastings algorithm 62
4.3.6 Burglary risk estimation 64
4.3.7 Covariates selection and construction 65
4.4 Results and discussion 67
4.5 Conclusions 70

5 The modifiable areal unit problem: Accounting for scale and zoning 72
5.1 Introduction 72
5.1.1 TAZ delineation and MAUP effects in traffic safety analysis . 73
5.2 Data 76
5.2.1 Crash dataset and road structure characteristics 76
5.2.2 Covariate definition 76
5.2.3 BSU definitions 77
5.3 Methodology 79
5.3.1 Software 79
5.3.2 Regionalization algorithm 79
5.3.3 Crash counts modelling 80
5.3.4 Assessment of model performance 83
5.3.5 Statistical tools for covariate exploration 83
5.4 Results and discussion 84
5.4.1 Effects of MAUP on input data 85
5.4.2 Parameter variations across aggregation levels 86
5.4.3 Parameter variations across BSU types 89
5.4.4 Model performance comparisons 90
5.4.5 Parameter interpretations 91

5.5 Conclusions 92

6 Modelling risk in relation to multiple sources: The effect of school locations on traffic accidents 100

6.1 Introduction 100
6.2 Data 102
6.2.1 Accident dataset 102
6.2.2 School dataset 102
6.2.3 Covariate definition 102
6.3 Methodology 104
6.3.1 Software 104
6.3.2 Time window of analysis 104
6.3.3 Observed vs. expected ratios 106
6.3.4 Macroscopic modelling 106
6.3.5 Risk modelling in relation to several point sources 107
6.3.6 Case-control study 108
6.4 Results and discussion 108
6.4.1 Risk ratios by school type 109
6.4.2 Spatial count models 109
6.4.3 Multiple source regression 111
6.4.4 Multivariate logistic regression 112
6.5 Conclusions 114

7 Geocoding quality and missing data: Reestimating a minimum acceptable hit rate 118

7.1 Introduction 118
7.2 Data 120
7.2.1 Crime events from Valencia 120
7.2.2 Spatial representations of Valencia 121
7.3 Methodology 121
7.3.1 Software 121
7.3.2 Estimation of a minimum acceptable hit rate 121
7.3.3 Beyond the Mann-Whitney test 123
7.3.4 Introduction of non-uniform geocoding error rates in Ratcliffe's procedure 124
7.3.5 Remarks on point process theory 125
7.4 Results 126
7.4.1 Crime datasets from Valencia 126
7.4.2 Simulation studies 130
7.5 Conclusions 135

8 Software development 138

8.1 SpNetPrep: Overview of the package 138
8.1.1 Technical issues 141
8.1.2 Network edition 143
8.1.3 Network direction 147
8.1.4 Point Pattern Revision 149

8.1.5 Summary 150
8.2 DRHotNet: Overview of the package 151
8.2.1 Using DRHotNet 152
8.2.2 Summary 165

9 Conclusions and future work 169

Resumen 172

References 179

List of Figures

2.1 Road structure of study displayed over a map of the city of Valencia (a) and its representation as a linear network made of links and vertices, with arrows indicating traffic flow directionality (b) 13

2.2 Graphical description of the split network showing the locations of IAZs and MAZs 15

2.3 Graphical description at the road segment level of the following network-related variables: (a) main road indicator, (b) parking slot presence, (c) traffic light presence, (d) bus stop presence, (e) bus lane presence, (f) number of lanes, (g) number of in-neighbours (h) number of out-neighbours and (i) AADT 16

2.4 Examples of types of neighbourhood in a directed linear network. The six road segments that are contiguous to road segment i allow the construction of the neighbourhoods $N(i) = \{a, b, c, d, h, j\}$, $N_{dir}(i) = \{b, d, h, j\}$, $N_{dir}^{in}(i) = \{b, h\}$ and $N_{dir}^{out}(i) = \{d, j\}$ 17

2.5 (a) Clustering of the road segments of the spatial network according to their neighbourhood geometry. (b) Detailed neighbourhood of a road segment, i, of the directed network. Six road segments share a vertex with i, but only four of these allow traffic to flow from i or to i. The values of the geometric variables for road segment i are: Straight$_{in}$ = 1, Right$_{in}$ = 0, Sharp$_{in}$ = 1, Straight$_{out}$ = 1, Right$_{out}$ = 0, Sharp$_{out}$ = 1, Short$_{in}$ = 1, Medium$_{in}$ = 1, Long$_{in}$ = 0, Short$_{out}$ = 1, Medium$_{out}$ = 0, Long$_{out}$ = 1, Length = 41.7, $|N_{dir}^{in}(i)| = 2$, $|N_{dir}^{out}(i)|$ = 2 18

2.6 Quintile distribution of the KDE values for $\sigma = 100$ m (a) and coldspots (green) and hotspots (red) detected after the computation of LISA statistics from these KDE values for the split network made of 20 m IAZ. In (a), each road segment of the split linear network is coloured according to the KDE value at its middle point 28

3.1 Boroughs of the area of the city of Valencia that was studied (a) and complete network structure used for the analysis, with arrows indicating the traffic flow (b). In (a), the nomenclature of each borough is overlayed on the map, with the first number indicating the district the borough belongs to, from 1 to 6: Ciutat Vella, l'Eixample, Extramurs, Campanar, la Saïdia and el Pla del Real 33

3.2 Graphical description of the procedure implemented in order to detect differential risk hostpots. Starting from a road network of interest (a), this is segmented into shorter elements called lixels (b). The traffic accidents available are then projected into the lixellized network in (b), some of which can be of one specific type (orange points), whereas the rest are of any other type (blue points), as shown in (c). Next, KDE is employed to produce an estimation of the probability of risk for the accident type being studied (symbolized by the orange points) at each of the middle points of the lixellized network (d). As an illustration, in (d) the lixels are coloured according to a sensible guess of these probabilities in view of the point pattern in (c), ranging from average or below average (gray) to very high (purple). The lixels satisfying the conditions imposed by the k and n parameters are selected (in red) and grouped, becoming differential risk hotspots (e). Finally, a Monte Carlo technique is applied to the hostspots in (e), yielding a value of significance for each of them that can occasionally lead to their rejection (f) 40

3.3 Estimated probabilities of observing an accident that involves a motorcycle for each lixel of the linear network (a) and differential risk hotspots detected after the application of the methodology with $k = 1$ and (b) $n = 20$, (c) $n = 30$ and (d) $n = 40$ 43

3.4 PAI$_{\text{motorcycle}}$ results (a) and proportion of traffic accidents involving a motorcycle (b) within the set of hotspots obtained with different values of k and n and a bandwidth value of $\sigma = 100$ m 45

3.5 Differential risk hotspots that are statistically significant (at the 0.05 level) for each collision type after the application of the detection procedure with $k = 1$, $n = 40$ and $\sigma = 100$ 46

3.6 Differential risk hotspots that are statistically significant (at the 0.05 level) for each vehicle type after the application of the detection procedure with $k = 1$, $n = 40$ and $\sigma = 100$ 47

3.7 Percentage of coincident road length presented between the set of differential risk hotspots obtained for the collision and vehicle types considered after the application of the procedure with $k = 1$, $n = 40$ and $\sigma = 100$. This percentage is computed with respect to the collision or vehicle type whose set of hotspots spans over a larger number of meters (through the quotient between coincident road length and total length of the longer set of hotspots) 49

4.1 Graphical description of burglaries in the city of Valencia through network-constrained KDE values 57

4.2 Time series representing the incidence of burglaries (grouped by time intervals of 10 days) during the period of two years considered. The series accounting for events whose date of occurrence was known is compared to the ones obtained by using the midpoint and aoristic strategies 59

4.3 Point pattern over the road network of Valencia representing the location of houses and buildings in the city. The colour and size of each point indicates the number of dwellings (exposure unit) the house or building comprise 61

4.4 Burglary counts at the hexagonal grid defined over the area of study (a) and example of construction of a hexagon around a location where a burglary took place (P) for the estimation of a burglary risk around it (b) 64

5.1 Points representing the locations of traffic crashes that occurred in Valencia during the years 2014 and 2015 (a) and time series (displayed by hour and weekday) of traffic crashes observed in Valencia during the same period (b) 76

5.2 Crash counts (CRASH) in Valencia considering a BSU configuration composed of CTs (a), TMs (b), TIs (c) and HEXAs (d). Districts of Valencia are overlayed (thicker lines, in black) for better readability and comparison 79

5.3 Moran's I values for the response (CRASH) and the covariates for each BSU type and aggregation level tested 86

5.4 Assessment of multicollinearity among the covariates through variance inflation factor (VIF) for each BSU type and aggregation level tested 87

5.5 Summary of the results obtained for the CAR models, considering the four types of BSUs and the levels of aggregation that were applied 93

5.6 Combined graph showing the distributions of local parameter estimates, for the covariates used in the GWR models (in rows) and each level of spatial aggregation (in columns) tested for the CTs 94

5.7 Combined graph showing the distributions of local parameter estimates, for the covariates used in the GWR models (in rows) and each level of spatial aggregation (in columns) tested for the TMs 95

5.8 Combined graph showing the distributions of local parameter estimates, for the covariates used in the GWR models (in rows) and each level of spatial aggregation (in columns) tested for the TIs 96

5.9 Combined graph showing the distributions of local parameter estimates, for the covariates used in the GWR models (in rows) and each level of spatial aggregation (in columns) tested for the HEXAs 96

5.10 Summary of the results obtained for the GWR models, considering the four types of BSUs and the levels of aggregation that were applied 97

5.11 Estimated local parameters for the GWR model considering CTs, TMs, TIs and HEXAs for EDU (a-d), BAR (e-h) and BUS (i-l) at AG300. Districts of Valencia are overlayed (thicker lines, in black) for better readability and comparison 97

5.12 Crash counts (a-c) and COMP values (d-f) for CTs at AG100, AG300 and AG500 (in order of appearance at each row, from more to less aggregated). Some central Districts of Valencia are highlighted in blue 98

5.13 Crash counts for CTs (a-c) and HEXAs (d-f) and OP values for CTs (g-i) and HEXAs (j-l) at AG100, AG200 and AG300 (in order of appearance at each row, from more to less aggregated). A CT in the South of Valencia is highlighted in blue 99

6.1 School locations in the city of Valencia distinguished by the academic level they offer. The central district of Valencia (city center) is highlighted with a thicker black line 104

6.2 Histograms and scaled densities along the timeline of school days (in 2014-2015) shown by traffic accidents that occurred close to a school and far from schools for a threshold distance of 50 m (a)-(b) and 150 m (c)-(d) 105

6.3 Hexagonal grid coloured according to accident counts observed at STW (a) and ETW (b). The school locations are represented with black squares (All-level), circles (Primary), diamonds (Preschool) and triangles (Secondary) 111

6.4 Relative risk curves derived from the source regression model with covariates for the four types of schools available in Valencia in STW (a) and ETW (b). The black dashed line indicates a relative risk of 1 112

6.5 Case and control buffer zones (100 m radius) defined around school locations in STW (a) and in ETW (b). The school locations are represented by black squares (All-level), circles (Primary), diamonds (Preschool) and triangles (Secondary) 116

7.1 The four spatial configurations of Valencia that were employed for the analysis: boroughs (a), Thiessen tessellations based on the main road intersections of the city (b), census tracts (c) and road network (d) 122

7.2 Four examples of simulated point patterns over a squared area of 2500 m^2: homogeneous Poisson process with $\lambda = 0.1$ (a), homogeneous Poisson process with $\lambda = 0.3$ (b), Matérn process with $(\lambda, s, \mu) = (0.1, 0.5, 1)$ (c) and Matérn process with $(\lambda, s, \mu) = (0.1, 4, 1)$ (d) . . . 126

7.3 Classification of the census tracts of Valencia according to the LISA indices obtained for robberies. Results from the complete dataset (a) and from a simulated 99% hit rate dataset (b). The census tracts whose border is coloured in green in (b) are affected by the removal of a sample of 106 events randomly selected (2% of the whole robbery dataset). The one in the city center is suddenly assigned to the high-high type, whereas the one in the northern area of the city is not detected anymore as a high-high zone 129

7.4 Locations of the robberies recorded in Valencia in 2014 and 2015. For experimental purposes, a higher geocoding error rate was assumed for the robberies occurred in avenues and squares (Experiment I) (a) and also for the robberies that took place in the city center and adjacent districts (Experiment II) (b) 130

7.5 Estimations of a minimum acceptable hit rate shown as a function of β (magnitude of the increased error) for the Experiment I and Experiment II, considering the four spatial structures available for Valencia for robberies (a) and car thefts (b) 130

7.6 Estimations of a minimum acceptable hit rate shown as a function of the intensity (λ) defined for each point pattern generated through an homogeneous Poisson process 132

7.7 Estimations of a minimum acceptable hit rate shown as a function of the intensity (λ) and clusters' radius (s) defined for each point pattern generated through a Matérn process. Empty circles correspond to the values obtained in the Simulation Study I for homogeneous Poisson processes 133

7.8 Estimations of a minimum acceptable hit rate shown as a function of the number of cells (n) and intensity (λ) defined for each point pattern generated through an homogeneous Poisson process. The values $\lambda = 5 \cdot 10^{-3}$ and $\lambda = 10^{-2}$ are represented with the same colour and shape because all the estimations were coincident. The points corresponding to $\lambda = 5 \cdot 10^{-4}$ and $\lambda = 10^{-3}$ for $n = 10$ also overlap 134

7.9 Logistic fit (blue curve) of the upper bounds of the 95% tolerance intervals obtained for the estimations of minimum acceptable hit rates obtained. The logistic curve corresponds to the parameters $\phi_1 = 99.07$, $\phi_2 = 9.46$ and $\phi_3 = 1.52$, according to Equation 7.2 135

8.1 Workflow that describes all the steps that could be carried out in order to perform a spatial statistics analysis on a point pattern that lies on a linear network. Some of these steps which lead to the final statistical analysis may be skipped but, at least, all of them should be considered. The blocks pointing the steps of the process include some of the R packages that would allow to successfully achieve each of them 141

8.2 Overview of the "Network Edition" section of the SpNetPrep application (a) and example of a road network uploaded into the application (b) 144

8.3 Use of the "Join vertex" (in green), "Remove edge" (in red) and "Add point" options (in green) in the SpNetPrep application (a), and resulting network after pressing "Rebuild linear network" (b) 144

8.4 Another use of the "Join vertex" (in green) option of the "Network Edition" section (a) that leads to the final corrected road structure in (b) 145

8.5 A road network introduced as input (a) and its simplified version after the application of the `SimplifyLinearNetwork` function with `Angle` = 25 and `Length` = 65 (b) 147

8.6 A zone of a road network introduced as an input in the "Network Direction" section of the SpNetPrep application (a) and manual addition of traffic flow to the network by using the options "Add flow" and "Add long flow" (b) 148
8.7 Example of a linear road network following usual notation for the edges (e_i) and vertex (v_i). Arrows represent the direction of traffic flow149
8.8 An example of a point pattern that lies on a road network as it can be visualized in SpNetPrep (a), and information that is displayed (marks of the point pattern, if available) when an event is clicked 150
8.9 Road network (in black) corresponding to the Near West Side Community Area of Chicago. Census tracts of the area are overlayed in blue. 153
8.10 Extracting a part of the road network structure analyzed from Chicago. Original structure extracted (left), made of 273 lines and 260 vertices, and simplified version of it (right) with 114 lines and 101 vertices. . . 156
8.11 Outputs from the function `PlotRelativeProbabilities` for the following choices of σ: 125 (a), 250 (b), 500 (c) and 1000 (d). 166
8.12 Outputs from the function `PlotHotspots` for the following choices of k and n: $k = 0.5$ and $n = 20$ (a), $k = 1.5$ and $n = 20$ (b), $k = 1$ and $n = 10$ (c) and $k = 1$ and $n = 30$ (d). 167
8.13 Outputs from the function `PlotHotspots` for the marks `year_after_2012`, `month_winter` and `hour_21_3` and the categorical value `Yes` for the three, considering the combinations of k and n that maximize the PAI_{type} (for the values of k and n tested). 168

List of Tables

2.1 Variables description and basic statistics, where SD denotes the standard deviation . . . 14

2.2 Mean values of the variables used to perform the clustering of the road segments according to their geometry . . . 18

2.3 Summary of the response (counts at the road segment level) for the original and the split road network . . . 24

2.4 Values obtained for the statistical tools employed for model comparison 25

2.5 Summary of the results obtained with Models 1 and 2. Coefficient estimates (β and γ) in bold represent covariates significant with 90% credibility, whereas Lo and Up denote the lower and upper bounds (respectively) of the 90% credible intervals for such estimates . . . 29

2.6 Relative frequencies of the covariates in the road segments that form the coldspots, average zones and hotspots detected. Each frequency is obtained by averaging the values of the covariates for all the road segments in each set, but weighting them according to their corresponding lengths. For the binary variables only the frequencies of presence at the road segment (value of 1) are shown . . . 30

3.1 Sample sizes, relative frequencies and Moran's I values (with associated p-values testing the null hypothesis of no spatial autocorrelation) of all the categories available for the two marks considered . . . 36

3.2 Description of the differential risk hotspots (denoted simply as hotspots in this Table) that are detected when applying the procedure with different values of k and n for $\sigma = 100$ m, where $N_{motorcycle}$ is the number of accidents involving motorcycles in the complete set of hotspots for two given values of the parameters, N is the total number of traffic accidents recorded in the same space, and $p_{motorcycle}$ is the proportion $N_{motorcycle}/N$. It can be appreciated that for some combinations of these two parameters no hotspots are determined . . . 44

3.3 Summary statistics for all the sets of differential risk hotspots (denoted simply as hotspots in this Table) that were determined as statistically significant by the Monte Carlo procedure (at the 0.05 level) for the collision and vehicle types considered, after the application of the detection methodology with $k = 1$ and $n = 40$. The table includes the number of differential risk hotspots that were found in each case, its total length (in kilometers), the number of traffic accidents of that type ($N_{type}^{hotspots}$) and in total ($N_{total}^{hotspots}$) that were recorded within the hotspots, the number of traffic accidents of that type ($N_{type}^{extended}$) and in total ($N_{total}^{extended}$) that were recorded within the hotspots or within an extension of 75 meters around them, the proportion $N_{type}^{hotspots}/N_{total}^{hotspots}$ ($p_{type}^{hotspots}$), the proportion $N_{type}^{extended}/N_{total}^{extended}$ ($p_{type}^{extended}$) and the global proportion of the type of accident along the complete road network $p_{type}^{network}$ (also available in Table 3.1) 48

3.4 Description of the sets of differential risk hotspots found for each collision and vehicle type considering the percentage of traffic accidents (for the corresponding type) and road length (in relation to the whole network) that the set represents. The quotient of these two percentages is equivalent to the PAI$_{type}$ value associated 48

4.1 Description of the covariates employed in order to adjust the results derived from the classical Knox test. For the formulas of β and π, e denotes the number of edges of a road network, v the number of vertex, T its diameter (longest distance reachable within the network) and L its total length 67

4.2 Summary statistics for the geometric and exposure related covariates at the hexagons forming the grid 67

4.3 Coefficients obtained from the modellization of burglary counts (using GLMs) with a set of geometric and exposure related covariates at different periods of the years 2016 and 2017. An asterisk (*) indicates significance for any of the coefficients at the 0.1 level 68

4.4 Adjusted ratios for $X_{\sigma,\tau}$ (observed/expected) with different selections of σ (in meters) and τ (in days) for the standard and adjusted Knox test, employing 5000 simulations. A double asterisk (**) indicates the significance of the ratio at the 0.01 level, whereas a single asterisk (*) means the same for the 0.05 level. The significance of the ratios is derived from the empirical p-values estimated for each σ and τ from the simulations provided by the Monte Carlo and Metropolis-Hastings procedures 69

4.5 Empirical p-values obtained for the Knox ratios shown in Table 4.4. The cells that are shaded correspond to $p < 0.01$ (darker colour) and $0.01 \leq p < 0.05$ (lighter colour) 69

5.1 Description of the covariates defined for the analysis and basic statistics of these covariates for the four BSU configurations tested (in their original configuration, prior to aggregation/regionalization) 77

5.2 Percentage of crashes located near BSU boundaries considering five distance thresholds (5, 10 and 20 m) and mean distance from crashes to BSU boundaries for the four BSU configurations employed in the analysis 78

5.3 Basic distributional properties of the response variable (crash counts) corresponding to each BSU type and aggregation level, where CV means coefficient of variation (standard deviation to mean ratio) . . . 88

5.4 Estimates with standard deviation (SD) for the parameters involved in the CAR model, considering CTs as BSUs. An asterisk (*) indicates the significance of a parameter with a 90% credibility 89

5.5 Estimates with standard deviation (SD) for the parameters involved in the CAR model, considering TMs as BSUs. An asterisk (*) indicates the significance of a parameter with a 90% credibility 90

5.6 Estimates with standard deviation (SD) for the parameters involved in the CAR model, considering TIs as BSUs. An asterisk (*) indicates the significance of a parameter with a 90% credibility 91

5.7 Estimates with standard deviation (SD) for the parameters involved in the CAR model, considering HEXAs as BSUs. An asterisk (*) indicates the significance of a parameter with a 90% credibility 92

5.8 Nearest neighbour indexes (NNI) and p-value associated with each index. An asterisk (*) indicates the statistical significance of the index ($p < 0.05$), which indicates clustering (NNI<1) or dispersion (NNI>1) 92

5.9 Models performance in terms of DIC and AIC (for CAR and GWR models, respectively), MAD, SAD and PMAD for all the BSU configurations and levels of aggregation tested 98

6.1 Description and classification of the covariates defined for the analysis 103

6.2 Risk ratios for each combination of distance threshold (σ), school type or combination of types and time window (STW or ETW). Risk ratios in bold were found significant at the 0.1 level. The ratios for $\sigma = 25$ m at STW were unreliable due to the small samples available and are not shown 109

6.3 Estimates (Est.) and 5th, 10th, 90th and 95th percentiles of the posterior distributions of the parameters involved in the CAR model for both time windows considered (STW and ETW). Estimates shown in bold are significant with 90% credibility 110

6.4 Parameter estimates obtained for the multiple source regression models in the two time windows, STW and ETW. Parameters λ_1, λ_2, λ_3 and λ_4 were associated with the number of complex intersections, main road length, number of bus stops and land use entropy, respectively. The indexes for the α's and β's represent: 1 = All-level, 2 = Preschool, 3 = Primary, 4 = Secondary 112

6.5 Estimates (Est.) and 5th, 10th, 90th and 95th percentiles of the posterior distributions of the parameters involved in the logistic model for both time windows considered (STW and ETW). Estimates shown in bold are significant with 90% credibility 115

7.1 Minimum acceptable hit rates (MAHR) estimated for the five crime datasets available for Valencia, along with the lower and upper bounds of the empirical 95% tolerance interval obtained for each estimation. The point patterns corresponding to each of the datasets are described according to their size (number of events), level of clustering (NNI) and intensity ($\hat{\lambda}$). The number of spatial units (areal or linear) of each spatial structure considered is also indicated. It is worth noting that $\hat{\lambda}$ is measured in $1/\text{m}^2$ in the case of boroughs, Thiessen polygons and census tracts, but in 1/m in the case of the road network structure127

7.2 Minimum acceptable hit rates (MAHR) estimated for the five crime datasets available for Valencia, along with the lower and upper bounds of the empirical 95% tolerance interval obtained for each estimation, considering the GLM approach and the clusters and outliers derived from LISA indices 128

7.3 Proposals of minimum acceptable hit rates (MAHR) for planar point patterns of certain intensity levels according to the model fit shown in Figure 7.9 135

Chapter 1

Introduction

In 1855, an English physician called John Snow presented a map showing the locations of cholera cases in the London epidemic of 1854 (Snow, 1855). He investigated the pattern formed by these locations and suggested that a public water pump on Broad Street was the source of the disease outbreak. The creation of this map and the subsequent analysis of the epidemic is widely considered as the first example of a spatial data analysis and one of the origins of epidemiology (Susser and Bresnahan, 2001).

Since 1855, many types of spatial data have emerged, especially during the last decades. Indeed, spatial statistics is the field of knowledge specifically devoted to the analysis of all kind of spatial data. There are three main types of spatial data, namely *point patterns*, *areal data* and *geostatistical data* (Cressie, 1993), which have given rise to the development of three main lines of research among spatial statistics researchers. The following paragraphs include a description of these three data types and a brief summary of the techniques that have traditionally been employed for the analysis of each of them. Before that, it is worth noting that spatial data usually allows incorporating a temporal component, which leads to the formation of spatio-temporal data. Hence, there exists many specific methodologies that generalize the fundamental tools coexisting for the three main subfields of spatial statistics to fit within the context of spatio-temporal data.

A *point pattern* is a set of points that lie on a certain space or region. Each point of a spatial point pattern is usually referred to as an event. The stochastic mechanism that generates a point pattern is called a *spatial point process* (Illian et al., 2008; Diggle, 2013). Given a point pattern, the most common goal is to investigate its structure. In particular, it is usually of interest to determine whether the points of the pattern present some dynamic of attraction or inhibition or, in contrast, if they satisfy the hypothesis of complete spatial randomness (Diggle, 2013). The joint use of non-parametric techniques such as the K-function (Ripley, 1977) with multiple model-based approaches (Diggle, 2013) facilitate the investigation of point patterns.

Areal data are observations from a random process over a countable number of spatial regions or units. The number of times that a type of event has occurred in each administrative unit of a city over a period of time constitutes a very typical example of areal data. The employment of spatially-lagged models (Anselin, 1988) and generalized linear models accounting for the spatial autocorrelation between the regions (Banerjee et al., 2004) is a common practice to explain a response variable on the basis of a set of covariates. Furthermore, detecting regions where a variable presents a particularly high value (hotspots) is another very important application (Anselin, 1995; Kulldorff, 1997).

Geostatistical data corresponds to a collection of measurements available over a continuous region. Geostatistics was initially conceived for ore reserve estimation, but the possible applications of this branch of spatial statistics are countless. The *variogram* and *kriging* are possibly the two capital concepts in the field of geostatistics (Cressie, 1993; Krige, 1960; Matheron, 1963).

Despite the persistence of this classification of spatial statistics into three subfields, it has been argued that the most important distinction should be made between spatially continuous and spatially discrete stochastic processes (Diggle et al., 2013). In addition, there is an existing trend to unify all kind of spatial data and techniques and, in particular, to use model-based geostatistics to deal with areal data and point patterns (Diggle and Giorgi, 2019). In order to work in this direction, log-Gaussian Cox processes are a fundamental tool (Møller et al., 1998).

Spatial and spatio-temporal statistics can be applied to many fields of research. Besides epidemiology and geology, which have already been mentioned, spatial data often arises in the context agriculture, astronomy, biology, criminology, hydrology, meteorology, remote sensing and traffic safety, among others. Indeed, the collaboration of both the Local Police of Valencia and the Spanish National Police with the Department of Statistics and Operations Research of the University of Valencia was a decisive factor to orient this thesis to the investigation of spatial and spatio-temporal data of special interest in the fields of criminology and traffic safety, and to the development of statistical techniques that facilitate the undertaking of this capital objective.

More specifically, the main objective of this thesis was to identify “research gaps” among the extensive literature on the use of spatial-temporal statistical methods for the analysis of crime and traffic safety data. As part of the development of this thesis, the investigation has focused on different support structures for data analysis, on statistical modelling of spatial data, on the spatio-temporal study of contagious events, on the quality of spatial data, or on the development of specific software, among others. In all cases, the aim was to improve the existing methodology, both to address some rather theoretical issues and to carry out practical studies more accurately.

The following sections provide a description of several relevant topics that are usually involved in the spatial and spatio-temporal analysis of crime and traffic safety data. First, section 1.1 is dedicated to introduce briefly a singular spatial structure called linear network. Although several of the analyses that are presented in this thesis are performed at the area level, the employment of linear networks spans across a substantial part of the thesis. Second, the different studies that have been carried out for the development of this thesis are described and contextualized throughout Sections 1.2.1 and 1.2.2, outlining the whole structure of the document.

1.1 Linear networks

Road network structures were introduced in the context of spatial statistics some years ago, providing the basis for analyzing events typically lying on such structures, which are more generaly referred to as linear networks. Indeed, a planar linear network, L, is defined as a finite collection of line segments, $L = \cup_{i=1}^{n} l_i$, in which each segment contains the points $l_i = [u_i, v_i] = \{tu_i + (1-t)v_i : t \in [0,1]\}$ (Ang et al., 2012; Baddeley et al., 2015, 2017). Following graph theory nomenclature, these segments are sometimes referred to as the edges of the linear network, whereas the points that determine the extremes of such segments are known as the vertices of the network.

Hence, a point process X on L is a finite point process in the plane such that all points of X lie on the network L (Ang et al., 2012; Baddeley et al., 2015, 2017). Similarly, a collection of events that is observed on L is known as a point pattern, x, in L.

The investigation of spatial patterns lying on linear networks is gaining attention in the last years. The design of new and more accurate/efficient kernel density estimators (McSwiggan et al., 2017; Moradi et al., 2018, 2019; Rakshit et al., 2019b), the introduction of graph-related intensity measures (Eckardt and Mateu, 2018), the construction of local indicators of spatial association (Eckardt and Mateu, 2017), or the estimation of relative risks (McSwiggan et al., 2019) are some topics that are starting to be developed for linear networks.

1.2 Applications

Spatial and spatio-temporal techniques are massively applied in the context of both criminology and traffic safety. Next sections highlight several types of analyses that are of particular interest in these two fields, from both a methodological and a practical perspective.

1.2.1 Traffic safety analysis

Research on traffic safety dates back at least to the 40s of last century. Gordon (1949) suggested the convenience of treating injuries and deaths that originate from traffic accidents as one more type of epidemiological data. Hence, the two main goals of traffic safety analysis have been the determination of factors that are associated with more traffic accidents (or with a certain typology of accident) and the accurate location of zones where traffic accidents are especially likely, which are usually called hotspots.

Both goals lead researchers to take a spatial approach. The choice of a spatial model to measure the relationship between a response variable representing accident counts or rates at a certain level of resolution and a set of covariates defined on the same scale is vital to avoid overlooking the likely presence of autocorrelation in the outcomes, and hence possibly bias and confound the estimations of model parameters. On the other hand, hotspot detection is a common practice in many fields that produce spatial data, particularly traffic accident data. Many methods are available for this purpose, which again require considering the spatial characteristics of the data.

Already in the 70s, the use of spatial techniques for analyzing traffic safety data became a usual practice. Moellering (1976) proposed the use of computer animated films to observe the spatio-temporal distribution of traffic accident patterns, which can be considered as a seminal contribution that was followed by many other studies published in the next years and decades. A few years ago, Lord and Mannering (2010) summarized many of the issues that deserve consideration for performing the statistical analysis of a traffic accident dataset. Besides certain questions such as controlling for overdispersion, underreporting or low sample sizes, the necessity of considering the presence of spatial and temporal autocorrelation in the data was also highlighted. In fact, more recently, Mannering et al. (2016) reviewed the methods and modelling approaches that can be chosen to incorporate spatial effects into traffic safety analysis. In the typical context of modelling the accident counts observed on a set of regions, the use of negative binomial and zero-inflated models is recommended to account for overdispersion and the possible excessive presence of zeros in the response. In addition, several suitable modelling techniques are indicated: random parameters count models, random parameters tobit model, random parameters generalized count models, latent-class (finite mixture) models, etc.

Linear networks are a suitable support to face the two main objectives that have been described in the previous paragraphs, although they entail certain technical difficulties. Indeed, modelling, for instance, accident counts at the road segment level is more subject to errors, by far, than analyzing accident outcomes on the basis of an areal subdivision of a region. More precisely, representing a road structure through a linear network implies simplifying certain urban features, which can complicate the location of some geocoded accident in the proper road segment. Given this difficulty, which arose at the first stage of the development of this thesis, the author developed the R package SpNetPrep (Spatial Network Preprocessing). This package allows users to perform the manual curation of a linear network representing a road network structure with the aid of a Shiny-based interactive application. This application provides users the possibility of adding and removing vertices and edges of the network, the capability of adding a direction to the network according to traffic flow, and the opportunity of inspecting and editing a point pattern lying on the road network. Furthermore, the package also contains a function that automatically simplifies the linear network structure by merging, under certain conditions of length and angle, the edges of the network that are joined by a second-degree vertex. Chapter 8 of this thesis contains a larger description of this package.

With the aid of SpNetPrep, several spatio-temporal analyses of traffic accident data have been carried out over the road network of Valencia, constituting a substantial part of this thesis. First, a district of Valencia was selected to perform a spatial modelling of traffic accident counts at the road segment level (Briz-Redón et al., 2019e). The necessity of curating both the road network structure and the traffic accident dataset provided by the Local Police of Valencia led to choose only one district of Valencia for this analysis. The main objective of this study was to find covariates or factors that increase the incidence of traffic accident counts at the street level. In particular, data modelling accounted for the presence of road intersections across the network under investigation. It is usually observed that many of the accidents that occur in either a urban or a rural area are located in or near to a road intersection, so it was convenient to consider the proximity of an intersection when defining the statistical models. More specifically, the original road network was split to distinguish between intersection and non-intersection road segments. Furthermore, the modelling of the data was combined with coldspot/hotspot detection, which was mainly based on network-constrained density estimation and local indicators of spatial association (Anselin, 1995), to gain a better understanding of the factors possibly involved with traffic accidents. Chapter 2 is totally dedicated to this analysis.

The second study that was carried out focused on the detection of hotspots along the road network of Valencia where a type of traffic accident is overrepresented (Briz-Redón et al., 2019b). The methodology proposed by Kelsall et al. (1995) to produce relative risk surfaces was adapted to allow the estimation of the relative probability of occurrence that corresponds to a type of event that takes place along a linear network. Then, an strategy is proposed to detect differential risk hotspots on the network, according to the relative probability values previously inferred. Both the magnitude of the probability estimate and the sample size involved in its calculation are accounted for the determination of the hotspots. The full procedure is described in Chapter 3, which includes an application with a traffic accident dataset providing information on the collision type and on the types of vehicles involved in each accident. The R package DRHotNet (Differential Risk Hotspots in a Linear Network), which is available in CRAN, implements the complete differential risk hotspot procedure. This R package is depicted in Chapter 8.

Even though the analysis of traffic accidents at the road segment level provides the highest level of accuracy for both researchers and traffic safety professionals, the study of accidents at an areal level is still of interest. Specifically, analyzing the traffic accident counts that are recorded on a collection of administrative units or police areas during a period of time facilitates the consideration of certain environmental or socio-demographic variables to assess their relationship with traffic accidents. In this regard, the choice of a particular areal unit of analysis can affect the statistical results that one carries out. This issue is known as the modifiable areal unit problem (MAUP) in the field of spatial statistics (Openshaw, 1984). This thesis includes a case study (Briz-Redón et al., 2019c) on the consequences of varying the scale or the zoning of a collection of regions in the context of traffic safety analysis (Chapter 5). The effects of modifying any of these two factors on both the response and the covariates, and the impact on model estimation and performance are investigated in detail.

1.2.2 Criminology

The study of crime from a spatial perspective started to gain certain importance in the 50s of last century. For instance, Shannon (1954) analyzed the spatial distribution of crime rates across American states. In the decade of the 70s the discipline started to consolidate through multiple studies focused on explaining the spatially-varying incidence of the most relevant crime types (Block, 1979; Brantingham and Brantingham, 1975; Georges, 1978; Harries, 1973; Stephenson, 1974). At the beginning of the 80s, the foundations of spatial criminology were definitely established by Patricia and Paul Brantingham (Brantingham and Brantingham, 1981, 1984).

The use of linear networks to perform a spatial analysis has become essential in criminology to properly capture the distribution of crime events across space. Indeed, last years are bringing to the field many studies focused on the evaluation of the law of crime concentration on a city or region. The law of crime concentration 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). Hence, it is necessary to analyze spatial crime data at a proper scale to truly appreciate the level of crime concentration. The employment of linear networks and hence the fact of locating crimes at the road segment level usually yields more realistic levels of event concentration than that provided by areal units of analysis.

There are several mechanisms that explain the spatially-varying incidence of crime in the short-, mid- and long-term. The presence of criminogenic features in a city can stimulate, facilitate or complicate criminals' activities and, as a consequence of this, explain crime rates in the mid- and long-term. Three main types of places are distinguished in literature: crime attractors, crime generators and crime detractors (Brantingham and Brantingham, 1995; Kinney et al., 2008). Crime attractors are places that constitute a singular context that naturally offers more opportunities with regard to committing a crime. Basically, the locations of a city where certain type of crime is common (for instance, someplace where drug dealers and consumers usually meet) are also likely to attract other criminal actions. On the other hand, crime generators are places that attract the presence of many people and hence provide more targets to criminals. A shopping mall or a football stadium are two examples of crime generators. Finally, a crime detractor is a place that makes the development of criminal activities more complicated. A police station is an example of crime detractor because it implies a higher presence of capable guardians (police officers) within a neighbourhood from the station. There exists multiple statistical modelling approaches to assess if one feature (or set of features) of a city have an influence on crime outcomes (Briz-Redón et al., 2019a). The different methods that are compared in Chapter 6, even though the application shown in it does not correspond to a crime dataset (it is again related to traffic safety), are all of potential use in this context.

On the other hand, explaining crime occurrence in the short term is intimately connected with the phenomenon of crime repetition. Virtually all types of crime have been investigated in the context of repetition or near-repetition, which refers to the increased likelihood of observing a crime in the proximity (in space and time) of a previous criminal event of the same type. Hence, some authors have referred to the phenomenon of crime repetition as a epidemiological contagious process (Loftin, 1986). Indeed, the Knox test originally conceived for epidemiological studies (Knox and Bartlett, 1964) is still the main tool to assess the magnitude and spatio-temporal extent of the near-repeat phenomenon for a type of crime and a spatio-temporal window, although new approaches are gaining importance (Mohler et al., 2011; Reinhart and Greenhouse, 2018).

The main drawback of the Knox test is that it does not account for spatio-temporal risk heterogeneity, which in the context of crime repetition complicates the distinction between near-repeats that are connected (explained by the so-called "boost" theory) and those that are unconnected (explained by the so-called "flag" theory). In Chapter 4 the classical version of the Knox test is adjusted following the work of Schmertmann (2015). One modification of the proposal made by Schmertmann is also provided (Briz-Redón et al., 2020). The adjusted version of the test is implemented to analyze the magnitude and extent of the near-repeat phenomenon considering a dataset of burglaries recorded in Valencia.

The final research that was conducted in the context of this thesis deals with geocoding quality and the presence of missing data as a consequence of non-geocoded events. This issue is of interest for every discipline that requires using spatial data, but it has been particularly emphasized in the context of quantitative criminology. In 2004, J. Ratcliffe declared that an 85% hit rate (percentage of geocoded events) was a minimum acceptable geocoding hit rate for conducting a spatial analysis (Ratcliffe, 2004a). Chapter 7 contains a reestimation of this rate by accounting for some spatial factors (intensity, clustering and aggregation levels) that were overlooked in the first estimation. Furthermore, the procedure proposed in Ratcliffe (2004a) (based on the Mann-Whitney test) is extended to other statistical methods of interest. Thus, it has been found that the 85% initially proposed may be too low under certain conditions (Briz-Redón et al., 2019d).

Chapter 2

Modelling traffic accident counts at the road segment level: Accounting for road intersections

This Chapter includes a statistical modelling of traffic accidents at the road segment level. In order to represent true neighbouring relationships between road segments, a directed road network structure accounting for traffic flow has been used. There were three methodological objectives: to present and discuss some issues that arise when conducting a spatial analysis of traffic accidents located on a road network, to analyze traffic accidents at road intersections, including a specific strategy that draws together both road intersection and non-intersection zones along the network, and to combine the results produced by the two statistical approaches finally chosen, spatial count models and coldspot/hotspot detection, in order to achieve more complete conclusions regarding the effect of various road characteristics on the occurrence of traffic accidents for the road network of interest.

2.1 Introduction

Traffic accidents are still a quite frequent cause of death for the European population, especially in the younger age groups. Even though the number of accidents has gradually decreased in the most developed countries of the world during the last decade, many efforts, in terms of prevention and road planning, are still being made to reduce their occurrence and severity. In this regard, studies aimed at analyzing the occurrence and distribution of traffic accidents can be very helpful, and could be broadly classified according to three main objectives: finding road and/or traffic characteristics associated with a higher occurrence of accidents, detecting zones with a high concentration of accidents and discovering the types of accidents that tend to produce more serious consequences for the vehicle passengers or road users involved. In this study, the modelling of accident counts at the road segment level with explanatory purposes is the main goal, although the detection of microzones of the network that show a singular risk of accident is also carried out. This section starts with a literature review on both topics: modelling traffic accidents outcomes and finding zones of high accident risk. This is followed by a review of the literature on the analysis of traffic accidents occurring in intersection and non-intersection zones. This issue has also been addressed in the analysis contained in this study.

2.1.1 Review of models and methods

Many important quantitative studies that have focused on factors that may be affecting traffic safety have been carried out through areal units of analysis. For instance, Quddus (2008) modelled traffic accident counts at the census ward level, which made it possible to explain the number of accidents from information related to traffic characteristics (volume and speed), road design and socio-demographic factors. Traffic volume and a proxy for poverty showed a significant positive association with traffic accidents. Similarly, Huang et al. (2010) studied traffic accident frequency at the county level considering traffic-related, demographic and socioeconomic characteristics of the counties being studied. This work focused on distinguishing two types of exposure variables: population and average daily vehicle miles travelled (DVMT) per county. The model using DVMT as the exposure yielded more significant associations with traffic accidents, some of which were positive (traffic intensity, density of principal and minor arterials, and percentage of young population) and others were negative (freeway density and average travel time to work).

In order to favour more accurate investigations, road networks have increasingly been used in traffic safety analysis in the last few years. The use of these structures, composed of links (segments) and vertices (points where two or more links meet), is becoming more popular, in spite of the technical difficulties their use entails. In this regard, it is worth noting that many of the factors that have been proved to generally increase the occurrence of traffic accidents require a road segment level analysis. Indeed, the list of infrastructure characteristics that were determined to be risky for drivers and users in a recent systematic review of published studies by Papadimitriou et al. (2019) included traffic volume, road surface (low friction), low curve radius, number of lanes, absence of paved shoulders, narrow shoulders, different junction types, etc. Although an areal-based analysis may also help to gain knowledge about the association of traffic accidents with any of these road characteristics, a road segment level analysis would be recommended to guarantee an appropriate investigation.

Given the convenience of using road networks for analyzing traffic accident outcomes, this paragraph includes a description of several studies that were performed at the road segment level, which enabled their corresponding authors to properly investigate certain infrastructure characteristics. For example, Aguero-Valverde and Jovanis (2008) found a positive association of traffic volume and certain shoulder widths with traffic accidents. In addition, Guo et al. (2017) developed a measure (integration) which reflects the accessibility of a node in the network, depending on its neighbourhood geometry. It was found that networks with a high integration value, which usually resemble a grid pattern, tend to be associated with more traffic accidents. Finally, Barua et al. (2016) analyzed severe and no-injury traffic accidents at the road segment level, finding that road segment length, average annual daily traffic, density of unsignalized intersections, business land use and the number of lanes showed a significant and positive association with both accident types.

On the other hand, several studies that have incorporated linear networks to treat accident datasets have only focused on detecting zones with a high concentration of accidents (hotspots). Indeed, Huang et al. (2016) suggested that the detection of hotspots at the micro-level is more accurate and useful for revealing risky road configurations than the use of areal macro-zones. For example, Xie and Yan (2013) applied kernel density estimation (KDE) to a linear network structure to evaluate distribution of traffic accidents and to find clusters of roads with a high proportion of accidents. They studied the impact of subdividing the network into shorter spatial units (segments), called lixels (Xie and Yan, 2008), and the variations observed depending on the choice of the kernel bandwidth parameter. A similar approach was taken by Nie et al. (2015) to prove that the application of network KDE improved the performance of local indicators of spatial association (LISA) to better identify accident hotspots.

To finish the literature review, it is of need to highlight certain aspects that deserve attention every time a statistical modelling of accident counts is performed. First, regardless of the choice of areal units or road segments for conducting the analysis, the consideration of spatial effects has almost become a requirement (Mannering and Bhat, 2014; Mannering et al., 2016). Getting back to some of the studies described above, some of them showed that the use of non-spatial models can lead to either spatially autocorrelated model residuals (Quddus, 2008; Huang et al., 2010) or to a significantly lower fit to the data (Aguero-Valverde and Jovanis, 2008). Both facts suggest that overlooking spatial effects is inappropriate. Moreover, Xu et al. (2017) tested a modification of the model proposed in Huang et al. (2010) that allowed the effects of the covariates to vary spatially. These authors observed that it is even advisable to include such variations, as otherwise biased estimates of the model's coefficients may arise.

Besides the consideration of spatial heterogeneity, other issue that often arises when performing a road segment level modelling of accident counts is the high presence of zeros (segments where no accident has been recorded). Zeng et al. (2017) used a Tobit model to control for left-censored accident rates that may be the consequence of under-reporting. Speed was associated with higher crash rates, whereas average annual daily traffic displayed a significant negative correlation. Anastasopoulos (2016) compared multivariate Tobit and zero-inflated models for modelling accident counts with a high percentage of zeros. Both strategies showed their own limitations, but each was capable of capturing zero-state heterogeneity across the road network.

2.1.2 Traffic safety at road intersections

The high rates of traffic accidents that are usually observed in proximity to road intersections is the reason for the existence of many studies on this topic. Thus, this paragraph includes a literature review (in chronological order) on the topic of modelling the occurrence of traffic accidents around road intersections. For instance, Castro et al. (2012) studied the spatio-temporal incidence of accident counts at urban intersections. It proved to be advisable to consider both the spatial and the temporal effect, and a significant effect was found for roadway configuration, approach roadway typology and traffic flow, among other factors. Xie et al. (2014) also developed several modelling approaches to analyze accident occurrence at intersections. The consideration of a hierarchical spatial model accounting for the effects produced at intersections by contiguous segments (corridor-level) clearly outperformed the rest of the models applied. Huang et al. (2017) analyzed accident counts at road intersections considering types of users (pedestrians, bicycles or motor vehicles) involved in accidents with a multivariate Poisson lognormal regression model. Moreover, Lee et al. (2017) used a mixed effects negative binomial model accounting for macro-level and micro-level factors to study accident counts at road intersections. Several covariates constructed at both levels of spatial resolution were found to be associated with more accidents at intersections. Cai et al. (2018) implemented a grouped random parameters multivariate spatial model at two levels, segments and intersections. Covariates were defined separately over segments, intersections and wider zones (allowing the inclusion of covariates, such as socio-economic characteristics, at a lower spatial resolution). Zhao et al. (2018) used multivariate Poisson log-normal and zero-inflated univariate and multivariate Poisson models to study accident frequency (by severity level) at signalized intersections, consisting of the road segments at 200 ft upstream from the signal controlling the intersection. Lastly, Alarifi et al. (2018) proposed the use of a multivariate hierarchical Poisson lognormal model that accounts for the spatial relationships between road segments and intersections located along the same corridor. Average annual daily traffic variables at roadway segments and intersections, absolute speed limit difference between a major and a minor road meeting at an intersection, and driveway density showed positive associations with the number of traffic accidents.

With regard to the distance threshold of 200 ft chosen by Zhao et al. (2018), it needs to be highlighted that the definition of intersection-related traffic accidents presents a low level of agreement. For instance, Miaou and Lord (2003) considered a distance of 15 m ($\simeq$50 ft) from intersection locations, Ye et al. (2009) 75 m ($\simeq$250 ft), Zhao et al. (2018) 60 m ($\simeq$200 ft) and Das et al. (2008) analyzed the range 0 to 60 m at increments of 15 m. Furthermore, besides the lack of agreement, intersection zones are not clearly defined in many of the papers in the field. Finally, it is also highly remarkable how several of these papers focused entirely on intersection entities alone, avoiding consideration of the road segments surrounding intersections. Although Lee et al. (2017) and Cai et al. (2018) considered zonal effects that were shared by close segments and intersections, only Alarifi et al. (2018) have conducted a unified consideration of road segments and intersections. In this regard, Miaou and Lord (2003) pointed out the advisability of modelling data taking all kind of road entities simultaneously, which according to these authors would include segments, intersections and ramps. The implicit assumption of independence between entities may lead to ignoring many important spatial relationships that are likely to exist between them.

The rest of the Chapter is structured as follows. The next section contains a complete description of the data employed for the analysis, including the traffic accident dataset recorded during the period of study and the network structure that represents the underlying space where these accidents occurred. This is followed by a methodological section that provides a description of the procedure followed to include the consideration of intersection zones, the definition of spatial neighbourhoods between road segments of the network, the specification of the spatial count models used to fit the data, the methods employed to assess the performance of such models, the definition of one class of network-constrained kernel density estimation and the procedure applied to locate zones of high and low risk along the network. Finally, there is a discussion of the performance and implications of the methods applied.

2.2 Data

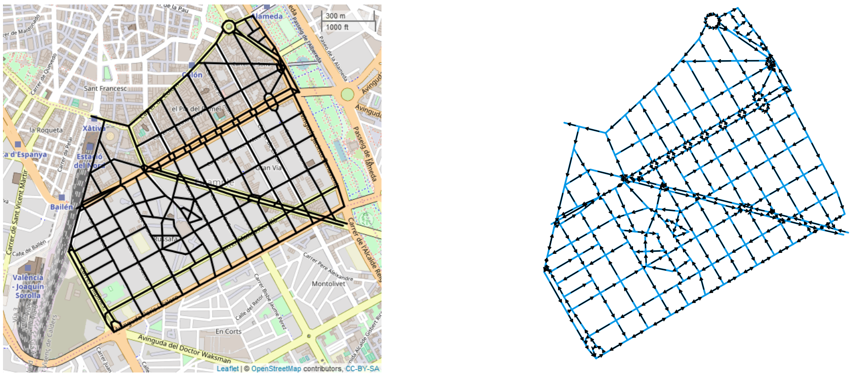
2.2.1 Accident information

A total of 5738 traffic accidents recorded by the Local Police Department of the city of Valencia (Spain) during the years 2005 to 2017 in the Eixample District of the city were used. Each of these accidents was geocoded from the address information recorded by the Police minutes after the accident had occurred. Once the coordinates of each accident were obtained, these were projected onto a linear network representing the traffic streets of the Eixample District of Valencia. This two-stage process was revised carefully in order to ensure a high level of accuracy.

2.2.2 Network structure

A linear network composed of 279 vertices and 444 road segments, representing a total length of 33.57 km, was used for the analysis. The vertices where more than two segments meet correspond to road intersections, which were 227 in the case of this network. Figure 2.1a contains a map (Graul, 2016; OpenStreetMap contributors, 2017) that shows the zone of the city of Valencia where the road network of interest is located. Some parts of this network were previously simplified without altering its basic geometrical structure in order to reduce the number of short road segments which could hinder the subsequent modelling of the data. Moreover, network preprocessing included the slight modification of highly complex intersections and the removal of pedestrian streets, which were performed with the SpNetPrep R package (Briz-Redón, 2019).

In addition, for the purpose of improving the analysis, the network was given directionality according to the traffic flow of this district of Valencia as of the end of December 2017 (see Figure 2.1b). Some of the road segments of the network were defined as bidirectional, representing two-way streets present in the district where no median strip separates the two flows of vehicles. However, bidirectional road segments were only 5% of the total, a fact that completely justifies the definition of traffic flow directionality along the network. In addition, for road segments divided



(a)

(b)

Figure 2.1: Road structure of study displayed over a map of the city of Valencia (a) and its representation as a linear network made of links and vertices, with arrows indicating traffic flow directionality (b)

by a median strip two (parallel) road segments were available in the network at a distance proportional to the width of the strip.

Finally, the possible changes in direction of traffic that could have been made during the period of years considered have not been taken into account due to the difficulty of tracking them. However, as Eixample District is very close to the centre of Valencia and is part of a very well-established area of the city, it can be assumed that changes of traffic direction must have been minimal in the period 2005-2017.

2.2.3 Network-related covariates

Several factors that could be associated with vehicle collisions are considered at the road segment level. These mainly include the presence of specific public services in the road segment (parking slots, traffic lights and bus stops) and basic characteristics of the roads that the links in the network represent. The latter include the number of lanes in the road, the presence of a bus lane (binary), the type of road (main or not, binary), the number of roads that directly connect to each road segment of the network, distinguishing whether they allow traffic to enter or leave it, a categorical covariate representing average annual daily traffic (AADT) and a categorical covariate assigning a geometric typology to each road segment (this one is described in the Methodology). In this regard, it is worth noting that AADT is not available for every road segment in Valencia, but only for the most travelled avenues and streets. Hence, the data available was used to define a 5-level categorical covariate representing the following ranges for AADT: <7000 (level 1), 7000-16000 (level 2), 16000-25000 (level 3), 25000-55000 (level 4) and >55000 (level 5). These ranges represent the least travelled road segments of the city for which scarce data is available (level 1) and the four quartile-based intervals that follow from the available AADT values (levels 2 to 5). It is worth noting that the strategy of categorizing AADT values has already been tested by several authors (Hao and Daniel, 2014; Fan et al., 2015; Yasmin et al., 2016).

Furthermore, numbers of lanes and neighbouring roads (referred to here as neighbours) were truncated and recoded for values higher than 5 and 3, respectively. To obtain the number of neighbours the network that was considered was actually an extension of the final one employed for the analysis, in order to avoid an unrealistic low number of neighbours for the road segments at the edge of the network. Finally, as the network of study represents a fairly small and homogeneous population area, it was concluded that the inclusion of demographic or socioeconomic variables was not of interest. Table 2.1 includes a description and statistical summary of the covariates introduced in this section.

Variable	**Description**	**Mean**	**SD**
Main road	Main road segment of the city (binary)	0.626	0.484
Parking slots	Existence of public parking slots in the road segment (binary)	0.613	0.671
Traffic light	Presence of a traffic light in the road segment (binary)	0.617	0.487
Bus stops	Existence of public bus stops in the road segment (binary)	0.110	0.314
Bus lane	Presence of a bus lane in the road segment (binary)	0.572	0.495
No. of lanes	Number of traffic lanes in the road segment	2.176	1.349
No. of in-neighbours	Number of neighbouring road segments allowing traffic to enter the road segment	1.770	0.631
No. of out-neighbours	Number of neighbouring road segments allowing traffic to leave the road segment	1.775	0.629
AADT	Average annual daily traffic (5 levels)	2.286	1.539

Table 2.1: Variables description and basic statistics, where SD denotes the standard deviation

2.3 Methodology

2.3.1 Software

The R programming language (3.4.1 version, R Development Core Team, Vienna, Austria) (R Core Team, 2018) was used to obtain all the results presented in this study. The R packages bayesplot (Gabry and Mahr, 2018), brms (Bürkner et al., 2017), ggmap (Kahle and Wickham, 2013a), spatstat (Baddeley et al., 2015), spded (Bivand and Piras, 2015) and SpNetPrep (Briz-Redón, 2019) were specifically required for performing the analysis and the data curation process.

2.3.2 Definition of intersection zones

In order to capture the differential risk between road locations around intersections and road segments between them, the original network structure was modified by creating shorter road segments in the proximity of each road intersection. The insertVertices function of the R package spatstat (Baddeley et al., 2015) was key for performing this task.

Specifically, road segments of 20 meters were inserted around intersection neighbourhoods (so that the furthest point of the segment from the intersection was at a distance of 20 m), which were determined to be intersection analysis zones (IAZs). On the other hand, segments not satisfying this condition, most of which are between two IAZs, were declared as middle analysis zones (MAZs). Thus, the original network of study was divided into 683 IAZs and 292 MAZs, leading to the formation of a new road network (referred to from now on as a split network) made up

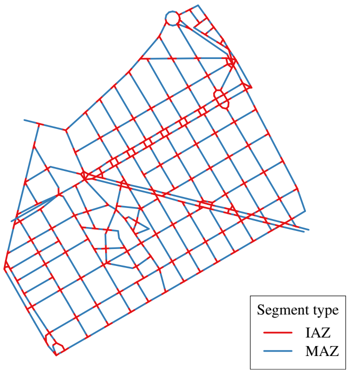


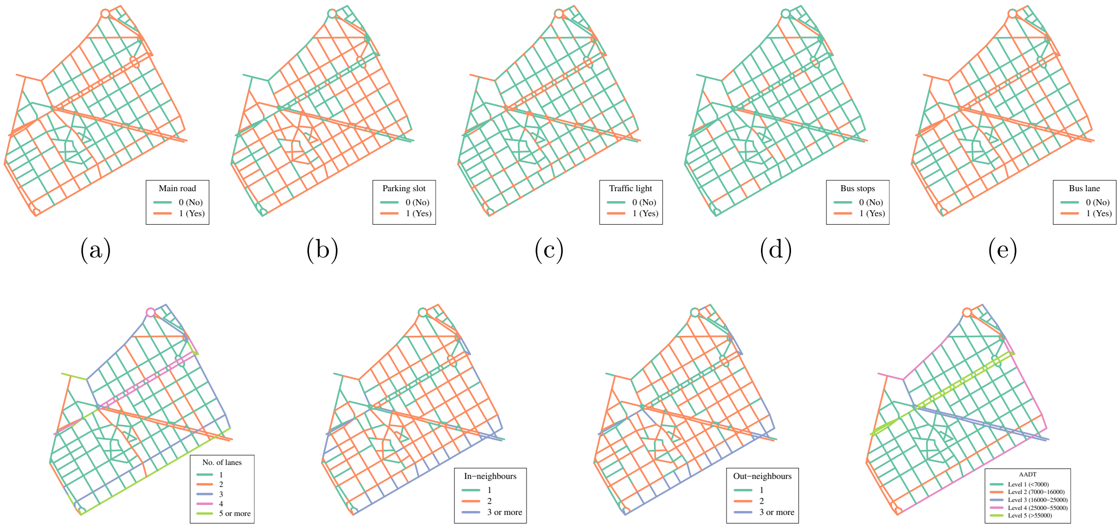
Figure 2.2: Graphical description of the split network showing the locations of IAZs and MAZs

of 810 vertices and 975 road segments (the original had 279 vertices and 444 road segments). As an illustration, Figure 2.2 displays the distribution of IAZ and MAZ along the split network.

Therefore, the definition provided for IAZ and MAZ allowed for the coexistence of street zones subject to different rules and causalities while being represented by a unique geometrical entity: the road segment (note that the sum of the number of IAZ and MAZ coincides with the number of road segments of the final network). This fact led to a unified definition of neighbourhood relationships and covariates for the two types of zone that mainly arise when dealing with traffic accident datasets. Indeed, the term segment is used without distinction for both types of zone throughout the Chapter, even though in related literature it is only used for what it has been defined as MAZs.

The choice of a distance of 20 meters was mainly based on knowledge of the road network of study and on similar distances used in literature (Miaou and Lord, 2003). Indeed, this distance allows a fair representation of IAZs as intersection-approaching or intersection-leaving segments. The selection of a shorter threshold distance was rejected due to the lack of sufficient certainty on the data collection procedure to guarantee the correct location of accidents at such a level of resolution around intersections. Furthermore, the objective was to employ the road segment as the only spatial unit of study, and this would be undermined if a threshold very close to 0 were chosen (as the IAZ would almost become the intersection point itself).

Regarding the definition of the covariates at the level of the new split road network, these simply follow the values available for the original network. Hence, each IAZ or MAZ of the split road network acquires the value (for a given covariate) of the corresponding whole road segment in the original non-split network. An exception was made with traffic lights, given their frequent location around road intersection zones. For this reason, a value of 1 was assigned to an IAZ or MAZ for the indicator related to traffic light presence if and only if a traffic light was present in the same road segment (before splitting) at a distance lower than 20 m from the middle point of the IAZ/MAZ. As an illustration, Figure 2.3 provides a graphical description



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

(f) (g) (h) (i)

Figure 2.3: Graphical description at the road segment level of the following network-related variables: (a) main road indicator, (b) parking slot presence, (c) traffic light presence, (d) bus stop presence, (e) bus lane presence, (f) number of lanes, (g) number of in-neighbours (h) number of out-neighbours and (i) AADT

of every covariate considered for the analysis, which enables us to appreciate the distinction made for traffic lights (Figure 2.3c).

2.3.3 Concept of neighbourhoods

The road segments that form the already defined directed network structure constitute the basic spatial units on which to perform the statistical analysis. Given a road segment, i, in the directed linear network, its neighbourhood, $N(i)$, can be defined in four different ways depending on whether the traffic flow information available is used. At the simplest level, if this information is not used, two road segments i and j are neighbours if they are connected by a vertex of the network. However, the use of the traffic flow leads to the definition of three other types of neighbourhoods. First, neighbourhood between i and j can be established if it is possible to travel from i to j or from j to i, in either direction, without passing through another road segment of the network; this is denoted $N_{dir}(i)$. In addition, if a distinction is made between travelling from i to j or vice versa, it is possible to separate the neighbouring road segments that allow you to reach i ($N_{dir}^{in}(i)$) from those that allow you to leave from i to another road segment of the network ($N_{dir}^{out}(i)$) (see Figure 2.4 for examples of all these types of neighbourhood). From now on these last two types of neighbours are referred to as in-neighbours and out-neighbours, respectively.

The four definitions of neighbourhood structures can lead to the construction of four different adjacency matrices. Thus, a W_{dir} matrix based on N_{dir} neighbourhoods was the only one employed as it was considered the most suitable for the goals established. Regarding this matrix, its entries, w_{ij}, are called weights and it holds

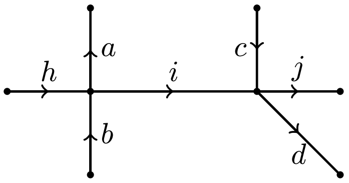


Figure 2.4: Examples of types of neighbourhood in a directed linear network. The six road segments that are contiguous to road segment i allow the construction of the neighbourhoods $N(i) = \{a, b, c, d, h, j\}$, $N_{dir}(i) = \{b, d, h, j\}$, $N_{dir}^{in}(i) = \{b, h\}$ and $N_{dir}^{out}(i) = \{d, j\}$

that $w_{ij} = 1/|N(i)|$, if $j \in N(i)$ (row normalization), and 0 otherwise.

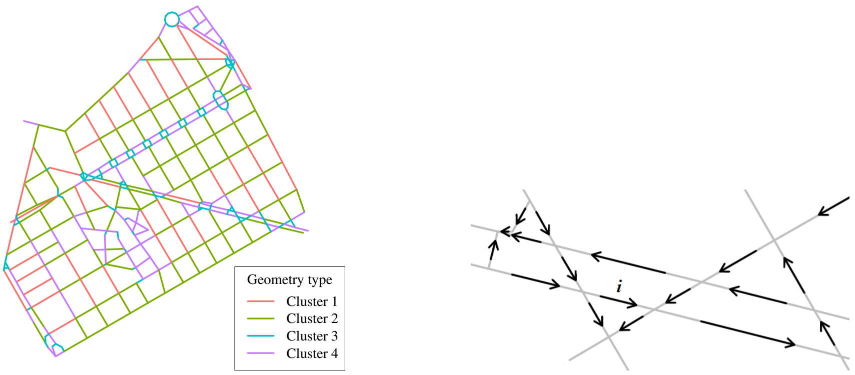
2.3.4 Road segment neighbourhood geometry

The geometric structure surrounding each road segment of the network was studied, a procedure made possible by the road network structure. Given a road segment of the network, the factors considered for each neighbouring road segment were the neighbourhood type (in or out) and the angles formed between the road segment and its neighbours. Road segment length and the number of in and out neighbours were also included to better discriminate between road segments. As was done with the other covariates previously defined (with the exception of the indicator factor for traffic lights), the geometry is studied from the perspective of the original network. Later, the values obtained are assigned to the road segments of the split network accordingly.

A total of six types of neighbours were defined by combining the angles of the road segments and the direction of the traffic. Angles between road segments were classified (measured in [0°, 180°]) into three groups: straight (]150°, 180°]), right (]60°, 120°[) and sharp ([0°, 60°]$\bigcup$[120°, 150°]). Each of these types of angle was then crossed with the in/out information associated with each neighbouring road segment to create the six possible scenarios.

The same strategy was followed with the lengths of the neighbouring road segments. In this case, the road segments were divided into three groups (short, medium and long) according to the 33.33% and 66.67% quantiles of the road segment length distribution. Again, the three groups created were crossed with the in/out information, producing six new classification groups.

The k-means algorithm (Hartigan and Wong, 1979) was then applied to a total set of fifteen geometric-related variables for each of the road segments: the number of neighbours belonging to each of the six angle-direction and length-direction combinations, the road segment length and the number of in and out neighbours. A value of $k = 4$ was chosen, since convergence was not reached for higher values of k, and this made it possible to form four clusters of 42, 130, 163 and 109 road segments, respectively.



(a)

(b)

Figure 2.5: (a) Clustering of the road segments of the spatial network according to their neighbourhood geometry. (b) Detailed neighbourhood of a road segment, i, of the directed network. Six road segments share a vertex with i, but only four of these allow traffic to flow from i or to i. The values of the geometric variables for road segment i are: Straight$_{in}$ = 1, Right$_{in}$ = 0, Sharp$_{in}$ = 1, Straight$_{out}$ = 1, Right$_{out}$ = 0, Sharp$_{out}$ = 1, Short$_{in}$ = 1, Medium$_{in}$ = 1, Long$_{in}$ = 0, Short$_{out}$ = 1, Medium$_{out}$ = 0, Long$_{out}$ = 1, Length = 41.7, $|N_{dir}^{in}(i)| = 2$, $|N_{dir}^{out}(i)| = 2$

Table 2.2 summarizes the mean values for the variables employed in the clustering procedure and Figure 2.5 includes the graphical representation of the four clusters and an example of construction of the geometric-related variables for a specific road segment. According to these results, Cluster 2 is mainly composed of medium-long road segments with a high average number of neighbouring road segments that form a right angle, which is associated with being part of a crossroads (90-degree intersection). Cluster 3 is formed by very short road segments, a high proportion of which involve acute angles, representing abrupt changes of direction in the directed network. Cluster 1 clearly presents the highest road segment length and an high number of neighbours. Finally, Cluster 4 is made up of short-medium length road segments and quite high connectivity with short-length road segments compared with Clusters 1 and 2.

Cluster	Straight$_{in}$	Right$_{in}$	Sharp$_{in}$	Straight$_{out}$	Right$_{out}$	Sharp$_{out}$	Short$_{in}$	Medium$_{in}$	Long$_{in}$	Short$_{out}$	Medium$_{out}$	Long$_{out}$	Length	$\|N_{dir}^{in}\|$	$\|N_{dir}^{out}\|$
1	0.45	0.81	0.48	0.40	0.74	0.74	0.21	0.64	0.88	0.29	0.69	0.90	170.66	1.83	1.98
2	0.45	0.93	0.63	0.42	0.98	0.59	0.28	0.58	1.15	0.27	0.57	1.15	115.84	2.17	2.11
3	0.21	0.46	0.88	0.21	0.52	0.86	0.75	0.48	0.31	0.74	0.50	0.34	23.80	1.58	1.61
4	0.25	0.59	0.84	0.29	0.53	0.78	0.53	0.75	0.39	0.52	0.77	0.31	68.40	1.77	1.78

Table 2.2: Mean values of the variables used to perform the clustering of the road segments according to their geometry

2.3.5 Accident count modelling

A Bayesian spatial model with Zero-Inflated Negative Binomial response (ZINB) was implemented to fit the observed accident counts for the split network structure (composed of 975 road segments). If $Y \sim \text{NB}(\mu, \psi)$ (basic negative binomial distribution of mean μ and shape ψ) then it holds that $E(Y) = \mu$, $V(Y) = \mu + \frac{\mu^2}{\psi}$ and $P(Y = x) = (\frac{x+\psi-1}{\psi-1})(\frac{\psi}{\mu+\psi})^{\psi}(\frac{\mu}{\mu+\psi})^{x}$. The zero-inflated version of the NB distribution acts as a double-stage process that makes it possible to increase the probability of value 0. Thus, if z denotes the structural probability of 0 for the ZINB distribution, its probability mass satisfies the next stepwise function:

$$P(Z=0) = \begin{cases} z + (1-z)P(Y=0) & , x = 0 \\ (1-z)P(Y=x) & , x > 0 \end{cases}$$

where $Y \sim \mathrm{NB}(\mu, \psi)$ and $Z \sim \mathrm{ZINB}(\mu, \psi, z)$.

Then, on the basis of the choice of a ZINB distribution for the response (accident counts) the next spatial model (Model 1) was specified:

$$Y_i \sim \mathrm{ZINB}(\mu_i, \psi, z)$$

$$\log(\mu_i) = \log(\mathrm{Length}_i) + \mathbf{x_i}\boldsymbol{\beta} + \phi_i \text{ (Model 1)}$$

where Y_i is the number of accidents observed at road segment i, μ_i and ψ are the mean (for road segment i) and overdispersion (shape) values for the ZINB distribution, z is the probability of value 0 for the ZINB distribution, the natural logarithm acts as a link function for the mean risk at segment i (μ_i), the natural logarithm of each segment's length is added as an offset, $\mathbf{x_i}$ is a vector that contains the values for the covariates described in Table 2.1 corresponding to segment i along with a factor indicating whether the road segment belongs to the IAZ class, $\boldsymbol{\beta}$ is a vector of coefficients to control the effect of these predictors and ϕ_i represents a spatial effect for road segment i.

The spatial effect was modelled using a conditional autoregressive (CAR) structure (Besag, 1974; Besag et al., 1991):

$$\phi_i \mid \phi_j, j \neq i \sim N(\alpha \sum_{j=1}^{n} w_{ij}\phi_j, \tau_i^{-1})$$

where $\alpha \in [0,1]$ is a spatial dependence parameter that measures the strength of spatial autocorrelation ($\alpha = 0$ reflects the complete absence of such effect), τ_i is a precision parameter that varies with i and w_{ij} the entry at the (i,j) position of the neighbourhood matrix W_{dir} ($w_{ii} = 0, \forall i$).

In particular, the joint distribution of $\Phi = (\phi_1, ..., \phi_n)$ satisfies the Gaussian multivariate probability distribution (Banerjee et al., 2004):

$$\Phi \sim N(0, [\tau(D - \alpha W_{dir})]^{-1}).$$

where D is a diagonal matrix that contains the number of in- and out-neighbours of each spatial unit.

Moreover, a second model (Model 2) specifically focused on the effect of the covariates on IAZs was implemented with the following linear predictor for $\log(\mu_i)$:

$$\log(\mu_i) = \log(\mathrm{Length}_i) + \mathbf{x_i}\boldsymbol{\beta} + \mathbf{x_i}\boldsymbol{\gamma} I_{\mathrm{IAZ}} + \phi_i \text{ (Model 2)}$$

where I_{IAZ} is an indicator function for IAZ and $\boldsymbol{\gamma}$ represents the vector of coefficients that measure the effect of the covariates at IAZ. Hence, Model 1 only considers the effect of IAZ as one of the factors being studied, whereas Model 2 allows the determination of the differential effect that each covariate can produce at the road segment level depending on the zone of analysis (IAF or MAF). Furthermore, for these two models, the inflation probability, z, was modelled through a logit equation that makes it possible to estimate a different value of z for each zone type:

$$\text{logit}(z) = z_{\text{Intercept}} + z_{\text{Slope}} I_{\text{IAZ}} \longleftrightarrow z = \frac{\exp(z_{\text{Intercept}} + z_{\text{Slope}} I_{\text{IAZ}})}{1 + \exp(z_{\text{Intercept}} + z_{\text{Slope}} I_{\text{IAZ}})} \tag{2.1}$$

where I_{IAZ} is again an indicator function for IAZ.

The estimation of the parameters of the two models was performed with the brms R package (Bürkner et al., 2017), which is based on the statistical software Stan (Carpenter et al., 2017).

2.3.6 Model checks

Several techniques were applied in order to check for the propriety of the different models employed for representing the observed accident counts. In this section, the methods used for this task, which included conditional predictive ordinate (CPO), general correlation coefficients and Moran's I, are briefly described.

The CPO method (Stern and Cressie, 2000; Marshall and Spiegelhalter, 2003) requires data simulation from the posterior distribution of a fitted Bayesian model. Indeed, if the values for the covariates of the models are left fixed as in the data used to fit the model, the accident counts simulated at each draw behave like replicates of the original counts (y) and are denoted by Y^{rep} (Gelman et al., 2013). If a model represents the counts properly, the observed counts should agree with the distribution of a simulated dataset of Y^{rep}. Then, a high departure between y and Y^{rep} may indicate a poor performance from the model. In this regard, CPO is a simulation-based tool that has already been used in similar research studies (Yang et al., 2013; Xie et al., 2014) with the main purpose of identifying outliers within the data, which in this case correspond to road segments. For this purpose, the distribution of Y_i^{rep} for every spatial unit (road segment) i is evaluated from all the original data except y_i itself (in a similar way to the leave-one-out cross-validation procedure). Thus, the goal is to find spatial units whose observed count value is far enough from the simulated distribution of $Y_i^{rep}|y_{-i}$, where y_{-i} denotes the original data with the exclusion of y_i. The determination of a p-value that tests this question for unit i is done through a reweighting of the Y^{rep} with the choice of the following weight:

$$\rho_{-i}^{(k)} = \frac{1}{P(y_i|\Lambda^{(k)})}$$

where k is the index for the simulation, y_i is the number of accidents observed for spatial unit i, $\Lambda^{(k)}$ represents the parameters sampled for the model at simulation number k (which includes the corresponding values for λ_i's, ψ, z and Φ) and P represents the probability function of a ZINB distribution that follows the parameters in $\Lambda^{(k)}$. Then, a p-value that allows outlier identification is approximated with the next expression (Marshall and Spiegelhalter, 2003):

$$P(Y_i^{rep} \leq y_i | y_{-i}) \approx \sum_{g=0}^{y_i-1} \frac{\sum_{k=1}^{K} P(Y_i^{rep} = g|\Lambda^{(k)})\rho_{-i}^{(k)}}{\sum_{k=1}^{K} \rho_{-i}^{(k)}} + \frac{1}{2} \frac{\sum_{k=1}^{K} P(Y_i^{rep} = y_i|\Lambda^{(k)})\rho_{-i}^{(k)}}{\sum_{k=1}^{K} \rho_{-i}^{(k)}} \tag{2.2}$$

General correlation coefficients are an extension of Pearson's correlation coefficient (Pearson, 1896) making it possible to compare two possibly related numerical vectors of the same length. Specifically, it can be employed to compare the distribution of ranks shown by the observed accident counts and the counts fitted by any statistical model applied. The formula for a general correlation coefficient, Γ, is:

$$\Gamma = \frac{\sum_{i=1,j=1}^{n} a_{ij} b_{ij}}{\sqrt{\sum_{i=1,j=1}^{n} a_{ij}^2 \sum_{i=1,j=1}^{n} b_{ij}^2}}$$

where the coefficients a_{ij} and b_{ij} must be anti-symmetric ($a_{ij} = -a_{ji}$, $b_{ij} = -b_{ji}$). As two important particular cases, if r^{obs} and r^{exp} denote the ranks (in decreasing order) of the observed and fitted accident counts per spatial unit (respectively), the following choices of a_{ij} and b_{ij} correspond to Kendall and Spearman correlation coefficients (Kendall, 1938; Spearman, 1904):

$$a_{ij} = \text{sgn}(r_i^{\text{obs}} - r_j^{\text{obs}}),\ b_{ij} = \text{sgn}(r_i^{\text{exp}} - r_j^{\text{exp}})$$

$$a_{ij} = r_i^{\text{obs}} - r_j^{\text{obs}},\ b_{ij} = r_i^{\text{exp}} - r_j^{\text{exp}}$$

where $\text{sgn}(x) = x/|x|$ (sign function). A high value of Γ, regardless of the specific selection of a_{ij} and b_{ij}, indicates a high level of agreement between the ranked observed accident counts and the ones predicted by a model. This is a clear sign of a good model fit.

Finally, Moran's I (Moran, 1950a,b) consists in a global estimation of the spatial autocorrelation of a variable indexed in according to a system of spatial units. Its definition is the following:

$$I = \frac{\sum_{i=1}^{n} \sum_{j \in N_{dir}(i)} \frac{1}{n_i}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

where x_i is a variable indexed by spatial unit and $\bar{x}$ its average. Thus, Moran's I makes use of the predefined neighbourhood structure and behaves as a correlation between the variable of interest and a variable that assigns to each of the spatial units a weighted average of the values of its neighbours. Under the hypothesis of no spatial autocorrelation, it holds that $E(I) = -1/(n-1)$, where n is the number of spatial units (975). Hence, negative Moran's I values for the residuals of the model would be an indicator of a good performance from the fitted Bayesian count models in capturing the spatial effect.

2.3.7 Kernel density estimation

Kernel density estimation (KDE) is commonly used to obtain the intensity of a point pattern that lies on a space. Particularly, it can be used to estimate the intensity of a point pattern along a linear network, requiring modifications of the classical formulas (valid for areal units) to account for the particularities of this spatial structure (Okabe et al., 2009; Okabe and Sugihara, 2012). In this study, the equal-split continuous kernel density defined by McSwiggan et al. (2017) is computed at the middle point of every road segment i of the linear network following the next equation, by which the $f_\sigma(i)$ values are obtained:

$$f_\sigma(i) = \sum_{x \in A(m_i,\sigma)} k(d_L(x, m_i))a^C(\pi) \tag{2.3}$$

where m_i is the middle point of the road segment i of the linear network, σ is the kernel's bandwidth, $A(m_i, \sigma)$ is the set of points of the network at a distance from m_i up to σ where an accident took place, $k(u) = \frac{1}{\sigma\sqrt{\pi}}e^{(\frac{-u}{\sigma})^2}$ is the kernel function (Gaussian), d_L is the distance along the network and $a^C(\pi) = \prod_{j=1}^{m} \frac{2}{\deg(v_j)}$, where $\pi = [v_1, ..., v_m]$ denotes the set of vertices of the network that have to be passed through to travel the shortest path that joins m_i with x and deg() represents the degree of a vertex of the network, meaning the number of road segments incident to the vertex. It needs to be remarked that the computation of the distance d_L between any two points of the network makes use of its directed structure, providing a realistic measure of the distance between the two points according to traffic flow.

A Gaussian kernel was selected because it is the most common option, and no other kernel functions were explored because this choice usually has little effect on the results (Silverman, 2018). On the choice of the bandwidth parameter, a value of around $\sigma = 50$ m would be optimal if the non-parametric test proposed by Cronie and Van Lieshout (2018) were followed. However, the larger value of $\sigma = 100$ m was applied in agreement with previous studies on road networks that have employed KDE for hotspot detection (Xie and Yan, 2013; Nie et al., 2015).

Finally, edge effects (Okabe and Sugihara, 2012) need to be discussed because the network used for the analysis is to a certain extent artificially bounded, being only a part of the larger road network of Valencia. First, it needs to be remarked that the kernel construction chosen (Equation 2.3) alleviates edge effects, as stated by McSwiggan et al. (2017). Second, Eixample District is delimited by pedestrian and secondary roads (to the north and south), a train station (to the west) and a green area (to the east), which to some extent make the district naturally bounded (Figure 2.1a enables us to appreciate some of these points). Furthermore, the two roads bordering the network of analysis to the north and south are important avenues of Valencia which account for most of the accidents in the vicinity (these avenues are part of the network analyzed). All these facts allow us to conclude that accident densities estimated along the four roads that form the border of the network are reasonable.

2.3.8 Coldspot/hotspot detection

The use of the count models was supplemented with a search for zones of the network with a particularly low or high incidence of traffic accidents; these are usually known as coldspots and hotspots, respectively. Several approaches to this problem coexist in recent literature on traffic accident data, including some of the studies already mentioned in the Introduction (Xie and Yan, 2013; Nie et al., 2015; Thakali et al., 2015; Harirforoush and Bellalite, 2016). These methods mainly agree in the use of KDE to obtain a smooth representation of the observed point pattern, a process which is commonly followed by the detection of zones of the network whose KDE values present a significant spatial autocorrelation.

Here, KDE was computed with $\sigma = 100$ m at the middle point of each road segment of the network considering the d_L distance along the network that accounts for traffic flow (following Equation 2.3). Then, the local version of Moran's I statistic known as LISA (Anselin, 1995) was obtained for each road segment following the next formula:

$$I_i = \frac{x_i - \bar{x}}{\sum_i (x_i - \bar{x})^2/n} \sum_j w_{ij}(x_j - \bar{x})$$

The road segments showing a significant local association (a threshold of 0.1 was used for the p-value instead of the usual 0.05 to minimally extend some of the coldspots/hotspots, allowing a wider part of the network to be analyzed) were selected and grouped according to their contiguity. Other inputs such as the accident count per road segment or the accident rates were also considered for computing the LISA values, but KDE was the only one capable of providing a sensible number of zones along the network presenting similar behaviour in terms of dangerousness. Finally, the basic average intensity of the point pattern (number of events per unit length) in each of the zones of interest was compared with the mean intensity in its first-order neighbourhood (the set of all the first-order neighbours of the road segments composing the zone), which made it possible to detect both low-intensity (coldspot) and high-intensity (hotspot) parts of the network showing a differential incidence of traffic accidents in comparison with the road segments in their surrounding areas.

Finally, once the coldspots and hotspots had been located in the network, the values of the covariates of the road segments that formed them were individually analyzed to confirm or put into question the conclusions that could be drawn from the use of the count models.

2.4 Results and discussion

A total of four Monte Carlo Markov chains (MCMC) of length 30000 were run for the two models starting from non-informative priors for the parameters involved.

The length of the chains was chosen to be large enough to ensure the convergence of the estimates of all the parameters involved in the models, which was afterwards checked using common validation tools (scale reduction factor close to 1 for all estimates and visual inspection of the chains). The choice of a ZINB model is sensible according to the values in Table 2.3, which was validated through subsequent predictive checks of a graphical nature. First, accident counts are clearly overdispersed with a variance-to-mean ratio of 31.22 for the original network. Second, 23.2% of the road segments (103 of the 444 road segments that form the original network) have no accidents recorded for the period being considered, which leads to the choice of a zero-inflated response. Furthermore, the inequality observed in the accident counts per road segment leads to a Gini index (Gini, 1912) of 0.67, with more than the 50% of the accidents recorded concentrated in 52 of the segments of the original network (these segments represent only 15% of the length of the network), in agreement with previous studies of a similar nature (although not focused on traffic accidents) referring to the law of crime concentration (Weisburd, 2015). These particularities of the data of study are also present when the network is split into IAZs and MAZs. However, whereas the variance-to-mean ratio is not so heavily affected, the number of zeros is much higher in IAZs. For this reason, the estimation of the zero-inflated probability was made dependent on the zone, as described in the previous methodological section regarding count model specifications.

	Original network	**Split network**	
		IAZs	**MAZs**
Mean	12.92	5.75	6.21
Variance	403.46	189.94	123.25
Variance/Mean	31.22	33.06	19.84
% Zeros	23.20	49.19	18.84
Gini Index	0.67	0.80	0.66
No. of accidents	5738	3924	1814
No. of segments	444	683	292
Road length (m)	33571.09	14166.96	19404.13

Table 2.3: Summary of the response (counts at the road segment level) for the original and the split road network

Table 2.4 displays the values obtained for the three kinds of validation tools that were applied. Moran's I values were negative for both models, which is a sign of good performance as it indicates the absence of spatial autocorrelation between model residuals. Correlation coefficients (Γ) derived from the comparison of observed and expected counts at the road segment were very similar and significantly greater than 0 for both models, although a slight improvement can be appreciated for Model 1. Regarding the percentage of potential outliers according to the CPO method (IAZs and MAZs showing a p-value lower than 0.05 according to Equation 2.2), the results are again very close, but better again for Model 1. In conclusion, Model 1 presents better results than Model 2 by a narrow margin, but both models offer a reasonable basis to allow conclusions to be drawn from them regarding the occurrence of traffic accidents in the network of analysis.

	Model 1	**Model 2**
Γ (Kendall)	0.37	0.34
Γ (Spearman)	0.47	0.43
% Potential outliers	13.85	15.28
Moran's I	-0.06	-0.04

Table 2.4: Values obtained for the statistical tools employed for model comparison

Therefore, let us now concentrate on model parameters and on the effects that the covariates being considered could have had on the accidents that occurred in the network of study during the period 2005-2017. Table 2.5 shows the results for Model 1 and Model 2, in which the missing levels of any covariate are implicitly present as they are considered the reference levels for the covariate (the other levels are estimated in relation to the missing one). If the estimation for the coefficient corresponding to a covariate (β's and γ's) or a structural parameter (ψ, z and α) lies in the 90% credible interval (all derived from the MCMC procedure), then the effect of that covariate or structural parameter is significant with 90% credibility. All structural parameters were found to be significant in all models. Hence, a modelling approach that includes spatial heterogeneity ($\alpha > 0$), overdispersion ($\psi > 0$) and a zero-inflated distribution that depends on the zone (MAZ or IAZ) is justified. Parameters ψ and z driving the ZINB distribution present very similar estimates for the two models. The higher percentage of zeros in IAZs is clear from the estimates obtained for the slope parameter ($z_{\text{Slope}} > 8$), which models z through a logit equation. This particularly means (for instance, for Model 1) that $z = \frac{\exp(-8.69+8.32)}{1+\exp(-8.69+8.32)} \simeq 0.41$ in IAZs (following Equation 2.1), a value that is not surprising in view of Table 2.3. On the other hand, $z = \frac{\exp(-8.69)}{1+\exp(-8.69)} < 2 \cdot 10^{-4}$ indicates that zero-inflation is not needed for modelling accident counts in MAZs, that is, a non-modified NB distribution would be suitable enough.

With regard to covariate effects, Model 1 indicates that main roads, roads containing a bus lane and approaching-intersection segments (IAZs) are associated with a higher accident count. In contrast, the existence of parking slots in the road and geometries of type 3 and 4 correlate with fewer traffic accidents at the road segment level for the network of study. Regarding these two geometry types, as mentioned previously, they mainly include short and sharp road segments (Cluster 3) and short-medium segments that are highly connected with the latter (Cluster 4). The sign of the coefficient related to Cluster 3 may be considered inconsistent with literature reporting higher crash risks for skewed road intersections (Harwood et al., 2000; Nightingale et al., 2017; Kumfer et al., 2019), which is supported by the fact that skewed intersections cause longer traverse times than 90-degree intersections and poor visibility for drivers, among other things (Gattis and Low, 1998). However, it is worth noting that most of the research regarding skewed intersections is based on high-speed rural intersections. The Eixample network analyzed in this study represents a low-to-moderate-speed urban area. No significant associations are found for the rest of the covariates, including the multilevel categorical ones representing the number of lanes in the road, the number of entrances/exits and the AADT level.

On the other hand, Model 2 provides a more complex depiction of the effect of the covariates being studied, as it considers a differential effect for each of them depending on the zone type (IAZ or MAZ). Among the β parameters, which now represent effects within MAZs, only the one representing the effect of Cluster 3 remains significant (with the same sign as in Model 1). Despite not being significant with 90% credibility, the effects of main roads and the presence of a bus lane should not be completely overlooked according to the confidence intervals obtained. In addition, Model 2 points out the different contribution that some factors may make to the risk of traffic accidents in IAZs or MAZs. Road geometry reflected by Cluster 4 now appears as significant only for IAZs, presenting a even more negative coefficient than in the case of Model 1. Moreover, the association of the two highest levels of AADT, 4 and 5, with traffic accidents presents a differential behaviour between MAZs and IAZs. Indeed, both β parameters are significant and positive, suggesting an increase in the number of traffic accidents in MAZs, but the two corresponding γ parameters are negative, indicating a protective effect (or, at least, a less detrimental effect) against traffic accidents in the most travelled road segments when a road intersection is near. Therefore, the distinction between IAZs and MAZs has allowed us to find a significant association between some AADT levels and traffic accidents that depends on the proximity to road intersections, a result that somehow compensates the surprising (according to previous research) non-significant estimates found for the AADT levels in Model 1.

The computation of network-constrained KDE values with $\sigma = 100$ m leads to the smooth representation of traffic accidents shown in Figure 2.6a. This Figure shows that one of the main avenues located in the network (which also has the highest values for AADT) has a very high accident rate along all its length. Similarly, avenues and main roads bordering the network contain some zones of high accident rates. In contrast, the central part of the network shows much lower values than neighbouring locations. Therefore, the KDE values computed at the middle points of the 975 segments forming the split network were used to find (through LISA values) coldspots and hotspots accurately located in the network, which are displayed in Figure 2.6b. Identifying coldspots and hotspots enables us to compare the values presented by the covariates in the road segments forming them, but also in the rest of the network. Table 2.6 contains the mean values (weighted by each road segment's length) of the covariates at coldspots, hotspots and average road segments (neither a coldspot nor a hotspot), which enables us to check the high relative frequency of main roads, bus lanes, 4 or more lanes, 3 or more in- and out-neighbours and levels 4 and 5 of AADT in the road segments belonging to hotspots in comparison to those in coldspots or in average microzones. On the latter, it must be remembered that many of these covariates or levels were not yielded as a significant factor by the count models. Finally, the geometries of type 3 (for IAZs), 2 and 4 (for MAZs) are particularly high in hotspots, which may be unexpected according to the results shown by the two models fitted. In this regard, it is worth remarking that one should not expect road characteristics particularly represented in hotspots/coldspots to display a significant association with traffic accidents from a global modelling perspective. Hence, the combination of two statistical methodologies can either strengthen the validity of the conclusions or call them into question.

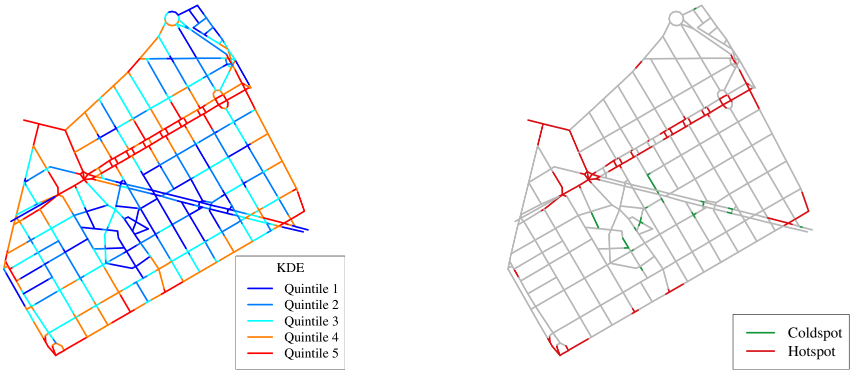
2.5 Conclusions

Traffic safety analyses set over areal spatial units have been of interest for many years, but the recent development of statistical techniques on linear networks is bringing new advances and challenges for this subject. Specifically, in this study, a linear network has been used to analyze a geocoded dataset of accidents that took place in the city of Valencia (Spain) during the period 2005-2017. In this regard, the proper consideration of road intersections and the combination of several statistical techniques have been emphasized.

Indeed, the study of traffic accidents around road intersections is of special interest given the high percentage of them that occur close to them. Typically, these analyses are done independently of the values observed for road segments between intersections. This strategy can potentially lead us to miss important relationships (mainly of a spatial type) between intersections and segments in between which may detract from the validity of the results. In this article, the definition of IAZs and MAZs along the directed linear network available has provided a unified approach (involving spatial relationships and the definition of covariates) to this issue that does not exclude any type of road entity.

On the other hand, from a modelling perspective, the coexistence of multiple methodologies to treat accidents datasets provides a flexible framework for analyzing many kinds of specific questions of interest, but this fact also leads to great difficulties when trying to decide on a particular approach. In this study, overdispersion of accident counts and the disparate effects that arise at road segments near intersections, producing both a high concentration of traffic accidents and a high presence of zeros, were addressed through a zero-inflated negative binomial distribution. In addition, spatial relationships between road segments were included with a CAR distribution based on a neighbourhood matrix that accounted for traffic flow. Later, model quality was assessed employing several validation tools, including checks based on simulated data that led to outlier detection, but also more classical techniques such as correlation coefficients and Moran's I.

Furthermore, this study has combined the use of spatial count models with the detection of coldspots and hotspots. The results derived from each of the approaches have been discussed and compared, providing coherent results even though some differences were noted. This kind of local analysis could be very useful for validating the results from the statistical models and and questioning some of the conclusions yielded by the former, increasing the robustness of the final results. In this regard, the nature of KDE alleviates the existence of geocoding inaccuracies that may arise when conducting a spatial analysis of this kind, especially when it is done at the road segment level. Indeed, the risk of making mistakes as a consequence of bad geocoding are higher for the construction of the response variable representing accident counts at the road segment level. Here, a small inaccuracy can lead to situating a traffic accident in the wrong road segment, altering the counts of two segments. Kernel density estimation, however, produces a smooth representation of the intensity of traffic accidents along the network that can even absorb some of the geocoding



(a) (b)

Figure 2.6: Quintile distribution of the KDE values for $\sigma = 100$ m (a) and coldspots (green) and hotspots (red) detected after the computation of LISA statistics from these KDE values for the split network made of 20 m IAZ. In (a), each road segment of the split linear network is coloured according to the KDE value at its middle point

inaccuracies that usually occur, as suggested by Harada and Shimada (2006) and Zandbergen (2009).

Overall, the modelling approach revealed that spatial heterogeneity, overdispersion and the effect of road intersections on adjacent road segments (including zero-inflation) must be accounted by analyzing the distribution of accident counts in the Eixample District of Valencia. The generalization of these findings to other urban areas may be risky, because this kind of analysis is always data-dependent, but it should always be reasonable to consider it. In addition, the detection of hotspots and coldspots identified the fact that main roads, the existence of a bus lane in the road, 4 or more lanes and high AADT values are associated with higher accident counts at the road segment level. The effect of other covariates remained unclear or non-significant and may require further analysis. In any case, this study was slightly limited in terms of covariates, which should be addressed in the future with the availability of more complete and accurate geographic information systems.

	Model 1			Model 2		
Covariate	β	Lo	Up	β	Lo	Up
(Intercept)	**-3.01**	-3.42	-2.60	**-3.43**	-3.99	-2.88
Main road	**0.41**	0.04	0.78	0.50	-0.09	1.09
Parking slots	**-0.27**	-0.51	-0.02	-0.14	-0.49	0.20
Traffic light	-0.08	-0.30	0.14	-0.20	-0.68	0.28
Bus stops	0.02	-0.25	0.30	0.21	-0.19	0.60
Bus lane	**0.36**	0.04	0.68	0.39	-0.09	0.88
No. of lanes (2)	-0.29	-0.64	0.05	0.01	-0.50	0.53
No. of lanes (3)	0.20	-0.23	0.65	0.00	-0.67	0.68
No. of lanes (4)	0.36	-0.22	0.94	-0.06	-1.02	0.89
No. of lanes (≥ 5)	-0.04	-0.64	0.56	0.01	-0.87	0.90
No. of in-neighbours (2)	0.05	-0.16	0.27	0.03	-0.31	0.36
No. of in-neighbours (≥ 3)	0.05	-0.38	0.48	-0.07	-0.70	0.56
No. of out-neighbours (2)	0.04	-0.19	0.26	0.03	-0.31	0.37
No. of out-neighbours (≥ 3)	0.29	-0.15	0.73	-0.04	-0.71	0.63
Cluster (2)	0.06	-0.23	0.35	0.04	-0.35	0.44
Cluster (3)	**-0.43**	-0.81	-0.06	**-1.24**	-2.25	-0.25
Cluster (4)	**-0.59**	-0.91	-0.28	0.05	-0.41	0.49
AADT (2)	0.10	-0.32	0.51	0.11	-0.52	0.74
AADT (3)	-0.15	-0.58	0.29	-0.26	-0.94	0.42
AADT (4)	0.28	-0.31	0.88	**1.22**	0.32	2.13
AADT (5)	0.48	-0.10	1.06	**1.34**	0.36	2.34
IAZ	**1.58**	1.35	1.79	**2.43**	1.67	3.18
Covariate|IAZ	γ	Lo	Up	γ	Lo	Up
Main road|IAZ	-	-	-	-0.27	-1.01	0.46
Parking slots|IAZ	-	-	-	-0.22	-0.68	0.25
Traffic light|IAZ	-	-	-	0.07	-0.47	0.61
Bus stops|IAZ	-	-	-	-0.30	-0.83	0.23
Bus lane|IAZ	-	-	-	-0.04	-0.66	0.58
No. of lanes (2)|IAZ	-	-	-	-0.33	-1.01	0.34
No. of lanes (3)|IAZ	-	-	-	0.42	-0.43	1.28
No. of lanes (4)|IAZ	-	-	-	1.05	-0.12	2.24
No. of lanes (≥ 5)|IAZ	-	-	-	-0.07	-1.21	1.08
No. of in-neighbours (2)|IAZ	-	-	-	0.09	-0.33	0.51
No. of in-neighbours (≥ 3)|IAZ	-	-	-	0.19	-0.64	1.01
No. of out-neighbours (2)|IAZ	-	-	-	0.01	-0.42	0.44
No. of out-neighbours (≥ 3)|IAZ	-	-	-	0.70	-0.16	1.57
Cluster (2)|IAZ	-	-	-	0.00	-0.54	0.55
Cluster (3)|IAZ	-	-	-	0.68	-0.40	1.79
Cluster (4)|IAZ	-	-	-	**-1.19**	-1.79	-0.59
AADT (2)|IAZ	-	-	-	0.03	-0.76	0.82
AADT (3)|IAZ	-	-	-	0.04	-0.81	0.89
AADT (4)|IAZ	-	-	-	**-1.76**	-2.91	-0.61
AADT (5)|IAZ	-	-	-	**-1.55**	-2.76	-0.36
Parameter	Est.	Lo	Up	Est.	Lo	Up
ψ	**1.71**	1.06	2.65	**1.56**	1.07	2.33
$z_{\text{Intercept}}$	**-8.69**	-15.87	-4.68	**-8.61**	-15.63	-4.63
z_{Slope}	**8.32**	4.30	15.49	**8.27**	4.28	15.30
α	**0.11**	0.01	0.32	**0.17**	0.01	0.47

Table 2.5: Summary of the results obtained with Models 1 and 2. Coefficient estimates (β and γ) in bold represent covariates significant with 90% credibility, whereas Lo and Up denote the lower and upper bounds (respectively) of the 90% credible intervals for such estimates

Covariate	**Coldspots**		**Average**		**Hotspots**	
	IAZ	**MAZ**	**IAZ**	**MAZ**	**IAZ**	**MAZ**
Main road (1)	0.33	0.00	0.53	0.47	0.94	0.94
Parking slots (1)	0.63	0.65	0.65	0.77	0.23	0.34
Traffic light (1)	0.25	0.00	0.52	0.07	0.63	0.29
Bus stops (1)	0.08	0.00	0.14	0.20	0.13	0.15
Bus lane (1)	0.33	0.00	0.52	0.50	0.82	0.98
No. of lanes (1)	0.71	1.00	0.56	0.59	0.17	0.07
No. of lanes (2)	0.25	0.00	0.21	0.21	0.09	0.11
No. of lanes (3)	0.04	0.00	0.13	0.12	0.12	0.23
No. of lanes (4)	0.00	0.00	0.07	0.04	0.40	0.21
No. of lanes ($\geq$ 5)	0.00	0.00	0.04	0.04	0.21	0.38
No. of in-neighbours (1)	0.28	0.26	0.26	0.24	0.46	0.34
No. of in-neighbours (2)	0.61	0.74	0.63	0.63	0.36	0.37
No. of in-neighbours ($\geq$ 3)	0.08	0.00	0.11	0.14	0.18	0.30
No. of out-neighbours (1)	0.16	0.39	0.27	0.21	0.42	0.33
No. of out-neighbours (2)	0.84	0.61	0.63	0.67	0.37	0.34
No. of out-neighbours ($\geq$ 3)	0.00	0.00	0.10	0.12	0.21	0.33
Cluster (1)	0.12	0.00	0.13	0.31	0.04	0.00
Cluster (2)	0.20	0.39	0.40	0.51	0.22	0.62
Cluster (3)	0.09	0.00	0.20	0.02	0.49	0.00
Cluster (4)	0.59	0.61	0.28	0.16	0.25	0.37
AADT (1)	0.71	1.00	0.60	0.67	0.11	0.06
AADT (2)	0.04	0.00	0.13	0.13	0.12	0.11
AADT (3)	0.25	0.00	0.12	0.08	0.09	0.13
AADT (4)	0.00	0.00	0.07	0.08	0.13	0.40
AADT (5)	0.00	0.00	0.07	0.04	0.55	0.30
Total road length (m)	506.87	144.95	10969.22	17626.74	2682.96	1632.44
No. of road segments	25	3	538	256	120	33
No. of accidents	7	1	1989	1066	1928	747

Table 2.6: Relative frequencies of the covariates in the road segments that form the coldspots, average zones and hotspots detected. Each frequency is obtained by averaging the values of the covariates for all the road segments in each set, but weighting them according to their corresponding lengths. For the binary variables only the frequencies of presence at the road segment (value of 1) are shown

Chapter 3

Identification of differential risk hotspots along a linear network

In this Chapter a procedure that allows detecting microzones of a road network where a specific type of accident is overrepresented is fully described. Then, the procedure is employed for the analysis of a geocoded traffic accident dataset including information on collision and vehicle types involved in each accident. Investigating the existence of microzones that are especially dangerous for a collision or vehicle type is a topic of interest from a prevention perspective due to the differential risk of severity that they usually trigger for the people involved in the accident, and also for better understanding how certain spatial road configurations may be correlated with certain collision or vehicle types.

3.1 Introduction

It is well known that collision and vehicle type are two capital variables that condition the severity of a traffic accident. To cite a couple of studies in this regard, Chang and Wang (2006) found vehicle type to be the most crucial factor that determines the severity of a traffic accident (which was determined to be superior for pedestrians, motorcyclists and cyclists), whereas Golob et al. (1987) verified that fixed-object and crossing were the most severe types of collisions. Thus, having the ability to detect the small sections of a road structure (which are called microzones in the remainder of the Chapter) that are particularly prone to present a specific type of accident is of great interest in order to implement preventive measures, specially for the type of accidents whose severity is expected to be higher. In the next paragraphs, a revision of previous studies that focused on the occurrence and causality of traffic accidents (considering the effect of collision or vehicle type) is carried out.

Regarding collision types, Dell'Acqua et al. (2013) defined several safety performance functions to establish the risk that associates to several collision types (head-on, rear-end, single-vehicle run-off-road) under a set of possible scenarios mainly involving surface conditions, light presence and the geometric structure of the street. Hosseinpour et al. (2014) specifically analyzed head-on collisions with a wide range of count-data models which showed that horizontal curvature, terrain type and heavy-vehicle traffic correlated with a high risk of observing this type of traffic accident. Finally, Wang et al. (2017) recently employed multivariate Poisson-lognormal models to analyze the factors contributing to traffic accidents with different severity, ranging from property-damage-only to fatal. In this work, the type of collision was mainly classified in terms of its directionality: same-direction, intersecting-direction, opposite-direction and single-vehicle. The results of the models dependent on these four types of collisions unveiled a differential effect of several road characteristics, as wide lanes associating with more opposite-direction accidents but less of single-vehicle type, or shoulder widths of more than 8 feet presenting a negative association with single-vehicle collisions, among other examples.

On the other hand, many studies have focused on the investigation of a specific type of vehicle, with an special emphasis being put on the estimation of the probability of injury and the severity associated with it. For instance, some studies have investigated these effects in accidents involving bicycles (Kim et al., 2007; Walker, 2007), motorcycles (Shankar and Mannering, 1996; Savolainen and Mannering, 2007) or heavy vehicles (Chang and Mannering, 1999; Anderson and Hernandez, 2017).

Motivated by these previous studies, the goal was designing a new methodology to detect microzones of a road network presenting a differential risk for a specific accident typology. The Chapter is structured as follows. First, the network structure that was used for the analysis is accurately described and located in its real context. Secondly, the traffic accident dataset that was employed is depicted, including a summary of the curation process and a detailed subsection on the information available regarding each traffic accident (by type of collision and vehicle). Then, a methodology is proposed to find microzones of the network that show a differential risk for a collision or vehicle type, including the posterior application of a Monte Carlo technique that serves to determine the statistical significance of each of these microzones. This procedure is fully displayed through an exemplification that makes use of accidents involving motorcycles. Finally, the methodology is applied and discussed for multiple collision and vehicle types.

3.2 Data

3.2.1 Network structure

A linear network composed of 1664 vertices and 2513 segments, which represented a total length of 191.14 km of road structure was analyzed. This network broadly included the city center of Valencia (Ciutat Vella District) and its five surrounding districts (l'Eixample, Extramurs, Campanar, la Saïdia and el Pla del Real), which are all shown, subdivided according to their boroughs, in Figure 3.1a (OpenStreetMap contributors, 2017; Graul, 2016). This area of Valencia contains some of the most travelled avenues of the city and represents, as a whole, an homogeneous and highly connected road structure.

Network complexity was reduced without altering its basic geometrical shape with the application of a simplification algorithm that merges segments sharing a vertex of second degree (through which only two segments are connected). Moreover, network preprocessing included the slight modification of highly complex intersections, the transformation of roundabouts into simple polygons of no more than six sides, and the removal of pedestrian streets. The complete process was performed with the help of the R package SpNetPrep (Briz-Redón, 2019).

Finally, the network was endowed with a direction according to traffic flow at this part of Valencia (also performed with SpNetPrep). Some of the segments of the network were defined as bidirectional, representing two-way roads where no median strip separates the two flows of vehicles. If a median strip is located in a road, two (parallel) segments are defined for the network. A representation of the final network that was employed in this study is available in Figure 3.1b, including the direction of traffic flow. The few differences that can be appreciated between the administrative structure portrayed in Figure 3.1a and the linear network in Figure 3.1b, were only executed to give the border of the whole structure a more solid representation.

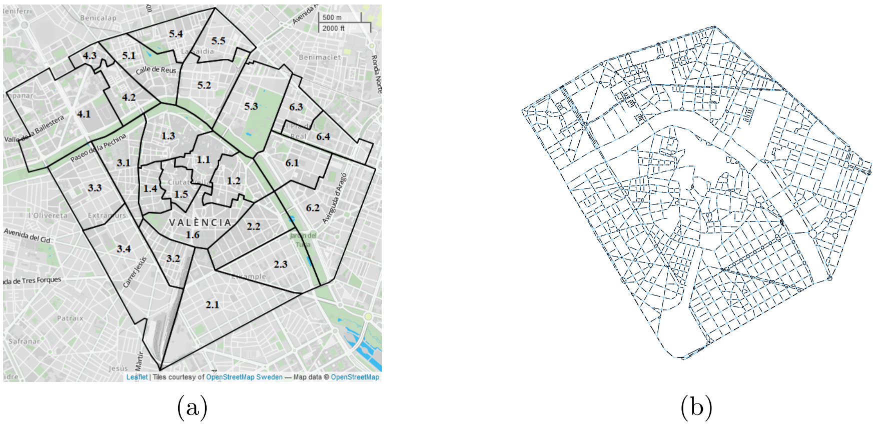


Figure 3.1: Boroughs of the area of the city of Valencia that was studied (a) and complete network structure used for the analysis, with arrows indicating the traffic flow (b). In (a), the nomenclature of each borough is overlayed on the map, with the first number indicating the district the borough belongs to, from 1 to 6: Ciutat Vella, l'Eixample, Extramurs, Campanar, la Saïdia and el Pla del Real

3.2.2 Accident dataset

A total of 11006 traffic accidents registered by the Police Department of the city of Valencia (Spain) during the years 2014 to 2017, which took place in the roads belonging to the area of the city described in the previous section were analyzed. Each of these accidents was accurately geocoded into the network from the address information recorded by the Police officers minutes after the accident had occurred. Manual curation of the data broadly included the following steps: selection of the accidents that took place in streets located within the districts of the area of analysis, geocoding (longitude-latitude coordinates) of the accidents from the street addresses reported by the Police via the Google Maps API and the R package ggmap (Kahle and Wickham, 2013a), revision of the obtained coordinates by applying reverse geocoding with the same package, and final inspection of the projection of these coordinates into the linear network.

As it was already mentioned in the introduction, each traffic accident in the dataset had information attached that made the point pattern become a marked point pattern. The marks regarding the collision type and the vehicles involved in each accident were chosen to be further analyzed and are described in the following two sections.

Types of collisions

Accidents are classified by Police officers according to the way they took place, leading to a quite complex variable with dozens of different categories. Considering the frequencies of these categories and examining the coincidence between some of them, only six collision types were established for the dataset: *Crossing*, *Fixed-object*, *Rear-end*, *Run-off-road*, *Run-over* and *Side*. The following lines include a brief description of each of these types.

- *Crossing*: The head of a vehicle collides to one of the sides (lateral) of another (moving) vehicle. Frontal collisions (between the front ends of two vehicles) were also assigned to this group as they were not abundant enough to establish a specific category.
- *Fixed-object*: A vehicle collides to a parked vehicle or to any fixed element in the street.
- *Rear-end*: The head of a vehicle collides to the end of another (moving) vehicle.
- *Run-off-road*: A vehicle loses control and leaves its traffic way, possibly invading the opposite direction of traffic flow or the sidewalk.
- *Run-over*: A vehicle hits a person and possibly drives over him/her. Sometimes it is also employed when a vehicle hits another one of clearly lower dimension and/or weight.
- *Side*: Lateral collision between two or more vehicles.

It must be remarked that some traffic accidents can be quite complex and involve several types of collisions and situations, making the treatment of this type of data quite challenging. In fact, a part of the accidents that took place in the road network considered during the period 2014-2017 remained without collision type being assigned due to the lack of information available in this regard, or because the literal concept annotated was completely different to the six predominant types that have been just defined (avoiding a proper recodification of the collision type). Table 3.1 shows the relative frequencies of the different collision types in the dataset (they represent 86.35% of the total because of the accidents uninformed or categorized with an infrequent type of collision) and the Moran's I values (Moran, 1950a,b) shown by each of the patterns, which confirm the significant spatial autocorrelation of every type of collision (at the 0.05 level, even though the *Run-off-road* type is at the limit).

Types of vehicles

Information regarding the types of vehicles implicated in each accident was also available. As expected, cars were the vehicles that appeared more frequently in the dataset by far, followed by motorcycles (mopeds were also recoded to this type). The involvement of cars in most of the traffic accidents led to discard its specific study, putting the focus on the rest of vehicle types.

Particularly, the other types of vehicles present in the dataset were bicycles, buses (private and public), lorries and vans. Accidents implicating public buses were separated from those involving private ones because the former are subject to specific routes, and also because of their singular importance as part of the public transportation system of the city.

It is needed to say that this mark of the point pattern was used in the form of a set of binary variables referring to each of the types of vehicles registered in the dataset. Therefore, each accident allowed the presence of several vehicle types, a situation that did not hold for the collision type. As a consequence, the addition of the relative frequencies in Table 3.1 exceeds 100.

The possibility that some microzones of the road network are particularly dangerous for some of these vehicle types was deeply analyzed, which seems reasonable according to the high values of the Moran's I in Table 3.1 that indicate the presence of spatial aggregation, specially for public buses.

3.3 Methodology

3.3.1 Software

The R programming language (3.4.1 version, R Development Core Team, Vienna, Austria) (R Core Team, 2018) was used to obtain all the results presented in this work. The R packages ggmap (Kahle and Wickham, 2013a), spatstat (Baddeley et al., 2015), spded (Bivand and Piras, 2015) and SpNetPrep (Briz-Redón, 2019) were specifically required for some parts of the analysis.

3.3.2 Estimating a relative probability for each type of accident

Kernel density estimation (KDE) is typically used in spatial statistics to estimate the intensity of a point pattern over a geographical space, as shown in Chapter

Mark	**Category**	***n***	**%**	***I*** **(*p*-value)**
Collision type	Crossing	3082	28.00	0.04 (0.00)
	Fixed-object	1518	13.79	0.06 (0.00)
	Rear-end	2691	24.45	0.16 (0.00)
	Run-off-road	198	1.80	0.03 (0.05)
	Run-over	701	6.37	0.13 (0.00)
	Side	1314	11.94	0.08 (0.00)
Vehicle type	Bicycle	635	5.77	0.09 (0.00)
	Car	9524	86.53	0.11 (0.00)
	Lorry	420	3.82	0.03 (0.05)
	Motorcycle	2811	25.54	0.12 (0.00)
	Private Bus	212	1.93	0.11 (0.00)
	Public Bus	747	6.79	0.21 (0.00)
	Van	916	8.32	0.05 (0.00)

Table 3.1: Sample sizes, relative frequencies and Moran's I values (with associated p-values testing the null hypothesis of no spatial autocorrelation) of all the categories available for the two marks considered

2 (Section 2.3.7). The equal-split continuous kernel density estimator for linear networks (McSwiggan et al., 2017) is now recalled:

$$\lambda_\sigma(x) = \sum_{z \in A(x,\sigma)} k(d_L(x,z))a^C(\pi) \tag{3.1}$$

where all the parameters and variables act as in Equation 2.3. For the development of the current methodology, the Gaussian kernel (represented by $k(\cdot)$ in Equation 3.1) was also used for simplicity, due to the known fact that this choice has little effect on the results (Silverman, 2018). Bandwidth parameters (σ) ranging from 50 m to 150 m were tested in order to check the effect of this election.

KDE can be used to estimate the spatially-varying relative probability of occurrence for a type of event or the relative risk between several event types, both two applications of special interest in case-control and related studies (Kelsall and Diggle, 1998; Diggle et al., 2005; Serra et al., 2013). Hence, if $\{y_i\}_{i=1}^n$ represents the binary outcomes (for example, case or control) of a collection of events observed at points $\{x_i\}_{i=1}^n$, KDE allows the derivation of a risk surface (Kelsall et al., 1995; Kelsall and Diggle, 1995) over space that can be interpreted as a conditional probability of observing a case ($Y_i = 1$) at a location X_i, with X_i and Y_i representing the random location of a spatial event and its outcome, respectively. In addition, in a more general setting, the case-control situation can be extended and one can set $y_i = 1$ if one mark of the point pattern takes certain value at point x_i, and 0 otherwise. This approach is taken, which enables defining a risk surface regarding the involvement of a specific collision or vehicle type in the neighbourhood of a point x_i.

Therefore, according to the formulas derived by Copas (1983) and Kelsall and Diggle (1998) for planar point patterns, an estimate (that depends on the bandwidth σ) of the risk presented by a type of event at a location x is:

$$p_\sigma(x) = \sum_{i=1}^{n} K_\sigma(x - x_i) y_i \Bigg/ \sum_{i=1}^{n} K_\sigma(x - x_i)$$

where n is the number of events observed and $K_\sigma(u) = h^{-2}K(\sigma^{-1}u)$ is a kernel function with $K(u) = (2\pi)^{-1}\exp(-\frac{1}{2}||u||^2)$ (Gaussian), being $|| \cdot ||$ the euclidean norm. This formula is then adapted to the case of the network KDE, which leads to the estimation of a relative probability of risk for any typology of accident at a location x of the road network:

$$p_\sigma(x) = \sum_{i=1}^{n} \lambda_\sigma(x_i) y_i \Bigg/ \sum_{i=1}^{n} \lambda_\sigma(x_i) \tag{3.2}$$

where $\lambda_\sigma(x)$ follows Equation 3.1. Hence, $p_\sigma(x)$ approximates the relative probability for the typology of traffic accident being represented by $y_i = 1$ to be observed at location x, which relies on the information provided by all the traffic accidents occurred within a linear radius (following the network structure) of σ meters from x.

3.3.3 Detecting differential risk hotspots

The main objective of the study was to design a methodology capable of identifying microzones of a road network where certain accident typology is overrepresented. From now on, these microzones will be referred to as differential risk hotspots (see Figure 3.2 for a full graphical description). Many previous studies have already dealt with the accurate detection of hotspots at the road segment level (Xie and Yan, 2013; Nie et al., 2015; Harirforoush and Bellalite, 2016), but focusing on a type of accident and assessing risk in relation to other types has been less investigated by far. Therefore, the following paragraphs include a description of the procedure proposed for differential risk hotspot detection, which relies on the estimation (across space) of a relative probability of occurrence for each event type, according to Equation 3.2. An implementation of this methodology is available in the R package DRHotNet, which is fully described in Chapter 8.

The first step of the procedure consists in estimating (with Equation 3.2) the relative probability of a specific accident typology along the complete spatial network. Such estimates need to be computed at a partition of the whole road network in order to obtain a estimation of the risk that the type of accident presents across space. In particular, the middle points of the 2513 segments of the road network from Valencia could have been chosen for this step, but in order to gain accuracy, specially near and around road intersections, the linear network was subdivided into shorter segments that are called *lixels* in literature (Xie and Yan, 2008) (Figure 3.2b). A value of 50 m was chosen such that the length of the segments of the new network (from now referred to as the *lixellized* network) did not exceed this threshold. The resulting lixellized network presented a total of 5099 segments and 4250 vertices, in which all traffic accidents available of any type were projected (Figure 3.2c).

Once the probabilities of risk were computed for a typology of accident at the middle points of the lixellized network (Figure 3.2d), several approaches were tested for attempting the definition of representative hazardous hotspots of road segments for each accident's typology. Firstly, following one of the most usual methodologies, local Moran's I (LISA) statistic values (Anselin, 1995) were computed for each of the lixels. Then, lixels showing a significant local association at the 0.05 level were selected and grouped according to their contiguity, leading to potential differential risk hotspots. However, probably due to the only moderate spatial autocorrelation showed by some of the types of accidents along the network and also to the small sample size at many microzones, this approach did not provide satisfactory results.

Therefore, an alternative method was finally established, which consisted in selecting the road segments with a superior relative probability estimate in comparison with the mean value for the whole network. More specifically, road segments associated with a relative probability exceeding the mean value for the network in more than k times the standard deviation presented by all the probabilities estimated were preselected. From this set of pieces of the network of length up to 50 m, only those where n or more accidents had occurred within a linear radius of σ meters (the bandwidth value chosen for estimating the probabilities) were finally selected for the construction of the hotspots (Figure 3.2e). Thus, the inclusion of the n parameter allows discarding some estimated probabilities that are not based on a large enough dataset of traffic accidents, which could be artificially large and consequently meaningless.

A sensitivity analysis was carried out on k and n with several values of the bandwidth parameter in order to find sensible choices of these two parameters that allow the obtention of a reasonable set of differential risk hotspots. The prediction accuracy index (PAI) developed by Chainey et al. (2008), which has already been used in some studies involving traffic accidents (Thakali et al., 2015) or crimes (Van Patten et al., 2009), was slightly modified to perform this analysis. Hence, a type-specific PAI (PAI_{type}) was defined as follows:

$$\text{PAI}_{\text{type}} = \frac{n_{type}/N_{type}}{m/M}$$

where n_{type} is the number of traffic accidents recorded in the hotspots with the type of interest, N_{type} the total number of traffic accidents in the study with the type of interest, m is the length (in meters) of all the hotspots detected, and M is the total length of the network structure being considered (191.14 km).

Thus, the PAI_{type} index computes the ratio between the proportion of accidents that took place in the set of differential risk hotspots obtained and the proportion of network length that is spanned by these hotspots. A higher value of the PAI_{type} index indicates that the hotspots found have proportionally condensed more traffic accidents for a given length of road, which represents a better performance in this context.

3.3.4 Assessing hotspot significance

The use of a KDE technique facilitates the estimation of the intensity of a point pattern over a space and the detection of microzones that show high/low values of intensity in comparison with their surroundings. However, the mere use of KDE does not provide a statistical significance value that helps to discriminate which hotspots show a more notorious differential behaviour for a type of collision or vehicle. In this regard, the methodology introduced by Bíl et al. (2013) is partially imitated to find which of the predefined differential risk hotspots deserves to be declared as a microzone of high dangerousness for a specific typology of accident, and not a consequence of the small sample size that the pattern presents at some microzones of the network, even though a right choice of the n parameter reduces the chances of this undesired possibility. Hence, the procedure described in the previous section serves to obtain a set of differential risk hotspots, whereas the process described in this one enables to assess a statistical significance value to each of these hotspots, which could then be ranked according to their importance, but also be rejected if they do not show enough statistical evidence of presenting a differential risk for the type of traffic accident being studied.

The methodology that is defined to check the statistical significance of each differential risk hotspot is now described. First, the point pattern available is left fixed in space whereas the marks (collision and vehicle type) of the events are changed randomly a total of 750 times. Following a Monte Carlo approach in order to test the null hypothesis of random mark assignment (meaning random type of collision or vehicles implicated), for every single simulation, and for every differential risk hotspot being analyzed, the probabilities of risk are computed at the middle points of the lixels forming each hotspot (considering each simulated marked point pattern). Then, a probability of risk for the type of accident is assigned to each hotspot at every simulation by simply averaging the probabilities estimated for every lixel (the average is weighted by lixel's length to increase the contribution of longer lixels, even though this decision barely alters the results). This process generates an empirical distribution for the probability of risk at every differential risk hotspot, allowing the construction of a statistical significance measure (a p-value) for each of them, completing the procedure that allows the detection differential risk hotspots in a directed road network. The assessment of a statistical significance to each differential risk hotspot can eventually lead to discard some of them or, at least, reduce their importance in favor of the most significant ones (Figure 3.2f).

On the choice of the value 750 as the number of simulations performed to assess statistical significance, the study of Robey and Barcikowski (1992) was followed. Assuming a type I error $\alpha = 0.05$, a type II error $\beta = 0.2$ (which implies a power for the test of 0.8) and a "liberal criterion" according to the definition of robustness for a statistical test introduced by Bradley (1978), a value around 750 would be recommended. A larger choice may be better depending on the level of uncertainty one wishes to undertake, but here the definition of 750 iterations was considered fair enough according to the results that were obtained and the computation time that was saved avoiding the use of a more stringent criterion.

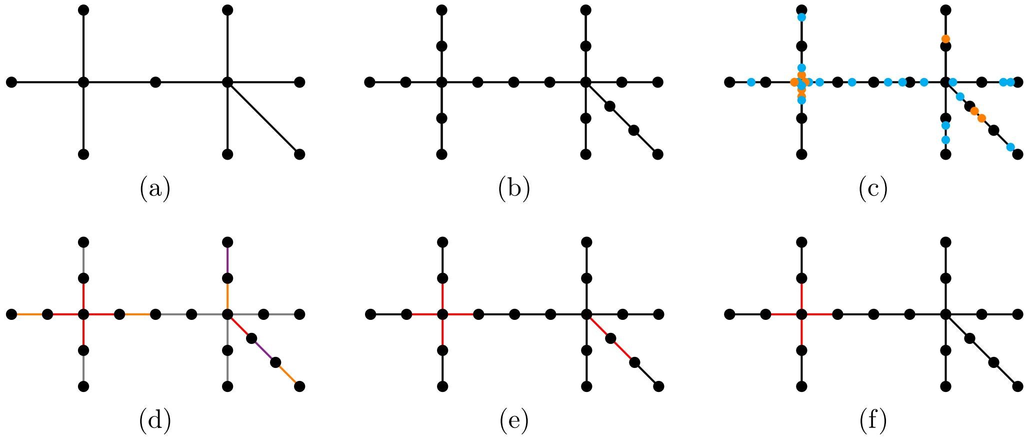


Figure 3.2: Graphical description of the procedure implemented in order to detect differential risk hostpots. Starting from a road network of interest (a), this is segmented into shorter elements called lixels (b). The traffic accidents available are then projected into the lixellized network in (b), some of which can be of one specific type (orange points), whereas the rest are of any other type (blue points), as shown in (c). Next, KDE is employed to produce an estimation of the probability of risk for the accident type being studied (symbolized by the orange points) at each of the middle points of the lixellized network (d). As an illustration, in (d) the lixels are coloured according to a sensible guess of these probabilities in view of the point pattern in (c), ranging from average or below average (gray) to very high (purple). The lixels satisfying the conditions imposed by the k and n parameters are selected (in red) and grouped, becoming differential risk hotspots (e). Finally, a Monte Carlo technique is applied to the hostspots in (e), yielding a value of significance for each of them that can occasionally lead to their rejection (f)

3.3.5 Ranking hotspots according to further criteria

The detection of microzones of a road network presenting a singular risk for a collision or vehicle type should be followed by the implementation of a set of countermeasures that attempt to improve their safety. As stated, the procedure introduced in this Chapter allows the ranking of the differential risk hotspots found according to their empirically determined statistical significance.

However, it would be sensible to bring into the equation several external factors, not considered explicitly by the procedure described in this Chapter, that are capital to establish a systematic decision-making process that optimizes the social cost-benefit associated to the application of traffic safety countermeasures. Hence, the absolute number of traffic accidents found in a hotspot, the proportion of severe accidents observed or some collision/vehicle types particularly overrepresented (which are highly correlated with severity outcomes) are only a few of the factors that may be incorporated for this matter. In this regard, some methodologies for hotspots ranking that allow the combination of several indicators, like the one developed by Coll et al. (2013), would be useful.

3.4 Results and discussion

This section starts with a subsection that illustrates the methodology proposed in this Chapter by showing all the steps that were followed to detect microzones of the linear network where traffic accidents involving motorcycles were overrepresented. Furthermore, this section includes a discussion on the choice of the k and n parameters and their appropriateness depending on the necessities of the researchers and professionals involved in the design of preventive measures. The second subsection contains a summary and analysis of the differential risk hotspots that are obtained for all the typologies of traffic accidents considered in the study.

3.4.1 Example of application: motorcycle accidents

Motorcycles contributed to 2811 of the 11006 accidents available in the dataset, representing 25.54% of the total of accidents. Following Equation 3.2, the relative probability of observing an accident involving a motorcycle was estimated along the complete lixellized network with a bandwidth value of $\sigma = 100$ m (Figure 3.3a displays a graphical representation of these estimations, with the lixels presenting a value higher than 0.3 being coloured accordingly). Several bandwidth parameters were tested, ranging from 50 to 150 m, and a value of 100 m was found a suitable choice according to the results obtained. The use of bandwidths lower than 100 m, specially those that are close to 50 m, leads to the frequent generation of many short hotspots that fail to connect to other microzones of the network (reducing the comprehensibility of the results), whereas the election of a larger value around 150 m seems to excessively extend the effect of the point pattern, likely producing fictitious microzones of differential risk.

As an illustration, the application of the hotspot detection methodology with $k = 1$ and several values of n is shown in Figures 3.3b-3.3d. As expected, for a fixed k, a larger value of n is more restrictive and less microzones of the network are pointed out.

The effect of the election of k and n is further investigated with the use of a set of values in the range $[0, 2]$ for k and varying n from 10 to 50. Figure 3.4 shows the values of PAI_{type} (Figure 3.4a) and the proportion of accidents involving a motorcycle (Figure 3.4b) for the sets of hotspots determined for the different values of k and n. Other vehicles (and types of collisions) were investigated with such graphical descriptions and provided similar results, so the arguments stated in the next paragraphs that are based on the case of motorcycle accidents stay true in general.

As it has already been explained, PAI_{type} index measures the ratio between the proportion of accidents of interest found within a set of hotspots (in this example, presenting a differential risk for motorcycles) and the proportion of length that these hotspots represent in the whole network structure. Therefore, the highest value of PAI_{type} would indicate the k and n values that optimize the procedure, in the sense of providing a minimal road length structure (the set of hotspots) given the number of motorcycle accidents that occurred on it. Figure 3.4a suggests that a choice around

$k = 1.5$ and $n = 45$ would be the optimal for PAI_{type} in this example. Moreover, in this study the interest lies in the relative probability of accident for a specific type of collision or vehicle, more than on the microzones of the network that contain the majority of them (PAI_{type} focuses on this). In this regard, Figure 3.4b shows how the proportion of accidents involving a motorcycle increases as k and n do, although a very restrictive selection of k and n would imply a too reduced number of hotspots.

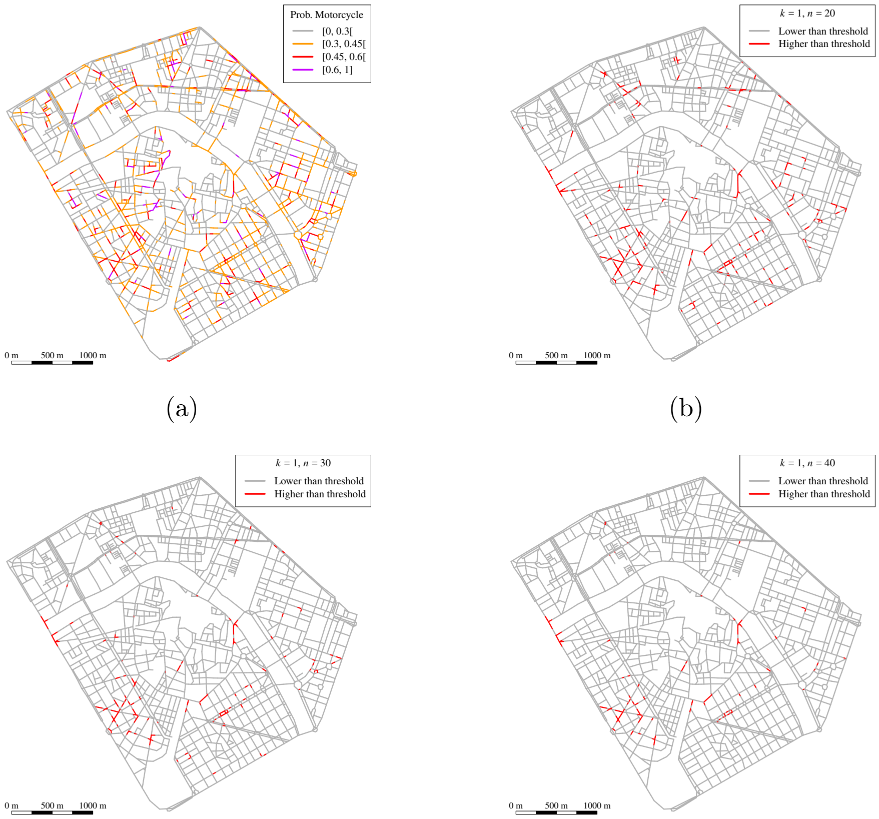
In conclusion, the variation of the k and n parameters can lead to very different sets of differential risk hotspots. A value for k in the interval [1,2] and $n \geq 30$ seem appropriate to obtain a reasonable number of microzones along the network presenting differential risk for one specific type of collision or vehicle, but the final election should remain to the decision of the researchers, and specially the professionals that could be in charge of establishing preventive measures according to the results obtained (police officers and traffic experts). To illustrate this idea, Table 3.2 describes the results that were obtained from the differential hotspot detection procedure for several values of k and n (still with accidents involving motorcycles). With the exception of the most restrictive parameter values, the microzone that is spanned by the set of hotspots extends for some (or many) kilometers, suggesting the use of a more exigent combination if the wish is to focus only in the most relevant parts of the network for the accident typology of interest.

Keeping this in mind, the procedure here described for motorcycle accidents was performed for most of the possible outcomes of the two marks attached to the point pattern (collision and vehicle type) with $k = 1$, $n = 40$ and $\sigma = 100$, including the posterior application of the Monte Carlo technique to obtain information regarding significance. These results are shown in the next section.

3.4.2 General application of the methodology

The procedure is employed with the different types of vehicles and collisions informed at the available dataset, providing the significant differential risk hotspots that are shown in Figure 3.5 and Figure 3.6 (at the 0.05 level). As in the example included in the previous section, the values $k = 1$ and $n = 40$ were chosen in order to obtain a initial set of differential risk hotspots. Later, the statistical significance of these hotspots was assessed by the Monte Carlo technique, yielding the final collection that appears in Figures 3.5 and 3.6 (only those with a p-value lower than 0.05 are selected). Specifically, in the case of motorcycle accidents, Figure 3.6a includes the hotspots from the group shown in Figure 3.3c that were yielded significant by the Monte Carlo procedure, even though the ones rejected could also be reconsidered once the former are inspected and treated by the authorities.

Occasionally, the differential risk procedure yields differential risk hotspots formed by only one lixel, which are specially notorious in the less frequent types of accidents that naturally involve a minimal number of hotspots. The examination of some of this one-lixel hotspots unveils some problematic situation that needs to be explained. The use of the KDE technique with a bandwidth of 100 m implies that the probability of risk that is estimated at the middle point of each lixel of the network is based



(a) (b)

(c) (d)

Figure 3.3: Estimated probabilities of observing an accident that involves a motorcycle for each lixel of the linear network (a) and differential risk hotspots detected after the application of the methodology with $k = 1$ and (b) $n = 20$, (c) $n = 30$ and (d) $n = 40$

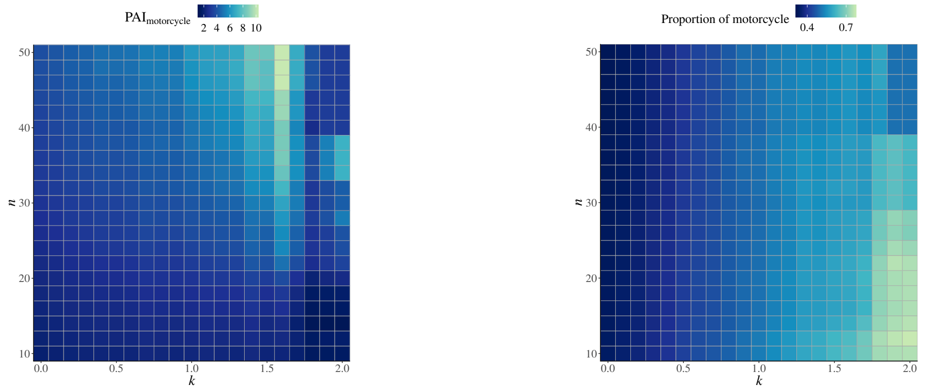
on a slightly wider microzone that the lixel itself (which can not exceed a length of 50 m by construction). This situation can eventually lead to the declaration of a singular lixel of the network as a differential risk hotspot, although only a little number of accidents of the type being studied have occurred within that lixel. Even though such a hotspot could be simply a false positive produced by the methodology being proposed, the problem addresses automatically sometimes if one thinks about the middle point of the lixel as the center of a linear radius of 100 m along the network (following traffic flow) where that type of accident is overrepresented. For instance, this situation takes place with the one-lixel differential hotspot detected for bicycle accidents in Borough 2.2 and with the one found in the border of the Boroughs 3.1 and 3.3 related to run-over collisions (see Figure 3.1a for locating these boroughs). In the former, no accidents involving bicycles were observed from 2014 to 2017, but 6 out of the 42 accidents that occurred in a radius of 100 m from the middle point of this lixel in the same period of time implicated a cyclist. Indeed, the estimated relative probability for an accident in this lixel to involve a bicycle was 0.17 (very close to the simple proportion), which is far superior to the proportion of bicycle accidents in the whole dataset (0.06). Similarly, no run-overs were recorded in the

k	n	**Hotspots**	$N_{motorcycle}$	N	$p_{motorcycle}$	**Length (m)**	$\mathbf{PAI}_{motorcycle}$
0.5	10	287	1247	2936	0.42	40692.88	2.08
1	10	181	583	1128	0.52	18217.86	2.18
1.5	10	98	200	320	0.62	6559.52	2.07
2	10	41	55	74	0.74	2320.5	1.61
2.5	10	14	14	15	0.93	588.62	1.62
0.5	20	235	1027	2461	0.42	26287.38	2.66
1	20	124	443	881	0.50	10277.83	2.93
1.5	20	48	129	220	0.59	2754.69	3.18
2	20	14	20	27	0.74	555.54	2.45
2.5	20	4	5	5	1.00	125.80	2.70
0.5	30	183	862	2107	0.41	17272.68	3.39
1	30	85	347	708	0.49	5853.93	4.03
1.5	30	25	100	178	0.56	1308.92	5.19
2	30	5	10	16	0.62	159.03	4.28
2.5	30	0	-	-	-	-	-
0.5	40	128	704	1739	0.40	12032.39	3.98
1	40	50	277	568	0.49	3824.72	4.92
1.5	40	16	84	155	0.54	880.23	6.49
2	40	2	4	8	0.50	86.04	3.16
2.5	40	0	-	-	-	-	-
0.5	50	101	583	1434	0.41	8262.58	4.80
1	50	35	240	489	0.49	2688.96	6.07
1.5	50	13	84	154	0.55	701.51	8.14
2	50	2	4	8	0.50	86.04	3.16
2.5	50	0	-	-	-	-	-

Table 3.2: Description of the differential risk hotspots (denoted simply as hotspots in this Table) that are detected when applying the procedure with different values of k and n for $\sigma = 100$ m, where $N_{motorcycle}$ is the number of accidents involving motorcycles in the complete set of hotspots for two given values of the parameters, N is the total number of traffic accidents recorded in the same space, and $p_{motorcycle}$ is the proportion $N_{motorcycle}/N$. It can be appreciated that for some combinations of these two parameters no hotspots are determined

aforementioned differential hotspot regarding this kind of collision, but in a radius of 100 m from its middle point a total of 11 out of 56 traffic accidents belonged to this type (the estimation of the relative probability of a run-over collision at this lixel was 0.21, much larger than the proportion of 0.06 for the complete network).

Therefore, one should consider each of the differential hotspots pointed out by the procedure as the core of a slightly extended microzone of the network that manifests a differential risk for one specific type of accident. This way of thinking about the hotspots is usually unnecessary, as the lixels that form the hotspot normally encompass most of the accidents recorded in that microzone of the network, but becomes a requirement to give sense to the shortest hotspots that are composed by only one lixel. In the same vein, with regard to the interpretation of a particular differential risk hotspot, it is worth to remark that the type of collision should be



(a) (b)

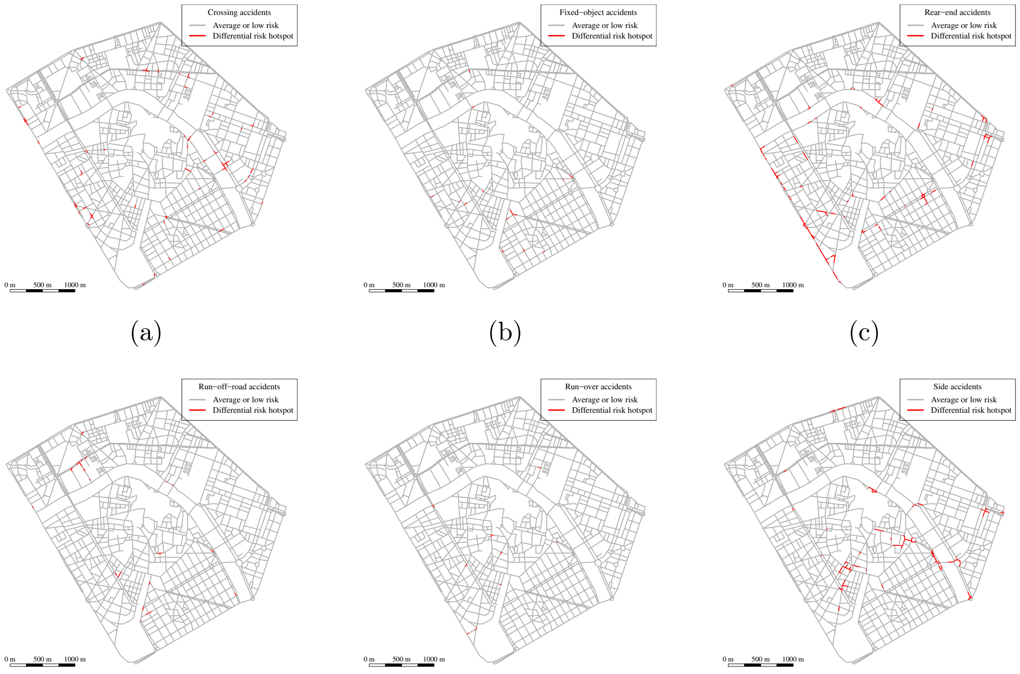
Figure 3.4: $\text{PAI}_{\text{motorcycle}}$ results (a) and proportion of traffic accidents involving a motorcycle (b) within the set of hotspots obtained with different values of k and n and a bandwidth value of $\sigma = 100$ m

also accounted. For instance, for run-off-road accidents, a detected microzone may be the consequence of a triggering condition located dozens of meters apart, in the same or in a connected road, which is where countermeasures should really focus.

As a summary, Table 3.3 describes the differential risk hotspots that were found for the different collision and vehicle types. This table includes the number of traffic accidents of each type that were recorded within the corresponding set of hotspots and the proportion of accidents they represented along this microzone of the road network. In order to minimize the possible presence of differential risk hotspots that point out a very specific microzone of the network but are really representing a wider part of it (issue treated in the previous paragraphs), the number of accidents of each type occurred within an extension of the hotspots is also indicated (a 75 m linear radius from the extreme points that include the complete hotspot).

Regarding the PAI_{type} values that were obtained, Table 3.4 indicates that these ranged from 5 to 16, approximately, being specially high for run-off-road collisions, bicycles, lorries and buses (private and public). Hence, the detection of differential risk hotspots was in average further optimized for the vehicles than for the collision types, suggesting an overall higher level of dispersion along the road network for the latter. Furthermore, Table 3.4 indicates that accidents involving public buses presented the higher level of concentration at differential risk hotspots, followed by private buses, side collisions and rear-ends. Table 3.4 also points out that the procedure presented in this Chapter is specifically focused on the finding of microzones along the road network that are particularly dangerous for a collision or vehicle type, and not on the microzones where the collision or vehicle type is more concentrated (with higher accident counts), which usually associates with arterial and busy roads regardless of the collision or vehicle type.

Finally, an exploratory analysis of the coincidence of differential risk hotspots between collision and vehicle types was performed. Figure 3.7 shows the percentage of road structure spanned by differential risk hotspots that is shared by each collision and vehicle type. This percentage is obtained in relation to the collision or vehicle type whose set of differential risk hotspots is longer. The highest percentages were found between side collisions and buses (both private and public), with a maximum percentage of coincidence close to 20% for side collision and public bus. Other remarkable but minor associations included crossing with motorcycles and rear-ends with motorcycles and private buses.



(a) (b) (c)

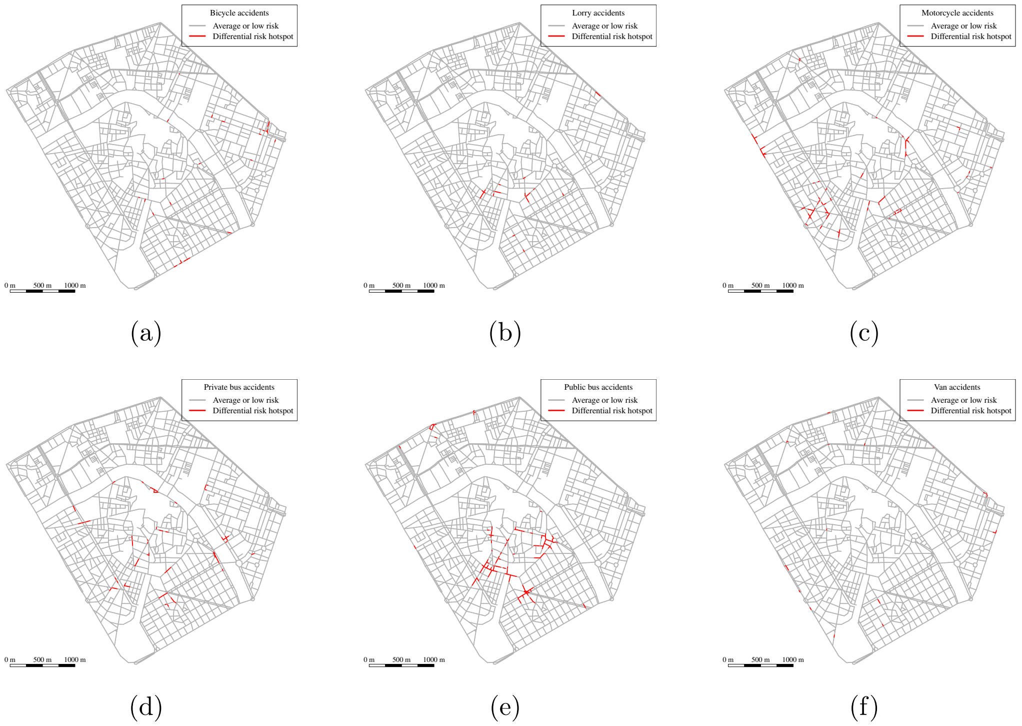
(d) (e) (f)

Figure 3.5: Differential risk hotspots that are statistically significant (at the 0.05 level) for each collision type after the application of the detection procedure with $k = 1$, $n = 40$ and $\sigma = 100$

3.5 Conclusions

This Chapter has fully described a methodology that provides microzones of a road network that present a significant high risk for a type of collision or vehicle. These microzones are called differential risk hotspots. The successful implementation of such methodology obviously requires that the road network itself is taken as the space where the traffic accidents are located (areal spaces are not accurate enough for this matter). Furthermore, it is recommended to fraction the road network into shorter pieces to increase precision.

The use of a KDE-based technique which is typically used in case-control studies has rendered possible to estimate a probability of risk along the whole network for every type of collision and vehicle properly informed in the available dataset. From these estimates, a procedure that accounts for the disparity between each estimate



(a) (b) (c)

(d) (e) (f)

Figure 3.6: Differential risk hotspots that are statistically significant (at the 0.05 level) for each vehicle type after the application of the detection procedure with $k = 1$, $n = 40$ and $\sigma = 100$

and the mean proportion of the type of accident in the complete road network (that also excludes the parts of the network lacking a representative sample) has been designed. The election of the two parameters that define the procedure is vital to obtain a sensible number of differential risk hotspots.

Indeed, the detection of microzones of a road network structure that present a differential risk for some type of collision or vehicle type should be a previous step to the definition of preventive measures that diminish its dangerousness. Depending on the resources available, the number of microzones of the road network that one would like to consider for further analysis and treatment could be very different. For this reason, the parameters of the algorithm could be chosen in a less restrictive way in order to find a larger number of microzones at which preventive measures could be implemented. Oppositely, if only a very reduced number of measures can be afforded, the values of the parameters should be increased and this naturally would lead to only a few microzones to be further studied.

Furthermore, the differential risk hotspot detection procedure is complemented with the posterior inclusion of a Monte Carlo technique that scores the importance of each hotspot. In this final step, the hotspots presenting a statistical significance under a fixed threshold could be discarded or left aside from a first package of preventive measures.

Type	Hotspots	Length	$N_{type}^{hotspots}$	$N_{total}^{hotspots}$	$N_{type}^{extended}$	$N_{total}^{extended}$	$p_{type}^{hotspots}$	$p_{type}^{extended}$	$p_{type}^{network}$
Crossing	48	2.97	251	416	660	1400	0.60	0.47	0.28
Fixed-object	13	0.63	27	54	106	373	0.50	0.28	0.14
Rear-end	43	5.15	409	866	823	2193	0.47	0.38	0.24
Run-off-road	19	1.13	15	108	39	537	0.14	0.07	0.02
Run-over	10	0.39	11	27	51	339	0.41	0.15	0.06
Side	30	4.42	237	867	410	2112	0.27	0.19	0.12
Bicycle	22	1.06	45	184	87	683	0.24	0.13	0.06
Lorry	11	0.98	34	185	63	543	0.18	0.12	0.04
Motorcycle	36	3.30	263	537	568	1485	0.49	0.38	0.26
Private Bus	27	3.00	45	383	81	1429	0.12	0.06	0.02
Public Bus	26	5.39	231	1047	339	1978	0.22	0.17	0.07
Van	15	0.77	31	100	73	423	0.31	0.17	0.08

Table 3.3: Summary statistics for all the sets of differential risk hotspots (denoted simply as hotspots in this Table) that were determined as statistically significant by the Monte Carlo procedure (at the 0.05 level) for the collision and vehicle types considered, after the application of the detection methodology with $k = 1$ and $n = 40$. The table includes the number of differential risk hotspots that were found in each case, its total length (in kilometers), the number of traffic accidents of that type ($N_{type}^{hotspots}$) and in total ($N_{total}^{hotspots}$) that were recorded within the hotspots, the number of traffic accidents of that type ($N_{type}^{extended}$) and in total ($N_{total}^{extended}$) that were recorded within the hotspots or within an extension of 75 meters around them, the proportion $N_{type}^{hotspots}/N_{total}^{hotspots}$ ($p_{type}^{hotspots}$), the proportion $N_{type}^{extended}/N_{total}^{extended}$ ($p_{type}^{extended}$) and the global proportion of the type of accident along the complete road network $p_{type}^{network}$ (also available in Table 3.1)

Type	% Traffic accidents	% Road length	$\mathbf{PAI}_{type}$
Crossing	8.14	1.55	5.25
Fixed-object	1.78	0.33	5.37
Rear-end	15.20	2.69	5.64
Run-off-road	7.58	0.59	12.83
Run-over	1.57	0.20	7.71
Side	18.04	2.31	7.79
Bicycle	7.09	0.56	12.72
Lorry	8.10	0.51	15.79
Motorcycle	9.36	1.73	5.41
Private Bus	21.23	1.57	13.49
Public Bus	30.92	2.82	10.96
Van	3.38	0.40	8.43

Table 3.4: Description of the sets of differential risk hotspots found for each collision and vehicle type considering the percentage of traffic accidents (for the corresponding type) and road length (in relation to the whole network) that the set represents. The quotient of these two percentages is equivalent to the PAI_{type} value associated

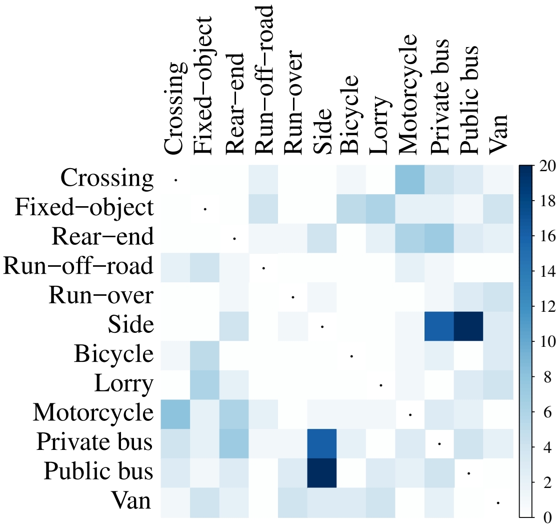


Figure 3.7: Percentage of coincident road length presented between the set of differential risk hotspots obtained for the collision and vehicle types considered after the application of the procedure with $k = 1$, $n = 40$ and $\sigma = 100$. This percentage is computed with respect to the collision or vehicle type whose set of hotspots spans over a larger number of meters (through the quotient between coincident road length and total length of the longer set of hotspots)

Chapter 4

Adjusting the Knox test for the analysis of the near-repeat phenomenon

In this Chapter, the primary goal is to highlight the necessity of adjusting the classic version of the Knox test for assessing the presence of space-time interaction to get a more accurate representation of the magnitude and duration of the near-repeat phenomenon in the context of criminology. The classical version of the Knox test and a modification of it based on the work of Schmertmann (2015), which allows adjusting the results of the original test under the more plausible assumption of varying risk in space and time, are described in detail. The adjustment of the Knox test is exemplified through the analysis of a dataset of burglaries occurred in Valencia (Spain) in the period 2016-2017. In particular, several covariates over an areal covering of the whole city are defined to measure their time-varying effect on the risk of burglary to make the adjustment of the original Knox test possible. Furthermore, regarding the dataset used, a methodological subsection focused on the time uncertainty issue that it presents has been also added, even though this is only a minor goal of the study.

4.1 Introduction

In the field of criminology, repeat victimization refers to the repetition of some criminal event against the same victim. The analysis of the repeat victimization phenomenon has been of great interest since decades ago (Loftin, 1986; Sparks, 1981). A more general framework regarding repeat victimization is the near-repeat phenomenon, which does not necessarily deal with those events that produce against the same victim (exact-repeat), but with those that occur at a close distance, in space and time, from the spatio-temporal location of a previous crime (near-repeat). This study is focused on the near-repeat rather than on the exact-repeat phenomenon.

The near-repeat phenomenon has been massively analyzed in the last decades for a wide range of crimes. For instance, armed street robberies (Haberman and Ratcliffe, 2012), shootings (Ratcliffe and Rengert, 2008), or gun assaults (Wells et al., 2012) have been researched under the near-repeat framework. Besides, some researchers have even analyzed the coexistence of several near-repeat dynamics for different crime types, as Youstin et al. (2011) did with shootings, robberies and auto thefts that had occurred in Jacksonville (Florida, USA).

Seeing the studies cited above, it follows that the near-repeat phenomenon can involve two general types of victims: people and private properties. In particular, the near-repeat phenomenon is usually observed for residential burglaries, as Johnson et al. (2007) showed through the analysis of burglary events occurred in a total of ten areas from five different countries. As the dataset used for supporting the goals of the present study is a collection of burglaries, hereinafter the focus is put on the mechanisms that drive the near-repeat phenomenon from the perspective of this type of crime.

Two main theories explain the existence of the near-repeat phenomenon with burglaries: “flag” and “boost” (Johnson, 2008). These two theories coexist and overlap in reality, becoming hard many times to determine which of the two is really driving the near-repeat phenomenon in a place. In fact, the presence of a near-repeat dynamic for the burglaries occurred in a city has been analyzed as a phenomenon dependent on the weekday (Glasner and Leitner, 2016), socioeconomic characteristics (Zhang et al., 2015) and housing homogeneity (Townsley et al., 2003). Now, the two theories that attempt to describe the near-repeat of burglaries are briefly described in the following paragraphs.

It is well known that the characteristics of a place alter the intrinsic risk associated to every crime type, including burglaries. Thus, a zone of a city can provide particular facilities to burglars that can lead to higher burglary rates and, as a consequence of this, to the repeat victimization of a property or a set of neighbouring properties. This mechanism of near-repeat generation is known as “flag” in criminology, which basically refers to the overall attractiveness (permanent or transitory) of a place for offenders.

The other well-established theory that explains the existence of the near-repeat situation is called “boost”. It has been observed that burglars tend to choose properties where they have broken in before, or very close ones to these because they can make use of their previous experience and then increase their chances of success (Polvi et al., 1991). Indeed, the “boost” hypothesis can sometimes be validated when a burglar (or group of burglars) are proved to be responsible for the commission of several offences that were close in space and time (Bernasco, 2008), or strongly suspected if a singular “modus operandi” is being used repeatedly in a particular zone (Ratcliffe and McCullagh, 1998).

Despite the “boost” and “flag” are two well-established theories in criminology that partially explain the near-repeat phenomenon, there is a lack of statistical methodology fully oriented to distinguish the “proportion” of near-repeats coming from each effect in a real context. For instance, Short et al. (2009) implemented a mathematical model to evaluate if the distribution of the time intervals between exact-repeat burglaries occurred in Long Beach (California, USA) only depended on risk heterogeneity and not on the previous occurrence of a similar offence. With this aim, these authors estimated a theoretical distribution for the time intervals between exact-repeats, which allowed them to reject the null hypothesis of complete temporal randomness and conclude that another mechanism (of a "boost" nature) had to be responsible for a proportion of the near-repeats.

4.1.1 Factors influencing crime risk: the case of burglaries

This section's goal is to show how burglary risk heterogeneity can arise as a consequence of a wide range of factors. Many of the studies pursuing the identification of covariates that may influence burglary risk follow a spatial approach making use of risk terrain modelling (RTM) techniques. These usually involve the segmentation of the whole area under investigation into spatial units that make possible to measure the variation across space of the factors considered for the analysis and hence model crime risk according to them. The following paragraph includes a description of several papers that used RTM to investigate some factors that are suspected of affecting burglary risk.

Moreto et al. (2014) distinguished between instigator (originator) burglaries, near-repeats, and isolated (in space and time) burglaries. These authors showed that instigators and near-repeats were more strongly associated with areas of high burglary risk than the rest of the burglaries, according to a set of covariates that considered the effect of land uses, pawn shops, drug market locations, among others. Caplan et al. (2015) investigated the influence on burglary rates of several features of the landscape including, for instance, bars, parks, or schools. They found that the proximity to foreclosures was the factor that increased the risk more severely. Nobles et al. (2016) analyzed the effect of several covariates on different types of burglaries within the context of the near-repat phenomenon. Thus, an index of concentrated disadvantage, residential instability, racial heterogeneity, and street connectivity showed a positive association with near-repeat burglaries. Andresen and Hodgkinson (2018) obtained multiple significant covariates, some of whom showed a positive association with burglaries (number of dwellings, unemployment rate, pawnbrokers, etc.), whereas others behaved oppositely (percentage of recent immigrants, median income).

Most of the factors analyzed in the studies cited above were closely related to crime attractors/generators and to demographic/socioeconomic characteristics of the population living in the place under investigation, leaving essential characteristics of the road network aside. The following lines fill this gap. Beavon et al. (1994) suggested that road network complexity acts as a natural barrier against crimes. Under the assumption that criminals tend to operate around the zones that are part of their regular activity space, complex (highly-dense and lowly-accessible) road structures should be less travelled, in general, and then provide fewer opportunities to criminals. In agreement with Beavon et al. (1994), Ye et al. (2018) recently found a negative correlation between road density and property crime. Besides road density, the connectivity of street segments within a city is another factor that has deserved investigation. Indeed, Davies and Johnson (2015) found a positive association between the number of burglaries observed at streets and the global connectivity level (betweenness) of such segments across the city of Birmingham (UK). However, Frith et al. (2017) found that this association was only holding for non-local connectivity in several towns of Buckinghamshire (UK) (non-local connectivity implies the computation of a street's connectivity only from those street segments that are at some predefined distance far). This finding was explained on the basis that streets highly-connected at a local level allow the presence of more usual pedestrians that behave as natural guardians of the place. Staying on this subject, dead-end streets have also been specifically studied within the context of searching covariates that explain burglary risk heterogeneity. The work from Johnson and Bowers (2010) suggests that the presence of dead-end streets reduce crime risk, although Hillier (2004) considered this statement valid only for linear and easy-to-police dead-end streets. Murakami et al. (2004) used four parameters coming from graph theory to characterize road network forms and investigate their relationship with convenience store robberies in Tokyo (Japan). Although these authors could not achieve solid conclusions regarding the effects of these parameters (given the results obtained for their data), the relationship between network's structure and the level of crime opportunity may deserve further investigation. The first two parameters that Murakami et al. (2004) used, $\beta = e/v$ and $\gamma = e/3(v\text{-}2)$ (where e and v represent the number of road segments and intersections, respectively), somehow measure network's topology, representing partial (β) and overall connectivity (γ). On the other hand, $\eta = L/e$ and $\pi = L/T$ (where L and T are the total road length and network's diameter) are metric indicators that account for the diameter and total length of the road structure being considered in relation to the number of edges and vertices. Thus, η and π can be interpreted as the average road length and the intricacy of the network, respectively. Kansky and Danscoine (1989) and Liu and Zhao (2015) provide a deeper description of all these parameters.

Specific characteristics of buildings and dwellings can lead burglars to search for better cost-benefit scenarios. Chang (2011) detected that buildings presenting an average condition showed a higher association with burglaries than the worst and best-looking buildings in a South Korean city. Malczewski and Poetz (2005) used a geographically weighted regression to assess that the value of dwellings produced significant local variations in the risk of burglary in London (Ontario, Canada). Mennis et al. (2011) obtained that a high percentage of vacant housing was producing more property crimes in Philadelphia (USA).

Finally, several demographic characteristics have been associated with a higher burglary risk at both the individual and spatially-aggregated level. Cohen and Cantor (1981) found that residents living in the city center, young residents, people with incomes in the highest and lowest categories, and people whose homes are unoccupied have all greater than average odds to suffer a residential burglary. On the other hand, at a spatially-aggregated level, some authors have also identified that racial heterogeneity may be a factor that leads to higher crime rates (Hipp, 2011).

Given this reasonably extensive literature review, it seems clear that there is a wide variety of factors that may modify the risk of burglary across space and time. The presence of such spatio-temporal risk heterogeneity can increase the number of near-repeats through a 'flag' type mechanism, rather than through a 'boost'. However, researchers overlook this aspect when applying the most used statistical test existing for measuring the near-repeat phenomenon: the Knox test (Knox and Bartlett, 1964).

4.1.2 The statistical assessment of near-repeats

In this section, several statistical methodologies that allow measuring the magnitude and significance of the near-repeat phenomenon given a dataset of events located in space and time are described.

The well-known Knox test proposed by Knox and Bartlett (1964) is the most usual tool in quantitative criminology to evaluate the presence of the near-repeat phenomenon in relation to one criminal event of interest, as shown in many studies from the last few years (Chainey et al., 2018; de Melo et al., 2018; Glasner and Leitner, 2016; Glasner et al., 2018; Haberman and Ratcliffe, 2012; Hino and Amemiya, 2019; Hoppe and Gerell, 2019; Johnson, 2010; Piza and Carter, 2018; Powell et al., 2018; Sturup et al., 2018). The Knox test is a non-parametric tool that mainly relies on a statistic based on the count of pairs of events that are close in space and time, given a prespecified spatio-temporal threshold. Thus, the Knox ratio between the number of close events observed to those expected under the null hypothesis of no space-time interaction (of no near-repeat phenomenon) serves to assess the statistical significance of the results. Similarly, Mantel (1967) constructed a test through the sum of the products between the temporal and the spatial distances associated with each pair of events available in the dataset. The null hypothesis of Mantel's test states that spatial and temporal distances are independent. Jacquez (1996) suggested the replacement of geographical-based distances by neighbour relations. Hence, he designed a test for space-time interaction based on the count of case pairs that are k nearest neighbours in space and time (simultaneously). Gabriel and Diggle (2009) developed the space-time inhomogeneous K-function (STIK), which is a generalization of the classical spatial K-function (Ripley, 1977). The STIK function enables to test for space-time clustering and interaction by comparing the values provided by this function for the dataset under investigation with the ones that would be yielded by a Poisson inhomogeneous spatio-temporal process. Finally, the self-exciting model developed by Mohler et al. (2011) is another alternative to analyze the near-repeat phenomenon. This kind of model has two main parts: a background rate (for instance, following a Poisson distribution) and a triggering component. The triggering component is usually driven by a kernel function (Ogata, 1998) that allows controlling the elevation and decay of the risk of the event possibly affected by the near-repeat phenomenon to occur again. Mohler (2014) also implemented one addition to the model described by Mohler et al. (2011) that consisted in the inclusion of a mark representing the capacity of each event to lead to more incidents of the same type. More recently, Reinhart and Greenhouse (2018) have proposed a self-exciting model that includes spatially-varying covariates for the estimation of the background rate.

However, despite the availability of new and promising methodologies, the Knox test is still a valid and practical option to measure the near-repeat phenomenon, being nowadays the most preferred methodology among many of the researchers in the field. Indeed, the Near Repeat Calculator (NRC), a software developed by Ratcliffe (2009) widely used for the computation of the Knox test among quantitative criminologists, has been cited 64 times (according to Google Scholar) since 2011 (as of June 2019). More specifically, the NRC reached a maximum number of citations in 2018, 16 of which came from criminology papers that declared the use of the NRC to conduct part of the research (this analysis was performed with the aid of the Publish or Perish software (Harzing, 2007)). Motivated by the constant and current use of the Knox test in crime-related studies, the focus has been put on assessing and showing the necessity of making some adjustments to the standard version of this test to produce more accurate results.

As already highlighted, the Knox test is the most used technique for assessing the presence of the near-repeat phenomenon, but it presents some limitations. Reinhart and Greenhouse (2018) pointed out that one drawback of the Knox test in comparison to self-exciting models is its incapability of globally measuring the rate and form of the decay of the near-repeat phenomenon. Furthermore, Ornstein and Hammond (2017) pointed out the necessity of isolating the effect produced by crime contagion ("boost" effect) from that resulting from spatio-temporal crime risk variations. This is an issue already discussed in previous studies from the perspective of bias presence in space-time interaction tests as a consequence of overlooking exposure shifts and risk heterogeneity (Kulldorff and Hjalmars, 1999; Mantel, 1967). Specifically, Ornstein and Hammond (2017) constructed an agent-based computational model to simulate burglary contagion and assess the performance of the Knox test. They confirmed that the Knox test can reliably detect the presence of some "boost" effect, but they also observed some deficiencies. First, the relationship between the real magnitude of the "boost" (a parameter of the model constructed) and the one determined by the Knox ratio is not linear. Furthermore, they found that even in the absence of a "boost", changes in baseline risk can produce Knox ratios consistently greater than 1, erroneously suggesting the existence of a contagion process.

Indeed, the standard Knox test relies on the hypothesis of constant relative risk (CRR) in space and time, which constitutes a very hard assumption, especially as the temporal window of observation is enlarged. In this regard, Kulldorff and Hjalmars (1999) constructed a modification (KH test) of the Knox test that accounts for the bias that arises when spatio-temporal risk variations are not considered. They suggested the consideration of the exposure unit related to the event of interest (usually human population) and its incorporation to the Knox test to produce more reliable results. However, years later, Schmertmann (2015) found the persistence of some bias in the results produced by the KH test, which were then corrected with the provision of a new empirical test based on the Metropolis-Hastings algorithm (Hastings, 1970; Metropolis et al., 1953) that does not modify the spatial and temporal margins of the observed map, avoiding the bias that sometimes affects the KH test. In summary, overlooking spatio-temporal risk variations can bias the results yielded by the standard Knox test and lead to the overdetection of significant spatio-temporal intervals under which the near-repeat phenomenon is not actually taking place, or to the underdetection of intervals that are really affected by the phenomenon (the overdetection or underdetection will depend on the direction of the bias).

4.2 Data

4.2.1 Burglaries dataset

A dataset provided by the Spanish National Police containing information about 2647 geocoded burglaries occurred in the city of Valencia during the years 2016 and 2017 was analyzed. The availability of geographical coordinates for every recorded burglary allowed analyzing of this crime at the most accurate level of spatial resolution. As an illustration, Figure 4.1 shows kernel density estimation (KDE) values (at the road segment level), which are derived from the locations of the burglaries following the network-constrained version introduced by McSwiggan et al. (2017). On the other hand, the temporal location of a good part of these events presented a higher level of uncertainty. The nature of burglaries is different from other common crimes such as assaults, in the sense of not being possible to determine, sometimes, the exact time at which a burglary has happened. A high percentage of burglaries take place when the owners of the property being invaded are away from home following their daily routines, or even on a long period of absence due to holiday travelling. For this reason, it is usual that the only information available regarding the moment a burglary has occurred is a pair of reference dates (or hours within a date) that delimit the temporal location of the burglary. These two reference times are commonly referred to as “from date” and “to date” (or “from hour” and “to hour” if the date is known). They represent, respectively, the last moment at which the owners can assure that the property had not been burgled yet and the moment at which the owners (or the Police, or some neighbours, etc.) found that the burglary had been committed.

4.2.2 Road network structure

In order to guarantee an accurate analysis of the occurrence of burglaries in the city of Valencia, the road structure of this city was employed for the analysis. Thus, the coordinates available for each crime were located into a spatial structure in the form of a graph made of edges and vertices. The use of this structure facilitates the performance of a more realistic analysis than the one that an areal representation of the city would enable us, as the spatial distance between any two events can be computed according to the way people can travel across the city. In addition, the availability of a digitized version of the road network allows the definition of some specific covariates based on street characteristics that would be otherwise missing from the analysis.

Technically, the road structure that was used in the analysis was made up of 9151 vertices and 12944 edges. These edges represent road segments of the city, with an

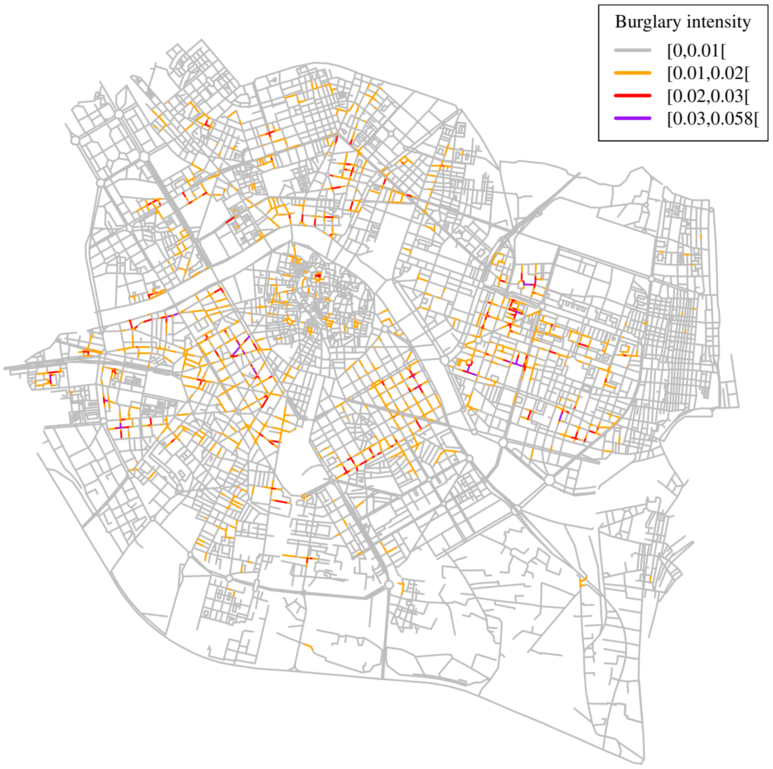


Figure 4.1: Graphical description of burglaries in the city of Valencia through network-constrained KDE values

average length of 65.13 meters. From the total number of vertices, 6129 (67%) had a degree higher or equal to three, meaning that at least three road segments of the network were incident to them (representing road intersections).

4.3 Methods

4.3.1 Software

The R programming language (3.5.0 version, R Development Core Team, Vienna, Austria) (R Core Team, 2018) was used to obtain all the results presented in this work. The R packages changepoint (Killick and Eckley, 2014), ggmap (Kahle and Wickham, 2013a), igraph (Csardi and Nepusz, 2006), rgeos (Bivand and Rundel, 2018a), spatstat (Baddeley et al., 2015) and SpNetPrep (Briz-Redón, 2019) were explicitly required for some parts of the analysis.

4.3.2 Time information uncertainty

The near-repeat phenomenon has been investigated for the burglary dataset introduced in the previous section on a daily basis, meaning that an uncertainty of some minutes (or hours) between the “from date” and the “to date” was unimportant if both dates were the same. However, for burglaries presenting a “from date” that was different from its corresponding “to date”, two strategies were tested to address this issue. First, the midpoint date between the “from date” and the “to date” was computed as the date halfway between the two available dates. The second strategy that was carried out was the employment of the concept of aoristic time, which has already been treated in several previous works in the field of criminology (Ratcliffe, 2002, 2004b). Indeed, the aoristic time has been suggested to be a more accurate choice, overall, than the midpoint or other approximations such as the “from date” or the “to date” themselves (Ashby and Bowers, 2013). Taking this approach into account, burglaries whose exact date of occurrence is not known are evenly distributed along the period determined by the “from date” and the “to date” recorded for them. For instance, a burglary occurred from date 2016-1-1 to date 2016-1-5 is weighted 0.2 along the five days that form this uncertain temporal interval.

Figure 4.2 shows the three time series that represent the evolution of the number of burglaries in Valencia for the years 2016 and 2017, considering the crimes whose dates are known precisely and the two estimations of the total number of burglaries that are derived from using the midpoint and aoristic approximations. The time series based on known dates lies below the other two because only 60.7% of the events recorded were known with the required precision (exact date). In addition, the gap between the curves is very notorious within the summer months which, in agreement with the longest vacation periods for Spanish citizens, undertake the highest level of temporal uncertainty.

A high similarity was found between the two time series derived from the use of the midpoint and aoristic dates (Figure 4.2). Furthermore, the spatial patterns of certain and uncertain burglaries did not show remarkable differences in terms of spatial clustering, which was checked through the nearest-neighbour index (Clark and Evans, 1954; Cressie, 1993) and visual exploration. Both facts led to discarding the aoristic approach for the rest of the study given the simplicity and inferior computational cost of the midpoint method and, more importantly, the fact that the management of time uncertainty is not the primary goal of the present investigation. Thus, the choice of the midpoint method is fair enough for achieving the research purposes of comparing the standard and adjusted version of the Knox test, despite the fact that the aoristic approach is probably the most accurate option available. Nevertheless, research on how the choice of a specific methodology for the imputation of uncertain dates can affect the results of the Knox test, which is currently missing in the literature to the best of my knowledge, may be of particular interest.

4.3.3 The Knox test

The seminal work from Knox and Bartlett (1964) established a statistical methodology that allows the identification of space-time interactions for any collection of events that are located in space and time. The so-called Knox test assesses the presence of such interactions by measuring the number of paired events that occur within a spatio-temporal interval and comparing this value with the one expected under the null hypothesis of no space-time interaction. More specifically, if σ and τ represent, respectively, a spatial and a temporal interval, the following statistic, $X_{\sigma,\tau}$,

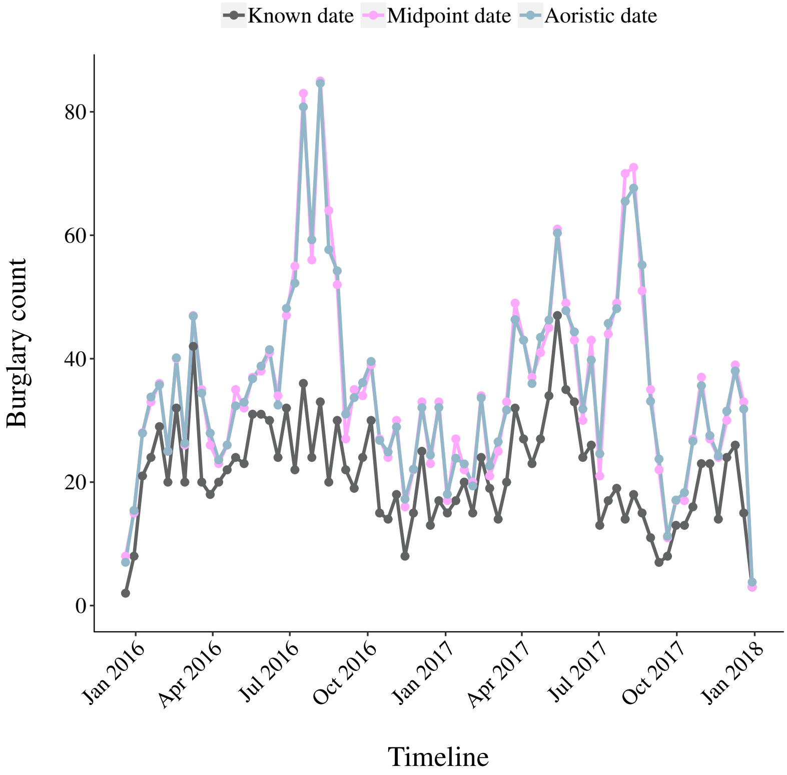


Figure 4.2: Time series representing the incidence of burglaries (grouped by time intervals of 10 days) during the period of two years considered. The series accounting for events whose date of occurrence was known is compared to the ones obtained by using the midpoint and aoristic strategies

represents the number of paired events that fall within the two specified intervals:

$$X_{\sigma,\tau} = \sum_{i=1}^{n} \sum_{j=i+1}^{n} S_{ij} T_{ij}$$

where n is the total number of events in the dataset, i and j are indexes for such events and S_{ij} (T_{ij}) is an indicator function that takes the value 1 if the spatial (temporal) distance between events i and j lies within the spatial (temporal) interval determined by σ (τ), and 0 otherwise. As the road network structure was available, the distance between two burglary events was measured in terms of the shortest-path distance, which is the minimal distance (in meters) required to travel from one event to the other while only passing through the edges and vertices that form such structure. The absence of a road network where the events of interest reside would oblige to use some alternatives as the Euclidean or Manhattan distances, but shortest-path provides a more exact measure. Indeed, Manhattan and especially Euclidean distances could increase the counts within a given spatio-temporal window significantly, as investigated by Groff and Taniguchi (2018).

The null hypothesis of the absence of space-time interaction is assessed with the estimation of an empirical probability distribution for $X_{\sigma,\tau}$ (for each specific combination of σ and τ considered). This distribution allows the construction of a p-value that evaluates the significance of this statistic and the estimation of the Knox ratio (observed/expected) that represents the departure of the observed value $X_{\sigma,\tau}$ from the one expected under the hypothesis of spatio-temporal randomness regarding the occurrence of burglaries.

To obtain the empirical distribution of the statistic $X_{\sigma,\tau}$, 5000 simulated datasets were created and their corresponding $X_{\sigma,\tau}$ values computed, following the classical Monte Carlo approach. For this task, the observed spatial locations for the burglaries were left fixed, whereas their time locations were randomly permutated across the complete dataset (Mantel, 1967). More precisely, if n is the number of burglaries in the dataset, this process consisted in the generation of 5000 random permutations of the natural numbers from 1 to n, reassign the observed dates to the spatial locations according to these permutations and compute the new $X_{\sigma,\tau}$ for this simulated combination of spatial and temporal locations. Then, p-values were estimated (for each choice of σ and τ) as $p = r/N$, where N represents the total number of simulations that were computed (5000) and r the number of simulated ratios that were greater than or equal to the observed statistic, $X_{\sigma,\tau}$.

4.3.4 Adjusting the Knox test

The Knox test was modified to account for the variations of burglary risk in space and time across the city of Valencia during the two years being considered. This is performed through the inclusion of several covariates that attempt to represent the variations in risk across space and time, which attempt to remove the effects that the wrong assumption of CRR may have on the results yielded by the Knox test. The approach presented in this Chapter is heavily based on the adjustment described in Schmertmann (2015).

Schmertmann (2015) proposed two main ways of adjusting the Knox test (via the Metropolis-Hastings algorithm) that were capable of addressing the bias issue that generally arises with other modifications of the Knox test: through exposure shifts and through covariates. The second option, based on covariate differential effects in space and time, was chosen to adjust the results of the Knox test associated to the dataset of burglaries analyzed. In the following paragraph this decision is justified.

If the context of an epidemiological study, the population size would be the exposure variable, and its use to adjust the Knox test would be entirely required, as it has been exemplified in previous studies (Schmertmann, 2015; Schmertmann et al., 2010). In this work, as burglaries are a crime on private properties where people live, the exposure should not be population but dwelling units instead. The 28682 housing units (buildings or independent houses) registered in Valencia for the year 2016, comprising a total of 382539 dwelling units (exposure), were geocoded for the analysis. Figure 4.3 shows the distribution of the housing units in Valencia in 2016, indicating the number of dwellings available at each unit. Independent houses and small buildings mainly located in the city center and in peripheral parts of the city of medium-low socioeconomic status. On the other hand, most of the large buildings are located far away from the city center, usually within the context of wealthier neighbourhoods. The fact that burglary data was only available for 2016-2017, and the negligible level of dwelling construction in Valencia during this period led to discard the modification of the Knox test proposed by Schmertmann (2015) solely based on exposure shifts and select the one accounting for covariate effects as the reference. Furthermore, it is worth noting that the inclusion of the covariates can be carried out contemplating two different scenarios. First, if the covariates available vary in space and time, then it could be determined the “global” effect that the covariate has (considering the whole period under investigation) on the existence of crime risk heterogeneity across space and time (as a consequence of the spatio-temporal variations in the covariates). This is exactly the approach depicted by Schmertmann (2015) regarding the adjustment of the Knox test through covariates. Second, covariates may have a fixed value over the period under analysis, but the relationship of each of them with crime outcomes could be variable over time, which is another mechanism of baseline crime risk modification. Given the nature of the covariates available for performing the case-study shown (most of which were unaltered during the period under investigation), the second option was selected. In the rest of the section the adjustment of the Knox test proposed by Schmertmann (2015) is specified and an alternative approach for modelling crime risk heterogeneity through covariate values is introduced.

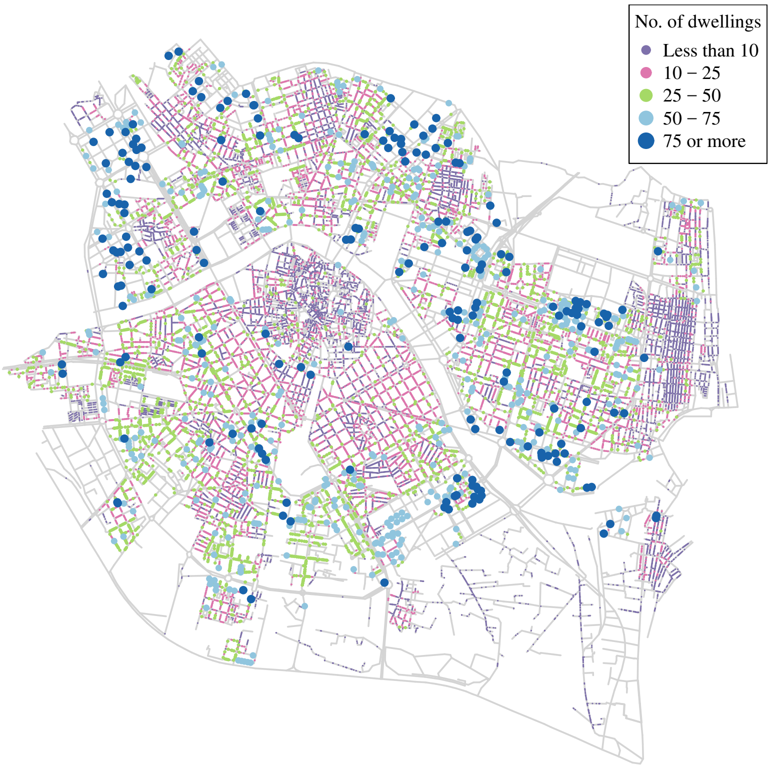


Figure 4.3: Point pattern over the road network of Valencia representing the location of houses and buildings in the city. The colour and size of each point indicates the number of dwellings (exposure unit) the house or building comprise

Some of the basic terminology introduced by Schmertmann (2015) is now presented.

The map of observed burglaries, M_0, is defined as the set of spatio-temporal locations, $(s,\ t)$, where a burglary occurred (at spatial location s, on date t). Then, given a permutation $\rho = (\rho(1),\ ...,\ \rho(n))$ of the first n natural numbers (where each $\rho(i)$ is a natural number from 1 to n, where repetitions are not allowed), a permutated map, M_ρ, is defined as the set of pairs, $(s,\ t_\rho)$, representing a burglary occurred at spatial location s on the permutated date given by t_ρ.

Following the previous notation, the probability of every possible permutated spatio-temporal map of burglaries in the dataset after a permutation of the dates can be related to exposure shifts and spatio-temporally varying covariates as follows (Schmertmann, 2015):

$$P\left(M_\rho\right) \propto \exp\left(\sum_{i=1}^{n} \log N(s_i, t_{\rho(i)})\right) \exp\left(\sum_{i=1}^{n} \gamma \cdot Z(s_i, t_{\rho(i)})\right) \tag{4.1}$$

where n is the number of burglaries available in the dataset, $(s{,}t)$ are the spatio-temporal coordinates of a burglary, ρ is a permutation of the first n natural numbers, $\cdot$ is the classical scalar product, $N(s{,}t)$ is the exposure value at $(s{,}t)$, $Z(s{,}t)$ is a vector of covariates evaluated at $(s{,}t)$, and γ' is a vector of coefficients representing the global effects on crime risk estimated for these covariates.

Equation 4.1 was then adapted to fit a context of no spatio-temporal variation in both the exposure and the covariates:

$$\log P(M_\rho) \propto \sum_{i=1}^{n} \alpha(t_{\rho(i)}) \cdot Z(s_i) \tag{4.2}$$

where n, s, t, ρ, and $\cdot$ represent the same than in Equation 4.1, $Z(s)$ is a vector of covariates evaluated at spatial location s, and $\alpha(t)$ is a vector of coefficients associated with the effect of these covariates on burglary risk that varies over time. Note that both Equation 4.1 and Equation 4.2 represent a proportionality relationship (a likelihood) and not an exact probability value.

Hence, Equation 4.2 focuses on the spatio-temporal variation of the effects that the covariates have on burglary risk. Putting Equation 4.2 into practice requires the partition of the whole space (Valencia) and time (2016-2017) to estimate burglary risk heterogeneity (this part is explained in Section 3.6). Hence, if t^* is a date within the time window, $\alpha(t^*)$ may be better understood as $\alpha(T)$, being T the period which contains date t^*. Section 3.6 clarifies how the coefficients that form $\alpha(t)$ are estimated.

4.3.5 Metropolis-Hastings algorithm

The basic Monte Carlo test presented in previous sections implicitly implies that every permutation of the dates observed (known exactly or available in the form of “from date” and “to date”) is equally likely for the locations where the burglaries occurred (which remain fixed in the construction of this test). This fact implies that each of the $n!$ possible permutations of the n temporal locations available in the dataset can be assigned with equal probability ($1/n!$) to the spatial locations that have been observed for the burglaries.

The Metropolis-Hastings algorithm (Chib and Greenberg, 1995; Hastings, 1970; Metropolis et al., 1953) modifies the standard Monte Carlo approach through the exploration of permutations that are more likely to be observed according to the road network and exposure characteristics.

Instead of producing a completely new permutation at each iteration of the process (as the basic Monte Carlo does), the Metropolis-Hastings algorithm generates a proposal permutation from the last one available, forming a chain of explored states that is usually referred to as a Monte Carlo Markov chain (MCMC). Starting from the map of burglaries observed, M_0, each proposal only permutes a subset of the n temporal locations of the current map in the chain, forming a new map (state) that is incorporated to the chain if and only if its relative probability in comparison with the previous one in the chain is large enough. Otherwise, the next state of the chain is defined with the same map available in the former iteration. These steps are specified with more precision in the following lines:

1. Take the identity permutation, $\rho_0 = (1,\ 2,\ \ldots,\ n)$, that corresponds to the observed spatio-temporal locations of the burglaries (all dates are fixed). Compute $X_{\sigma,\tau}$, and start the chain with ρ_0 and the likelihood of M_0 (following Equation 4.2).

2. The iterative process starts. For K iterations do:

- Generate a proposal distribution that permutes J (even number) of the n temporal locations of the current map in the chain. Specifically, if a_1, ..., a_J are the J indexes sampled, the permutation consists in the interchange of positions $a_1 \leftrightarrow a_{2,}$ $a_{J-1} \leftrightarrow a_J$.
- Compute the relative likelihood, R, between the new proposal and the last element of the chain (again, Equation 4.2 is invoked).
- Obtain a value, u, following a uniform distribution in the interval [0,1]. If $u \leq \min(R,1)$, the proposal is accepted and added to the chain. If $u > \min(R,1)$, the proposal is not accepted and the chain is updated with the same proposal and map than in the previous iteration.

3. Once the chain of states is formed, the first B are eliminated (a process called "burn-in"). Furthermore, the chain can be "thinned" by extracting a subset of the states that are far enough from each other, avoiding an excessive correlation between close states.

4. The final chain of states (after the steps followed in 3.) provides an adjusted empirical distribution of the $X_{\sigma,\tau}$ statistics that allows making inference on it as in the standard Monte Carlo procedure. The simulated Knox ratios that are extracted

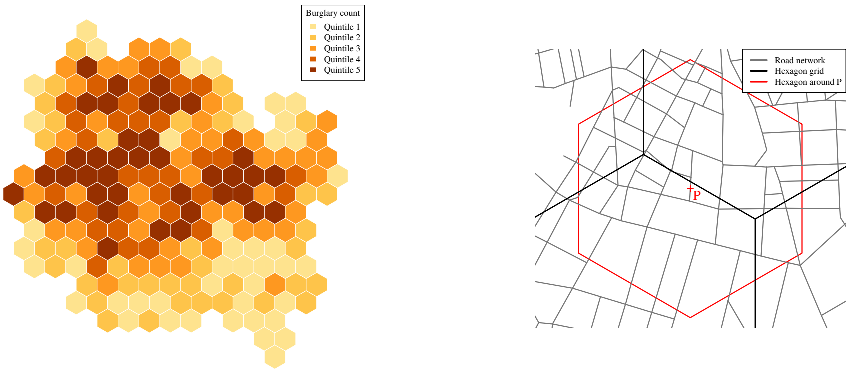
from the chain can be used to compute an empirical p-value as in the case of the Monte Carlo procedure.

Regarding the values of the parameters involved in the implementation of the algorithm, in this study the choices of $K = 27000$, $J = 50$, $B = 2000$ and a thinning achieved by choosing one of every five values of the chain were finally selected. Under these conditions, the effective sample size of each chain (for each combination of σ and τ) was $(27000-2000)/5=5000$.

4.3.6 Burglary risk estimation

With the goal of estimating burglary risk across space (Valencia) and time (2016-2017), the partition of both components was required. This section explains how this was performed. Besides, several other options are indicated.

The whole area of Valencia being analyzed was subdivided into 201 equal-sized hexagons of around 0.22 km^2 (288.68 m of side length). The choice of a grid was preferred over administrative divisions, such as boroughs or census tracks, for the avoidance of boundary effects (boundaries of administrative units are usually located along the main roads of the city) and the presence of highly unequal spatial units of analysis. Furthermore, the use of hexagonal grids is really frequent in spatial modelling and can even provide certain advantages over squared or rectangular grids at some situations (Birch et al., 2007). Figure 4.4a shows the composition of the hexagonal grid and the quintile within the distribution of burglary counts each hexagon belongs to.



(a)

(b)

Figure 4.4: Burglary counts at the hexagonal grid defined over the area of study (a) and example of construction of a hexagon around a location where a burglary took place (P) for the estimation of a burglary risk around it (b)

With regard to the choice of several periods along with the whole time window, in the analysis of the burglaries occurred in Valencia this was solved through changepoint detection, which is a methodology developed for time series analysis that pursues the identification of abrupt changes in a time series (Killick et al., 2012). Other sensible choices for the segmentation of the complete time window (not based on the own data) would be the definition of periodic intervals (months or group or months), or even more specific periods based on external information (following police advice). Any choice of the periods serves to adjust the standard version of the Knox test, as a varying effect (in space and time) will be provided to each covariate, allowing the adjustment of the test to produce.

Thus, the coefficients in $\alpha(t)$ (Equation 4.2) were estimated through a quasi Poisson generalized linear model (GLM), because of the overdispersion observed for the burglaries at the hexagon level (the negative binomial or the Poisson-lognormal models would be other two reasonable choices). If Y_i represents the burglary counts at hexagon i and $Z_j(i)$ a covariate evaluated at hexagon i, a GLM model is based on the linear relationship between a link-function (as the natural logarithm) and a set of covariates, which can be expressed as follows (Faraway, 2016):

$$\log\left(r_i\right) \ = \alpha_0 + E\left(i\right) + \alpha_1 Z_1\left(i\right) + \cdots + \alpha_k Z_k\left(i\right) \tag{4.3}$$

where r_i satisfies $\mathrm{E}\left[Y_i \mid Z_1\left(i\right), \ldots, Z_k\left(i\right)\right] = r_i$ (conditional expectation of Y_i), representing burglary risk at hexagon i, α_0 is the constant coefficient of the model, $E(i)$ is an offset term that represents the number of dwelling units available at hexagon i (level of exposure), $Z_j(i)$ is the value of the jth covariate at hexagon i and α_j is the coefficient that represents the effect of the covariate Z_j. The quasi Poisson modelling approach captures the overdispersion present in the counts by allowing the variance of the Y_i's to be increased by a factor $\varphi > 1$.

Therefore, if Equation 4.3 is assumed as a mechanism that explains burglary risk for the area of study, Equation 4.2 can then be rewritten in the following way:

$$\log P(M_\rho) \propto \sum_{i=1}^{n} \log(r(s_i)) \tag{4.4}$$

where $r(\mathrm{s}_i)$ is the value of burglary risk predicted for a hexagon centered at s_i and having the same size than the hexagons that form the grid (Figure 4.4b), according to Equation 4.3. The definition of a GLM model in the form of Equation 4.3 for each of the periods considered within the years 2016 and 2017 allowed determining of a vector α for each of them. Thus, the temporal variations in the effects produced by the covariates to burglary intensity are accounted for. Now, a detailed description of the set of covariates that was selected for the analysis is included in the following section.

4.3.7 Covariates selection and construction

Adjusting the standard Knox test to account for the intrinsic spatio-temporal interaction in burglary risk makes it necessary to perform a careful selection of the set of covariates that will be involved in the adjustment. In this case, the search focused on three aspects that can cause risk heterogeneity: the geometry of the road network, the characteristics of the units of exposure, which are the dwellings available in the city, and some demographic characteristics of the residents of the city.

As a summary, Table 4.1 contains a brief description of all the covariates considered with the aim of adjusting the standard Knox test. The values for these covariates were obtained for each of the hexagons forming the grid defined over the city of Valencia. Hence, the use of the term “road network” in Table 4.1 refers to the specific zone resulting from the intersection of the whole road network and each hexagon of the grid. On the other hand, Table 4.2 displays a summary of the event of interest (burglaries), the exposure unit (dwellings) and the set of covariates over the hexagonal grid that was used for the analysis.

This set of covariates attempts to gather several of the factors (described in Section 4.1.1) that have been investigated concerning their influence on burglary risk. Data unavailability usually obliges researchers to adapt or approximate some of the covariates used by other authors when performing their studies. Indeed, while several geometric-related and population-related covariates were defined in a very similar way as it has been done in prior studies, the exposure-related (dwelling-related) covariates required the introduction of some modifications, which are now depicted.

The level of attractiveness that a building or dwelling may have for burglars could have been measured through the visual inspection of streets of Valencia. However, as this process is highly complex and costly, the cadastral value and the antiquity of the dwellings were chosen as a proxy of attractiveness for the characterization of the dwellings in Valencia. Cadastral values are assigned by official appraisers of the administrative office in charge of this task. It is worth noting that the cadastral value of a dwelling can occasionally be quite different from the current market price of it, but the correlation between the two values is usually high. Similarly, the percentage of dwellings that are occupied could only be estimated from the two databases that contain, respectively, the registry of people inscribed in the city and the registry of built dwellings.

Regarding the geometric-related covariates, these were all constructed at the hexagonal level (road density, dead-ends, number of intersections, β and π). In this regard, betweenness was finally discarded for the analysis, as its consideration at the hexagonal level may be misleading. A road segment level analysis would be the right context for using this measure.

The rest of the covariates related to network's geometry were also constructed at the hexagonal level (road density, dead-ends, number of intersections, β and π). Despite this strategy leads to lose the road-segment level resolution that naturally fits some of these covariates, obtaining an aggregated value over a grid of areal units also helps to gain knowledge on network's structure at a different scale.

Finally, some basic demographic characteristics of the residents living at each hexagon of the grid were considered, which included the percentage of the population in the age range 16-24 years and over 65 years, and the percentage of immigrants. The availability of a dataset including both the number of residents in every building or house in Valencia and the geographical coordinates corresponding to each of the latter allowed computing these covariates with high accuracy.

Type of factor	Covariate	Description
Geometry-related	Road density	Total length (in meters) of the road network
	Dead-ends	Number of dead-end streets located along the road network
	No. of intersections	Number of vertex in the road network that connect three or more streets
	β	e/v
	π	L/T
Exposure-related	Cadastral value	Average official value (in euros per m^2) for dwellings in the road network
	Antiquity	Average antiquity (in years) for dwellings in the road network
	Occupation	Average percentage of occupied dwellings in the road network
Population-related	16-24 population	Percentage of population in the age range 16-24 years
	$\geq$ 65 population	Percentage of population in the age range $\geq$ 65 years
	Immigrant population	Percentage of immigrant population

Table 4.1: Description of the covariates employed in order to adjust the results derived from the classical Knox test. For the formulas of β and π, e denotes the number of edges of a road network, v the number of vertex, T its diameter (longest distance reachable within the network) and L its total length

		Mean	SD	Min	Max
Event	No. of burglaries	15.29	13.06	0.00	58.00
Exposure	No. of dwellings	2211.22	1366.93	53.00	5125.00
Covariates	Road density	4396.99	1535.57	1048.26	10038.66
	Dead-ends	3.09	2.70	0.00	11.00
	No. of intersections	35.04	22.94	4.00	186.00
	β	1.16	0.09	0.90	1.40
	π	5.01	2.00	1.14	12.23
	Antiquity	48.73	25.63	6.52	201.53
	Cadastral value	262.91	83.62	28.75	486.80
	Occupation	68.60	13.18	17.88	90.20
	16-24 population	8.25	1.92	0.00	14.29
	$\geq$ 65 population	20.03	8.35	3.08	93.75
	Immigrant population	11.77	5.23	0.00	32.35

Table 4.2: Summary statistics for the geometric and exposure related covariates at the hexagons forming the grid

4.4 Results and discussion

Adjusting the standard Knox test on the basis of the covariates considered required following Equation 4.3. In this regard, the time-dependent vector of coefficients $\alpha(t)$ was obtained through the fit of several GLMs (Equation 4.4) including the covariates described in Section 4.3.7 (previously standardized), considering the burglary counts observed (over the hexagonal grid of Figure 4.4a) at different periods within the years 2016 and 2017 as the response of the models. Specifically, the following five periods were determined through the changepoint analysis of the time series of burglary counts: January 2016-June 2016, July 2016-August 2016, September 2016-March 2017, April 2017-August 2017 and September 2017-December 2017. As it can be appreciated in Figure 4.2, the periods July 2016-August 2016 and April 2017-August 2017 witnessed the highest number of burglaries. The other three periods, however, represented medium-to-low risk periods for burglaries in Valencia for the years 2016 and 2017. Table 4.3 shows the coefficients derived from each of the GLMs models defined for each period within 2016-2017 indicated.

Covariate	**Jan 16-Jun 16**	**Jul 16-Aug 16**	**Sep 16-Mar 17**	**Apr 17-Aug 17**	**Sep 17-Dec 17**
Constant	-6.682*	-6.9878*	-6.5966*	-6.4245*	-7.1763*
Road density	0.3268	0.0616	0.3319	-0.0692	-0.1907
Dead-ends	-0.1131	-0.2185*	-0.0760	0.0502	-0.0241
No. of intersections	-0.0007	0.2134	-0.0305	0.1722	0.1663
β	-0.1510	-0.0839	-0.1744	-0.0514	-0.0906
π	-0.0041	-0.0197	0.0221	-0.0123	0.1490
Antiguity	0.0386	-0.0542	0.0732	0.2479*	0.0530
Cadastral value	0.1943	0.2609	0.0237	0.2188	0.1410
Occupation	0.3223*	0.1781	0.0727	0.2577*	0.1050
16-24 population	0.0001	-0.0293	0.0466	-0.1008	0.1819
$\geq$ 65 population	0.3418*	0.2045	0.2821*	0.3231*	0.2519
Immigrant population	0.0650	-0.0179	-0.1482	-0.0100	0.1296

Table 4.3: Coefficients obtained from the modellization of burglary counts (using GLMs) with a set of geometric and exposure related covariates at different periods of the years 2016 and 2017. An asterisk (*) indicates significance for any of the coefficients at the 0.1 level

Several spatio-temporal windows were defined in order to assess the existence of a near-repeat effect for burglaries in the city of Valencia, and to compare how both the standard and the adjusted version of the Knox test behave for the burglary dataset. The spatial component was represented by intervals of 120 meters length. Furthermore, in order to allow for the detection of spatially closer incidents, the first interval [0,120[was divided into three parts: 0 (same location),]0,60[and [60,120[meters. The temporal component was considered through several intervals (in days) representing a distance of one or several weeks: [0,7[, [7,14[, etc.

The results obtained from the application of the classical version of the Knox test considering the midpoint dates of burglaries in order to deal with time uncertainty are displayed in Table 4.4 (left). As expected, the near-repeat phenomenon drove to some extent the burglaries occurred in Valencia during 2016 and 2017, with an observed number of repeated burglaries (in the same location) within the same week around ten times higher than what would be expected by chance. With regard to the rest of spatio-temporal windows, the existence of a near-repeat phenomenon was clear for τ = [0,7[and every of the spatial intervals tested (in decreasing magnitude as the distance increases). In addition, near-repeats were also observed when considering longer periods. Specifically, the phenomenon was remarkably pronounced for $\sigma = 0$ and τ = [7,14[, and for σ =]0,60[with τ = [7,14[, [14,21[and [28,35[days.

On the other hand, adjusting the Knox test led to the values that are provided in Table 4.4 (right), next to the ones obtained for the standard test. The comparison of both sub-tables allows checking that the adjustment of the standard version of the test reduced the magnitude of many of the Knox ratios. For instance, the Knox ratio decreased from 10.88 to 10.42 (4.2% reduction) for $\sigma = 0$ and τ = [7,14[.

More interestingly, the level of statistical significance displayed by some of the estimated Knox ratios changes remarkably after the adjustment is made. In this regard, Table 4.5 shows the p-values associated with each Knox ratio, which are obtained from the simulated ratios obtained from the Monte Carlo (standard test) and

	Knox test					Adjusted Knox test				
σ τ	$[0,7[$	$[7,14[$	$[14,21[$	$[21,28[$	$[28,35[$	$[0,7[$	$[7,14[$	$[14,21[$	$[21,28[$	$[28,35[$
0	10.88**	2.16**	0.84	1.21	1.01	10.42**	2.09**	0.81	1.19	0.98
$]0,60[$	2.00**	1.59**	1.72**	1.09	1.80**	1.95**	1.59**	1.68**	1.09	1.76**
$[60,120[$	1.43**	1.11	1.22	0.96	0.99	1.36**	1.06	1.19	0.92	0.98
$[120,240[$	1.29**	1.10	1.05	1.25**	1.02	1.23**	1.06	1.01	1.22**	0.97
$[240,360[$	1.23**	1.18**	1.06	1.12*	1.12*	1.18**	1.14**	1.02	1.10*	1.10*
$[360,480[$	1.20**	1.13**	1.05	1.06	1.08*	1.16**	1.10*	1.02	1.03	1.07
$[480,600[$	1.11**	1.07*	1.06	1.10**	1.02	1.08*	1.04	1.03	1.08*	1.00

Table 4.4: Adjusted ratios for $X_{\sigma,\tau}$ (observed/expected) with different selections of σ (in meters) and τ (in days) for the standard and adjusted Knox test, employing 5000 simulations. A double asterisk (**) indicates the significance of the ratio at the 0.01 level, whereas a single asterisk (*) means the same for the 0.05 level. The significance of the ratios is derived from the empirical p-values estimated for each σ and τ from the simulations provided by the Monte Carlo and Metropolis-Hastings procedures

	Knox test					Adjusted Knox test				
σ τ	$[0,7[$	$[7,14[$	$[14,21[$	$[21,28[$	$[28,35[$	$[0,7[$	$[7,14[$	$[14,21[$	$[21,28[$	$[28,35[$
0	0.0000	0.0001	0.7452	0.2988	0.5260	0.0000	0.0000	0.7550	0.3140	0.5720
$]0,60[$	0.0000	0.0011	0.0005	0.3421	0.0000	0.0000	0.0048	0.0006	0.3388	0.0000
$[60,120[$	0.0018	0.2119	0.0508	0.6289	0.5511	0.0008	0.3328	0.0694	0.7558	0.5828
$[120,240[$	0.0000	0.0604	0.2355	0.0000	0.3773	0.0006	0.1564	0.4418	0.0008	0.6038
$[240,360[$	0.0000	0.0006	0.1160	0.0100	0.0113	0.0000	0.0020	0.3746	0.0162	0.0290
$[360,480[$	0.0000	0.0011	0.1117	0.0890	0.0285	0.0000	0.0204	0.2814	0.2292	0.0758
$[480,600[$	0.0039	0.0333	0.0683	0.0054	0.3016	0.0160	0.1500	0.2322	0.0430	0.5008

Table 4.5: Empirical p-values obtained for the Knox ratios shown in Table 4.4. The cells that are shaded correspond to $p < 0.01$ (darker colour) and $0.01 \leq p < 0.05$ (lighter colour)

Metropolis-Hastings (adjusted test) procedures. Indeed, if two common significance levels as 0.01 and 0.05 are selected (for illustrative purposes) to decide which spatio-temporal intervals are implicated in the near-repeat phenomenon, several differences arise (Table 4.5). Concretely, at the 0.01 level, the Knox ratios corresponding to three combinations of σ and τ (σ = [480,600[with τ = [0,7[, σ = [360,480[with τ = [7,14[, and σ = [480,600[with τ = [21,28[) are not statistically significant when considering the adjusted p-values. Similarly, at the 0.05 level there are two combinations of σ and τ that are affected by the adjustment of the test (σ = [480,600[with τ = [7,14[, and σ = [360,480[with τ = [28,35[).

In view of the consequences that adjusting the Knox test had on the results for the burglary dataset analyzed, one could think that the effort carried out to adjust the classical version of the Knox test was not worth it. In fact, the Knox ratios corresponding to the values of σ and τ that are more strongly involved with the near-repeat of burglaries in Valencia were only moderately affected by the adjustment. There seems to be no way to predict how much the adjustment of the Knox test will change the results that the standard test yields but, in any case, adjusting the Knox test should not be overlooked under the presence of crime risk heterogeneity in space and time. Furthermore, it is worth noting that from a practical point of view, discarding only one spatio-temporal with regard to its connection with the near-repeat phenomenon may be of value. In this analysis, for instance, adjusting the Knox test revealed that although the first week after a burglary's occurrence presents an elevated risk of repetition for all the spatial intervals considered up to 480 m, this effect seems to dissipate from this threshold. In addition, it can also be appreciated that the statistical significance of the Knox ratio corresponding to $\sigma = [480{,}600[$ and $\tau = [21{,}28[$ gets dramatically reduced if the test is adjusted. These "small" differences can be truly important in practice.

4.5 Conclusions

Despite its limitations, the Knox test seems to be still the most used tool in the field of quantitative criminology to assess the near-repeat phenomenon. The greatest limitation that arises when using the Knox test is the impossibility of bringing crime heterogeneity into the equation.

The primary goal of this study was purely methodological: highlighting the necessity of adjusting the basic version of the Knox test to account for crime heterogeneity to obtain a more realistic picture of the magnitude and spatio-temporal extent of the near-event phenomenon in the area under analysis. To this aim, this methodological advice was put into practice using a dataset of burglaries occurred in Valencia (Spain).

In order to adjust the Knox test for this dataset, the method proposed in Schmertmann (2015) was modified in order to adjust the classical version of the Knox test through covariate effects varying in time. Under the presence of exposure shifts and spatio-temporal variations among the covariates (which was not the case), the adjustment proposed by Schmertmann (2015) could be followed straightly. The adjustment required the use of the Metropolis-Hastings algorithm, which allowed approximating the empirical distribution of the Knox ratio for each of the spatio-temporal intervals selected for the analysis.

Several covariates related to road network's geometry, dwelling characteristics and demography were considered. The selection of the covariates relied on a literature review on the topic of crime risk heterogeneity particularly focused on burglary events. Despite the fact that the construction of certain covariates is not always possible, the use of sensible proxies should be generally accurate enough to fulfill the purpose of adjusting the standard Knox test according to some of the factors that may influence crime risk.

The adjustment of the Knox test described in this Chapter requires the partition of the space-time. Space can be segmented following an established administrative division, or using a regular grid. Regarding the segmentation of the whole period under investigation into several subperiods, a data-based procedure was followed (changepoint methodology for time series analysis). This step could be carried out in a different way choosing, for instance, a periodic segmentation of the temporal window. The effect of the choice of a certain spatial or temporal subdivision on the adjustment may be severe, so sensitivity analyses would be advisable. The finding of an “optimal” space-time segmentation, somehow meaning that it maximizes space-time variation, was out of the scope of this investigation, but it is a topic of interest.

Regarding the analysis performed on the burglary dataset described, the differences yielded by the adjusted version of the Knox test (in comparison with the standard test) may be perceived as only modest, although several remarkable variations were actually observed. It is worth noting that under a considerably larger temporal window (for instance, a decade), the spatio-temporal variation of both exposure (dwellings) and covariate effects would have been greater, and then the consequences of adjusting the Knox test would likely have been more notorious. In any case, under a scenario of spatio-temporal crime risk heterogeneity, considering the use of an adjusted version of the Knox test to favor a more accurate analysis of the near-repeat phenomenon is strongly recommended.

Chapter 5

The modifiable areal unit problem: Accounting for scale and zoning

In this Chapter, a complete investigation of MAUP effects is carried out using a dataset of traffic crashes occurred in Valencia (Spain). Whereas some related papers in the field have only focused on scale (Lee et al., 2014; Xu et al., 2014), most of them have tested different zonings without controlling the scale factor explicitly. This fact makes it challenging to determine if MAUP effects are a consequence of scale, zoning or the interaction between the two. This study tries to fill this gap with a simultaneous investigation of several basic spatial units (BSUs) and aggregation levels that allow the distinction between scale and zoning effects, in seeking to provide a more complete depiction of the phenomenon. Two modelling approaches, conditional autoregressive models and geographically weighted regressions have been used for this objective, following the choices of similar analyses. Furthermore, it has been specifically investigated how the changes in scale or zoning affect several questions involved in any macroscopic statistical modelling. These include the spatial autocorrelation of the covariates, multicollinearity among covariates and the basic distributional characteristics of the response variable. The investigation of MAUP usually focuses on the changes that finally arise in the estimation of model parameters after a switch of scale or zoning, but the changes in the underlying characteristics of the data being modelled are frequently overlooked. More insights on this issue are also provided.

5.1 Introduction

Traffic safety analysis at the macroscopic level requires the definition of a basic spatial unit (BSU) for performing the analysis. Hence, the whole area of investigation needs to be covered by BSUs that allow researchers to analyze the incidence and causality of traffic crashes across it. The definition of BSUs can be done both manually, through the advice of experts of the field, or automatically on the basis of an algorithm specifically designed for BSU delineation.

The choice of a certain BSU over an area of interest is closely related to the well-known modifiable areal unit problem (MAUP). MAUP refers to the effects that carries the change from a collection of BSUs to another with regard to statistical inference and interpretation (Openshaw, 1984). In a seminal paper, Openshaw (1977) presented the two main factors that need to be addressed for the delineation of an area into BSUs: scale and zoning. Scale, or aggregation level, refers to the number of zones in which the whole area of study is subdivided for performing the analysis. Hence, given a scale, zoning is the way the BSUs are joined forming the zones of analysis while preserving the specified scale. Openshaw (1977) proved that the election of the zones has an effect on spatial interaction models, in terms of fitting and parameter estimates. For this reason, he proposed a methodology in order to find the zoning subdivision of the area of analysis that optimizes model performance. More specifically, Openshaw also studied the consequences of MAUP on linear regression (Openshaw, 1978) and correlation coefficients (Openshaw, 1979), although he recognized that there are high difficulties for assessing the problem theoretically, leaving simulation studies as the main tool available for its approach. Years later, Fotheringham and Wong (1991) extended the examination of MAUP to multivariate statistical analysis, considering multiple linear regression and multiple logistic regression within the context of two classical administrative divisions: block groups and census tracts. The aggregation of both divisions at different scales allowed observing that MAUP was capable of creating a severe instability in parameter estimates when the zoning or, more remarkably, the scale were modified. These authors found that the interpretation of some of the variables included in the models could be dramatically altered due to MAUP, as changes in the signs of the parameter estimates were appreciated. In addition, goodness of fit was observed to grow monotonically as aggregation level got increased. In a more observational work, Lee et al. (2018) have investigated the effects of MAUP in means, variances and Moran coefficients, considering several scales and levels of autocorrelation for the variables involved. They have concluded that MAUP effects are not strong on means, unless a very high spatial autocorrelation is present, and that higher levels of aggregation tend to decrease the variance. Finally, it is worth noting that despite the fact that the study of the MAUP and its consequences in statistical inference have mainly been of descriptive or exploratory nature, recent works are trying to fill this gap by providing more accurate measurements of MAUP effects. Remarkably, Duque et al. (2018) have proposed a nonparametric test, S-maup, that measures the sensitivity to MAUP of a spatially intensive variable. Therefore, the S-maup test can be used to determine the level of aggregation at which MAUP effects do not impact the statistical analysis severely. As a drawback, the S-maup test lacks a theoretical definition. Indeed, an extensive simulation procedure was implemented by the authors in order to be able to supply critical values for different levels of scale and autocorrelation for the spatial variable.

5.1.1 TAZ delineation and MAUP effects in traffic safety analysis

The convenient delineation of an area into traffic analysis zones (TAZs), which behave as BSUs from the perspective of the present research, requires several considerations to be made, although the guidelines suggested in literature are usually varied and even contradictory. In a pioneering work, O'Neill (1991) proposed six criteria for the delineation of TAZs, including the homogeneity of socioeconomic characteristics, population and trip attraction levels, the minimization of intrazonal trips and the employment of physical, historical or administrative boundaries. Martínez et al. (2009) made use of a mobility survey available for the city of Lisbon (Portugal) in order to design TAZs fitting the following four criteria: boundaries are set over roads presenting a low trip generation density, intra-TAZ trips are minimized, TAZs with a very low or large number of trips are avoided and homogeneity within a TAZ is pursued as much as possible. Dong et al. (2015) used a K-means algorithm to classify a predefined set of cell areas according to primary features (traffic volume, hourly inflow, outflow and incremental flow) and optimizing features (peak and valley values for the primary features).

Efforts have also been made in order to account for the boundary effect in TAZ delineation. Siddiqui and Abdel-Aty (2012) proposed the distinction between boundary and interior pedestrian crashes considering a buffer of 100 ft from TAZs boundaries. Covariate information was weighted in the case of boundary crashes depending on the length of shared boundary between contiguous TAZs. Then, the specification of two analogous models for both types of crashes allowed the detection of differential effects for some of the covariates included in the study. Furthermore, there exists some simple methods that allow the allocation of traffic crashes occurred near TAZ boundaries, including half-and-half ratio, one-to-one ratio and ratio of exposure. Very recently, Zhai et al. (2018a) proposed a novel model-based iterative method for assigning crashes located close to boundaries. This was proven to produce better predictions at the BSU level than the other boundary assignation methods and to increase the number of significant covariates detected.

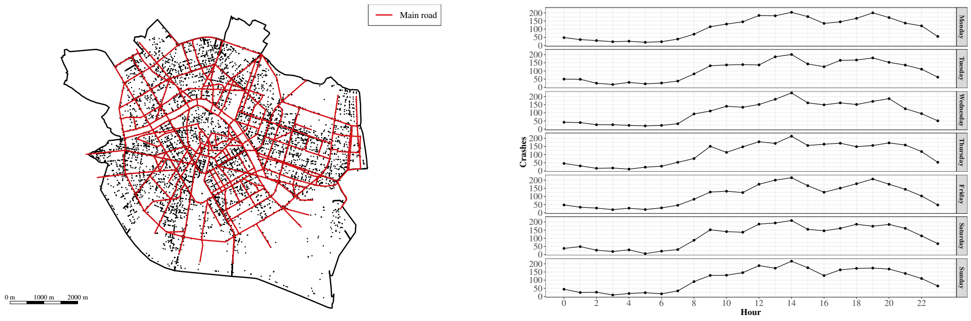
Several authors have investigated MAUP in the context of traffic safety analysis, which are now briefly discussed. First, Thomas (1996) noted that changes in scale may alter the probability distribution that best fits the nature of the available crash counts. Indeed, Thomas (1996) worked at the road segment level in order to infer three length thresholds that would require a distinct modelling strategy for crash counts: a Poisson distribution for very short segments (less than 1 hm), an intermediate empirical distribution for middle segments and a normal distribution for long segments (more than 20 hm).

In the last decade, however, most of the research studies related to MAUP were settled in the context of areal units of analysis. For instance, Lee et al. (2014) took the more than 1000 TAZs already defined for several counties in Central Florida (USA) and combined them into new subdivisions of the space containing from 100 to 1000 TAZs (in intervals of 100 TAZs). Specifically, total crash rates available for the period of study were employed by the regionalization algorithm for the obtention of sets of homogeneous zones containing the different number of TAZs specified. The Brown-Forsythe test was applied in order to check how the changes in TAZs affected the variance of crash rates. A moderate value of this test with a certain TAZ system represents an optimal situation, which means that the scale at which the TAZs are defined is suitable for detecting both local and global variation. Thus, the definition of 500-700 TAZs was found optimal for the data analyzed in Lee et al. (2014). Xu et al. (2014) tested different TAZ schemes including from 50 to 738 units of analysis. They suggested the use of 350 or more TAZs (for their case study) in order to reduce MAUP effects because for this scale they found a superior number of significant covariates and more stable coefficient estimations. Similarly, Ukkusuri et al. (2012), used ZIP codes and census tracts as TAZs, determined that a finer aggregation level (census tracts in their study) was more suitable for data modelling as it enables a higher data variability and greater explanatory power. Abdel-Aty et al. (2013) modelled crash counts occurred at two American counties at the level of TAZs, block groups and census tracts, considering total, severe and pedestrian crashes. Relevant differences were found in terms of the number of significant variables that were yielded by models based on different spatial units and, more specifically, in the type of factor (roadway related vs. commute related) providing more significant variables. Amoh-Gyimah et al. (2017) investigated several spatial units that may be used as TAZs, including statistical area levels, postal areas, state electoral divisions, grid cells and Thiessen polygons developed around the centroids of Melbourne Integrated Transport Model. These authors made use of several statistical models to provide a more complete perspective of the effects that the choice of TAZs can lead to. The presence of MAUP was evident as they observed that a reduction in the number of zones produced an increase in the number of significant variables. Furthermore, they concluded that the selection of the modelling technique is another important factor that may reduce MAUP. Indeed, geographically weighted Poisson regression appeared to be less affected by MAUP than random parameter negative binomial. These authors also suggest the use of Thiessen-based and grid cells for prediction purposes, according to their results. Zhai et al. (2018b) applied a multivariate Poisson lognormal model with multivariate conditional auto-regressive prior on block groups, census tracts, zip codes and predefined BSUs for the analysis of traffic crashes occurred in a county of Florida (USA). Important variations were found in relation to coefficient sign, magnitude and significance, and the larger units showed a superior forecasting performance. In addition, the detection of high-crash locations revealed some unexpected situations, as certain zones that were shared by all the BSU configurations showed a completely opposite behaviour depending on the underlying BSU-dependent model being used for such assessment. Finally, in a review paper regarding the effects of MAUP in traffic safety analysis, Xu et al. (2018) proposed four potential solutions: avoiding data aggregation, considering the spatial variation of the covariates employed for data modelling (an issue that is usually skipped), defining an optimal zoning system for the analysis and conducting sensitivity analyses in order to check for MAUP presence and magnitude, regardless of the strategies undertaken for attempting its reduction.

5.2 Data

5.2.1 Crash dataset and road structure characteristics

A total of 18037 traffic crashes that took place in the city of Valencia (Spain) during the years 2014 and 2015 were analyzed (Figure 5.1a). Geographical coordinates for each of these crashes and information regarding the date and hour of occurrence were provided by the Local Police of Valencia. The available coordinates were used to locate the crashes on a spatial representation of the road network of the city (linear network), as a guarantee of accuracy. This road network has a length of 840.3 km (with a diameter of almost 11.6 km) and contains 6110 road intersections. Arterial roads of Valencia, which were employed to define BSUs, extend up to 168.3 km and are also displayed in Figure 5.1a.



(a) (b)

Figure 5.1: Points representing the locations of traffic crashes that occurred in Valencia during the years 2014 and 2015 (a) and time series (displayed by hour and weekday) of traffic crashes observed in Valencia during the same period (b)

5.2.2 Covariate definition

Several covariates were constructed to explain the incidence of traffic crashes among BSUs for the years of study, which were classified into environmental, network-related and socioeconomic. Environmental covariates included the consideration of different services (public or private) that are located along the road network which are known to influence the dynamics of traffic flow and in consequence are likely to affect crash rates. The services selected were schools (from preschool to high school level), bars/restaurants, hotels, private companies (mainly financial, legal or insurance), bus and tram stops.

Network-related covariates were precisely derived from the information provided by the road network structure, and included non-pedestrian road length, which was considered as an exposure, average road betweenness and number of road intersections (involving any road type, main or not). Betweenness is a measure of network connectivity and was computed for each road segment of the network according to the next formula (Freeman, 1977):

$$B_e = \sum_{i \sim j} \frac{\sigma_{ij}(e)}{\sigma_{ij}}$$

where i and j are vertex of the network that are connected by a path $(i \sim j)$, σ_{ij} the number of shortest paths between i and j and $\sigma_{ij}(e)$ the number of shortest paths that connect i and j while passing through the edge e of the network.

Finally, socioeconomic and demographic information was introduced through the percentages of population in the range 16-24 and over 65 years, and also with the average power of cars (in hp), which clearly correlates with economic status.

It is of need to highlight that the data that was used to construct this set of covariates for different zonal schemes and levels of aggregation was available in point-referenced format. Hence, it was possible to aggregate the data at any desired level of aggregation or zoning system. In addition, the availability of the digitized version of the road network of the city allowed the computation of the betweenness or the number of road intersections. All these steps were carried out through specific GIS functions available in the R programming language (R Core Team, 2018).

		BSU configuration							
		CTs		**TMs**		**TIs**		**HEXAs**	
Type	**Variable**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Crashes	No. of traffic crashes (CRASH)	31.74	36.11	31.29	30.50	47.52	46.52	34.43	35.86
Environmental	No. of undergraduate educational centres (EDU)	0.62	0.87	0.61	1.03	0.93	1.29	0.68	1.05
	No. of bars/restaurants (BAR)	8.49	9.35	8.37	10.56	12.71	13.97	9.31	13.54
	No. of companies (COMP)	29.54	49.66	29.13	34.28	44.23	49.12	32.41	60.72
	No. of hotel rooms (HOT)	14.34	71.58	14.14	53.21	21.47	66.85	15.76	63.04
	No. of parking zones (PARK)	0.16	0.47	0.16	0.44	0.24	0.62	0.18	0.52
	No. of bus stops (BUS)	1.67	2.18	1.64	1.74	2.50	2.39	1.82	1.77
	No. of tram stops (TRAM)	0.05	0.30	0.05	0.26	0.08	0.31	0.06	0.25
Traffic-related	Average betweenness (BETW)	518.97	1013.62	786.17	2020.76	760.32	1703.99	384.92	853.54
	Intersection density per road km (INT)	8.71	6.33	8.61	6.42	9.01	6.78	6.96	5.51
Socioeconomic/demographic	% of young (16-24 years) population (YP)	9.87	2.12	9.06	4.43	9.41	3.49	7.96	5.47
	% of old ($\geq$ 65 years) population (OP)	24.02	5.97	23.27	10.12	24.54	11.11	21.74	17.23
	Average horsepower of cars (HP)	12.25	0.72	11.78	2.88	12.09	2.13	10.67	4.22

Table 5.1: Description of the covariates defined for the analysis and basic statistics of these covariates for the four BSU configurations tested (in their original configuration, prior to aggregation/regionalization)

5.2.3 BSU definitions

In the absence of an established TAZ configuration for the city of Valencia (which is probably the most used areal unit in the field of traffic safety analysis), several possibilities were explored for the investigation of MAUP effects. The use of census tracts (CTs) of Valencia and a grid of hexagonal BSUs (HEXAs) are two easy-to-implement options that were tested. Particularly, CTs have been investigated in many previous studies (Wier et al., 2009; Abdel-Aty et al., 2013; Cai et al., 2017). Regarding the use of hexagons, these have been recommended over square grids in related literature on traffic safety and MAUP given its more compact shape (Loidl et al., 2016). The scarcity of road network at some areas in the North of Valencia led to join some of the hexagonal units that were covering them, but these modifications were minimal in relation to the whole hexagonal grid.

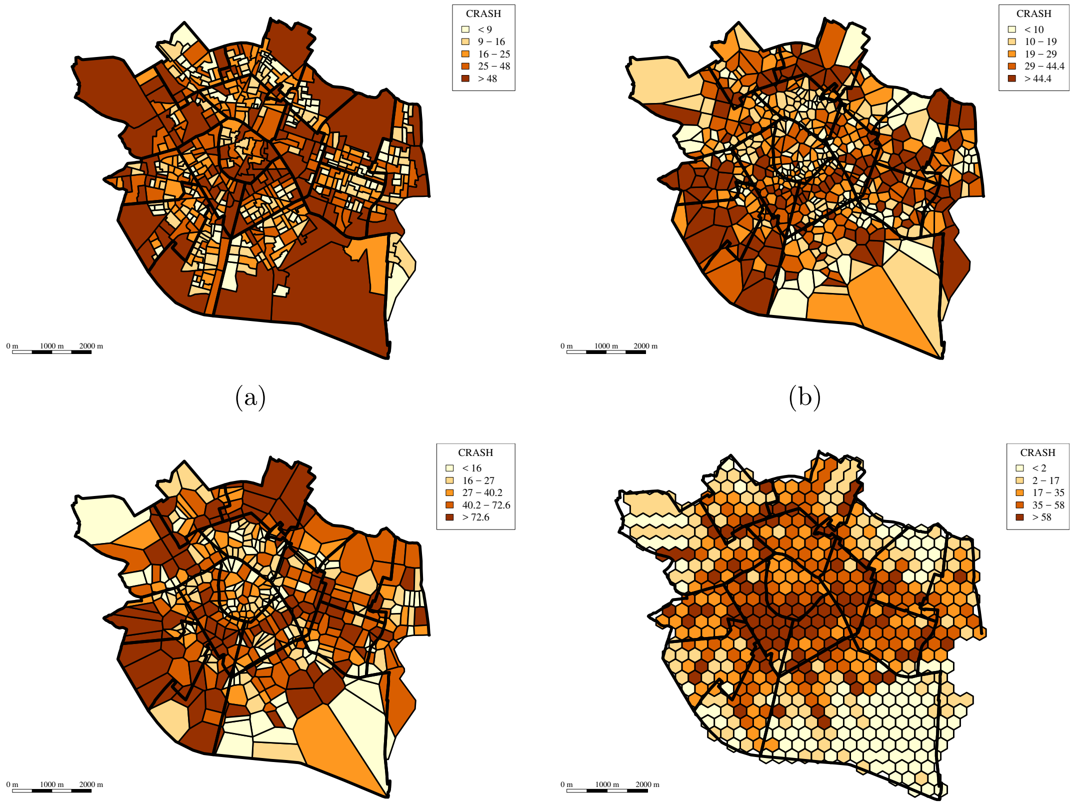
Furthermore, two more specific BSU schemes were delineated on the basis of two capital elements of any urban traffic network: main roads (segments) and intersections between main roads (points). It is known that these two road entities absorb a substantial percentage of traffic crashes, being the case of road intersections especially treated in literature (Miaou and Lord, 2003; Huang et al., 2017; Lee et al., 2017). Thiessen polygons (also known as Voronoi or Dirichlet polygons) were constructed around points located along main roads of the city and exactly at main intersections, generating two BSU types that hereinafter are referred to as TMs (Thiessen polygons based on main roads) and TIs (Thiessen polygons based on intersections between main roads). Given a collection of locations in a planar space, the Thiessen polygon built from one of these locations, P, contains all the points of the space that are closer to P than to any of the other locations established. Hence, each of the Thiessen polygons defined as a BSU was associated with a particular point along the main road structure (in-between a main road) or to a main intersection of the city. The use of TMs, TIs or HEXAs clearly alleviates the uncertainties derived from crashes located near BSU boundaries, which may have a strong effect in the case of CTs given the historical tendency of defining administrative divisions along main roads, where many crashes occur (Table 5.2).

Then, Figure 5.2 includes the four types of BSU configurations that were defined over the region of study, which provide enough evidence of how the change of the system alters substantially the spatial distribution of traffic crashes across the city. For instance, the central district of Valencia includes several CTs and HEXAs where the crash rate belongs to the highest quintile (Figures 5.2a and 5.2d), but this effect clearly reduces when the TMs and TIs are considered (Figures 5.2b and 5.2c).

The number of CTs in Valencia at the beginning of the year 2015 (566) served as a guide in order to define the other three BSU systems in a way they presented a comparable scale (similar number of BSUs). In the case of TIs, the initial scale was conditioned by the number of main intersections in Valencia, rendering it impossible the implementation of a finer BSU scheme of this nature, than that presented in Figure 5.2c. Thus, the four baseline configurations in Figure 5.2 composed of 566 CTs, 574 TMs, 378 TIs and 515 HEXAs were chosen to analyze the MAUP effect in the modelling of traffic crash counts for the available dataset.

BSU configuration	**< 5 m (%)**	**< 10 m (%)**	**< 20 m (%)**	**Mean distance (m)**
CTs	35.36	45.29	59.56	27.99
TMs	8.75	18.32	30.23	48.77
TIs	6.65	11.54	22.73	64.78
HEXAs	5.83	11.91	22.40	52.19

Table 5.2: Percentage of crashes located near BSU boundaries considering five distance thresholds (5, 10 and 20 m) and mean distance from crashes to BSU boundaries for the four BSU configurations employed in the analysis



(a)

(b)

(c)

(d)

Figure 5.2: Crash counts (CRASH) in Valencia considering a BSU configuration composed of CTs (a), TMs (b), TIs (c) and HEXAs (d). Districts of Valencia are overlayed (thicker lines, in black) for better readability and comparison

5.3 Methodology

5.3.1 Software

The R programming language (3.5.1 version, R Development Core Team, Vienna, Austria) (R Core Team, 2018) was used to obtain all the results presented in this work. The R packages ClustGeo (Chavent et al., 2017b), ggplot2 (Wickham, 2016), INLA (Rue et al., 2009; Martins et al., 2013; Lindgren and Rue, 2015), rgeos (Bivand and Rundel, 2018a), spatstat (Baddeley et al., 2015), spded (Bivand and Piras, 2015), spgwr (Bivand and Yu, 2017) and SpNetPrep (Briz-Redón, 2019) were specifically required for performing the analysis.

5.3.2 Regionalization algorithm

The term regionalization was defined by Guo (2008) as the process of aggregating a set of spatial entities into a reduced number of regions in a way that a predefined objective function is optimized. There are several important regionalization algorithms, including SKATER (Assunção et al., 2006), REDCAP (Guo, 2008) and ClustGeo (Chavent et al., 2017b). In this study the latter was chosen, which is implemented in the R package ClustGeo (Chavent et al., 2017a). The next paragraphs contain a brief description of how this method works and how it was used.

Given a number of clusters, K, to be formed and two matrices, D_0 and D_1, that represent the homogeneity and physical distances (respectively) between the spatial units available before regionalization, the ClustGeo algorithm relies on the minimization of a measure called mixed within-cluster inertia, defined as the sum of the mixed inertias of all of the clusters established. The mixed inertia of a cluster, C_k^α, follows the next expression (Chavent et al., 2017b):

$$I_\alpha(C_k^\alpha) = (1-\alpha)\sum_{i\in C_k^\alpha}\sum_{j\in C_k^\alpha}\frac{w_i w_j}{2\mu_k^\alpha}d_{0,ij}^2 + \alpha\sum_{i\in C_k^\alpha}\sum_{j\in C_k^\alpha}\frac{w_i w_j}{2\mu_k^\alpha}d_{1,ij}^2$$

where $\alpha \in [0,1]$ is a parameter that controls the importance that the homogeneity and physical distances (represented by D_0 and D_1) have in the clustering procedure, $k \in \{1,...,K\}$ is the index for the cluster, w_i is the weight of spatial unit i, $\mu_k^\alpha = \sum_{i\in C_k^\alpha} w_i$ and $d_{0,ij}$ (resp. $d_{1,ij}$) is the normalized dissimilarity between spatial units i and j in D_0 (resp. D_1).

In this study, the total number of crash counts registered per BSU during the period 2014-2015 at four time slots (23h-7h, 7h-14h, 14h-20h and 20h-23h, which were selected according to the daily trends observable in Figure 5.1b) and at the weekends were used to define the dissimilarity matrix D_0. Regarding D_1, this matrix was constructed from the Euclidean distances between the centroids of the BSUs. Furthermore, the weights (w_i) were set equal for all BSUs and a value of $\alpha = 0.1$ was chosen, giving much importance to the spatial distances between the BSUs during the aggregation procedure (the investigation of the optimal value of α suggested by ClustGeo led to this choice).

It needs to be remarked that there is a technical difference between SKATER and REDCAP algorithms and the method implemented in ClustGeo. Indeed, the choice of K in ClustGeo does not represent the number of contiguous and homogeneous regions that are created, but the number of homogeneous regions (according to the variables provided to the algorithm) that need to be regrouped later in order to fully satisfy the contiguity constraints. Hence, the input K in ClustGeo is a lower bound of the number of BSUs that are generated at the end of the process, although both values barely differ. The use of several values of K, from 100 to 500 in intervals of 100, allowed MAUP to be investigated in the present study with five different levels of spatial aggregation, which are denoted by AG100, AG200, AG300, AG400 and AG500 within the rest of the analysis.

5.3.3 Crash counts modelling

Conditional autoregressive model

The modelling of crash counts at the macroscopic level requires the consideration of data overdispersion. Two common choices in the field of traffic safety analysis to address this issue are negative binomial (also known as Poisson-Gamma) and Poisson lognormal probability distributions (Lord and Mannering, 2010). Both have their own advantages and disadvantages, Poisson lognormal being more recommended in cases of high overdispersion (particularly skewed distribution of the counts), whereas negative binomial has been suggested to be more suitable for moderately overdispersed counts, and also for counts with a large number of zeros (Khazraee et al., 2018; Shirazi and Lord, 2018). In the context of this study, it is hard to choose between one distribution or the other, as the change in scale or zoning alters the statistical properties that are involved in this decision. Anyhow, as the crash counts available under the different combinations of aggregation level and BSU type were overall only moderately overdispersed, a negative binomial distribution was selected.

Therefore, a conditional autoregressive (CAR) model with negative binomial (NB) response was chosen to fit the crash counts recorded for each BSU. The use of a CAR structure for modelling crash counts is a usual strategy in traffic safety analysis to account for spatial heterogeneity (Quddus, 2008; Huang et al., 2010).

If $Y \sim \text{NB}(\mu, \psi)$ (basic NB distribution of mean μ and shape ψ) then it holds that $E(Y) = \mu$, $V(Y) = \mu + \frac{\mu^2}{\psi}$ and $P(Y = x) = (\frac{x+\psi-1}{\psi-1})(\frac{\psi}{\mu+\psi})^{\psi}(\frac{\mu}{\mu+\psi})^{x}$. Then, assuming a NB distribution for the response (crash counts) the following spatial model was implemented:

$$\begin{aligned} Y_i &\sim \text{NB}(\mu_i, \psi) \\ \log(\mu_i) &= \log(E_i) + \beta_0 + \sum_{m=1}^{p} \beta_m X_{im} + \phi_i \end{aligned} \tag{5.1}$$

where Y_i is the number of crashes observed at BSU i, μ_i and ψ are, respectively, the mean risk (for BSU i) and overdispersion ($1/\psi$) values for the NB distribution, the natural logarithm acts as a link function for μ_i, E_i (exposure at BSU i) is the length of non-pedestrian road at BSU i which acts as an offset of the equation, X_{im} represents the value of the m-th covariate at BSU i, β_m is the coefficient that controls the effect of the m-th covariate and ϕ_i represents a spatial effect for BSU i. Regarding the selection of the exposure, the unavailability of vehicle miles travelled data (traffic volume) for non-main roads of Valencia left non-pedestrian road length as the natural choice, a possibility already considered in previous research studies (Qin et al., 2004; Imprialou et al., 2016).

The spatial effect in Equation 5.1 was modelled using the following CAR structure (Besag, 1974; Besag et al., 1991):

$$\phi_i \mid \phi_j, j \neq i \sim N(\sum_{j=1}^{n} w_{ij}\phi_j, \tau_i^{-1})$$

where w_{ij} is an indicator parameter that is 1 if BSUs i and j are contiguous and 0 otherwise, and τ_i is a precision parameter that varies with BSU i.

Geographically weighted regression

Geographically weighted regression (GWR) is a form of linear regression that captures the spatial heterogeneity present in the data by allowing model parameters to vary locally (Brunsdon et al., 1996; Fotheringham et al., 2002; Nakaya et al., 2005). GWR has already been used in traffic safety analysis (Hadayeghi et al., 2010; Matkan et al., 2011; Xu and Huang, 2015; Gomes et al., 2017), including some analyses from the perspective of MAUP effects (Amoh-Gyimah et al., 2017).

The mathematical expression that corresponds to the GWR model is the following:

$$\log(\mu_i) = \log(E_i) + \beta_0(\mathrm{BSU}_i) + \sum_{m=1}^{p} \beta_m(\mathrm{BSU}_i) X_{im} \tag{5.2}$$

where μ_i, E_i and X_{im} are as in Equation 5.1. The main feature of GWR is the consideration of local regression parameters (in contrast to global parameters of Equation 5.1) which are denoted by $\beta_m(\mathrm{BSU}_i)$ in Equation 5.2. As in Equation 5.1, a NB distribution was used in the definition of the model to consider overdispersion. A modification of GWR called semiparametric GWR consisting in the combination of fixed and spatially-varying effects for the covariates involved in the model has also been used in traffic safety analysis (Xu and Huang, 2015; Amoh-Gyimah et al., 2017). However, the classical version of the GWR model was employed in order to provide a more unified framework for the comparison of the set of models obtained for each aggregation level and zoning, which is the main purpose of this analysis.

Hence, a GWR model behaves similarly to a generalized linear model (GLM), although for the former the parameters that compose the model are estimated locally, at each BSU, depending on the crash counts and covariate values at the surrounding areal units. The influence that BSU i produces on another BSU j (denoted as w_{ij}) was controlled by the following Gaussian kernel function:

$$w_{ij} = e^{-0.5\frac{d_{ij}^2}{\sigma^2}}$$

where d_{ij} is the Euclidean distance between BSU i and BSU j (between their centroids) and σ is the fixed bandwidth employed by the kernel function, which represents the level of influence that the rest of BSUs have on a given BSU with regard to model fitting (a higher value for σ means that model parameters are estimated on the basis of a wider zone around each BSU). Several other kernel functions are available instead of the Gaussian (bisquare, for instance), but this choice is usually not responsible of strong effects on the results (Silverman, 2018).

Regarding the bandwidth, a value of $\sigma = 2$ km was chosen in this study for all the BSUs and aggregation levels being considered. Other authors opted for the choice of a specific optimal bandwidth for each BSU and aggregation level (Amoh-Gyimah et al., 2017), but here a fixed value was used in order to avoid the presence of a source of variation other than scale or zoning, which are the focus of the analysis. The value of 2 km was chosen because it was close to the optimal values that were observed for the different BSUs and aggregation levels tested.

5.3.4 Assessment of model performance

The goodness of fit of the CAR models was assessed through Bayesian deviance information criterion (DIC) (Spiegelhalter et al., 2002). Similarly, AIC was used for GWR models. In addition, several measurements of model performance typically used in other traffic safety analysis papers on the MAUP for model comparison were considered for both CAR and GWR models: mean absolute deviation (MAD) (Lee et al., 2014; Xu et al., 2014; Amoh-Gyimah et al., 2017; Zhai et al., 2018b), sum of absolute deviation (SAD) (Lee et al., 2014; Xu et al., 2018; Zhai et al., 2018b) and percent mean absolute deviation (PMAD) (Lee et al., 2014; Xu et al., 2018).

The formulas for MAD, SAD and PMAD are the following:

$$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n} |e_i|$$

$$\text{SAD} = \sum_{i=1}^{n} |e_i|$$

$$\text{PMAD} = \frac{\sum_{i=1}^{n} |e_i|}{\sum_{i=1}^{n} |y_i|}$$

where n is the number of BSUs and $e_i = y_i - \hat{y_i}$ represents the difference between the number of crashes observed at BSU i (y_i) and the number fitted by the model ($\hat{y_i}$).

From the perspective of interpreting the results, a lower value of any of the aforementioned statistics (DIC, AIC, MAD, SAD or PMAD) indicates a better fit to the available data.

5.3.5 Statistical tools for covariate exploration

Several statistical tools were used to explore the covariates provided to the models at different scales and zonings, and hence provide more instruments to analyze their sensitivity to the MAUP. This section includes a brief description of these tools.

Average nearest neighbour index (NNI) measures the level of clustering/dispersion of a point pattern. Hence, it is suitable for the exploration of a covariate constructed from a pattern of points located across the area of investigation (EDU, BAR, COMP, HOT, PARK, BUS and TRAM among the set of covariates used in the present research). The definition of the NNI is the following (Clark and Evans, 1954; Cressie, 1993):

$$\text{NNI} = \frac{\frac{1}{P}\sum_{i=1}^{P} D_{NN}(i)}{\frac{1}{2}\sqrt{\frac{A}{P}}}$$

where P is the number of points that form the pattern, A the area of the whole space where the pattern lies and $D_{NN}(i)$ is the distance from point i to its nearest neighbour (the closest point to i in the pattern). The NNI represents a ratio between the average nearest-neighbour distance observed for the pattern and the value that would be expected under the hypothesis of random spatial distribution. A NNI lower than 1 indicates that the pattern is clustered, whereas a value higher than 1 is a sign of the dispersion of the pattern. More specifically, the following z-score ($z \sim N(0,1)$) is associated to the computation of the NNI, which allows one to assess the statistical significance of the index under the null hypothesis of NNI $= 1$ (Clark and Evans, 1954):

$$z = \frac{\frac{1}{P}\sum_{i=1}^{P} D_{NN}(i) - \frac{1}{2}\sqrt{\frac{A}{P}}}{\frac{0.26136}{\sqrt{P^2/A}}}$$

Moran's I (Moran, 1950a,b) was computed for every combination of BSU, aggregation level and covariate available. Moran's I is defined as follows:

$$I = \frac{\displaystyle\sum_{i=1}^{n} \sum_{j \in N(i)} \frac{1}{n_i}(x_i - \bar{x})(x_j - \bar{x})}{\displaystyle\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

where $N(i)$ represents the set formed by the neighbours of BSU i, which has cardinality n_i, x_i is the value of a covariate at BSU i and $\bar{x}$ the mean value of the covariate across all spatial units available. Moran's I behaves as a spatial autocorrelation coefficient for areal-based data. Under the hypothesis of no spatial autocorrelation, it holds that $E(I) = -1/(n-1)$, where n is the number of spatial units in each case. A higher Moran's I value indicates a higher tendency of the covariate to show strongly associated values for neighbouring BSUs.

Multicollinearity among the covariates considered was investigated through the Variance Inflation Factor (VIF) (Fox, 1991), which is defined as follows for a given covariate or predictor, X_j:

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the R^2 found when regressing all other covariates onto X_j (Miles, 2014). A higher value of VIF suggests that the covariate is more susceptible to lead to multicollinearity issues. A value of VIF not greater than 10 is usually interpreted in literature as a sign of no severe multicollinearity (Miles, 2014).

5.4 Results and discussion

Tables 5.4-5.7 include the parameters estimated for the CAR models considering each BSU configuration and aggregation level. On the other hand, Figures 5.6-5.9 display the densities (distributions) of the local parameter estimates obtained from the GWR models for each BSU type, aggregation level and covariate. The distributions of these local parameters were scaled (divided by their standard deviation) to facilitate the graphical comparison.

The main conclusion that yields from all these results is that MAUP effects have heavily affected the macroscopic traffic safety analysis performed. In the present section, an initial subsection gives insights into the varying nature of the data (response and covariates) as one changes scale or zoning. The subsequent subsections contain a description of how the variations in aggregation level (scale) or BSU type (zoning) changed model parameter estimates and a performance comparison of all the models implemented. After that, the associations derived from all the models and BSU configurations tested are globally evaluated in pursuit of more solid conclusions (despite the consequences of MAUP) that allow the identification of some factors that correlate with more traffic crashes.

5.4.1 Effects of MAUP on input data

Before analyzing scale and zoning effects on parameter estimations considering both CAR and GWR models, an investigation of the consequences of MAUP on the input data that is afterwards modelled (response and covariates) was performed.

The distributional characteristics of the response (crash counts) for different BSUs and aggregation levels are shown in Table 5.3, from which it yields that overdispersion (through the coefficient of variation) and kurtosis reduced as the level of aggregation was increased. Furthermore, it can also be observed in Table 5.3 that the percentage of zeros (percentage of BSUs where no traffic accident was recorded) was almost negligible for most of the combinations between a BSU and an aggregation level, excluding the HEXAs.

The magnitude of the spatial autocorrelation shown by the response variable and the covariates also suffers from scale and zoning changes. Figure 5.3 displays all Moran's I indexes computed for each combination of a BSU and aggregation level, which suggests that it is hard to predict the level of spatial autocorrelation after a change of scale. Indeed, some covariates tend to be more spatially autocorrelated as aggregation increases (the number of parking zones, for TMs), whereas other covariates show the opposite behaviour (the number of companies, for HEXAs). In addition, CTs and HEXAs show overall higher levels of spatial autocorrelation for the covariates than TMs and TIs. This is remarkable for traffic crashes, which display a particularly high spatial autocorrelation in the case of HEXAs, rather than in the other three BSUs considered for investigation. The spatial autocorrelation of some of the covariates is more deeply discussed in the following subsections.

Finally, the computation of variance inflation factors (VIF) leads to the conclusion that multicollinearity issues were not present for the distinct dataset analyzed (at different scales and zonings), as VIF factors were always below 10 (Figure 5.4). However, it is important to appreciate that VIF consistently increased as the level of aggregation increased, specially for some covariates such as the number of bars/restaurants (BAR), the number of companies (COMP) and intersection density (INT). Hence, this analysis suggests that the level of aggregation should not be excessively increased in order to avoid multicollinearity among the covariates.

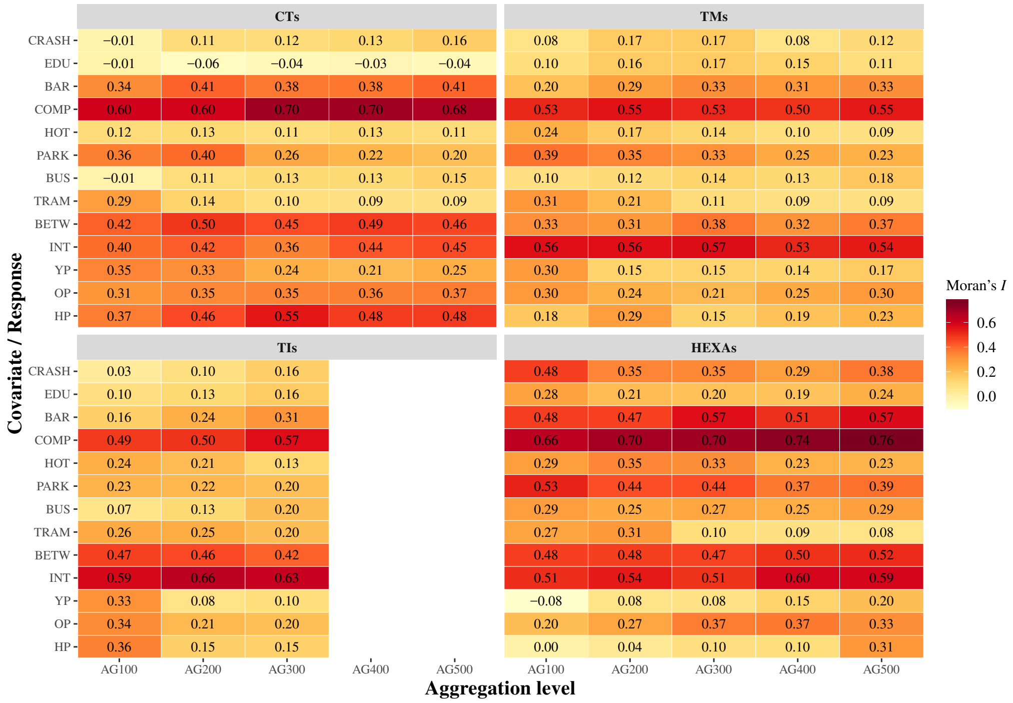


Figure 5.3: Moran's I values for the response (CRASH) and the covariates for each BSU type and aggregation level tested

5.4.2 Parameter variations across aggregation levels

Given the moderate level of significance achieved by the set of covariates, an 80% credibility level was also considered along with the most usual 90% level for the estimations yielded by the CAR models. Figure 5.5 displays a graphical summary of the significance achieved by all the covariates involved in the analysis. According to Figure 5.5, the hypothesis of obtaining more significant variables and the consequent higher interpretation power at lower levels of aggregation suggested by Xu et al. (2014) was true in the case of CTs, but it was not clear, at all, for the rest of BSU configurations. Anyhow, this question could have been better addressed in the presence of a higher number of significant covariates.

The level of aggregation applied to each BSU configuration through the regionalization algorithm produced moderate-to-severe effects in parameter estimates for the CAR models. Hence, although the parameter estimates evolved moderately with changes in the scale (Tables 5.4-5.7), several covariates were only significant at some of the aggregation levels tested. However, some of the covariates did not seem affected by MAUP and remained significant with each aggregation level (old population percentage and average horsepower for CTs, and number of educational centers and average horsepower for both TMs and TIs). Despite not being significant for the most aggregated scheme considered (AG100), the positive association between traffic crashes and the number of bus stops for HEXAs was also consistent. Remarkably, none of the covariates that were found significant (at 80% of credibility) experimented a change of effect (from positive to negative, or vice versa) after

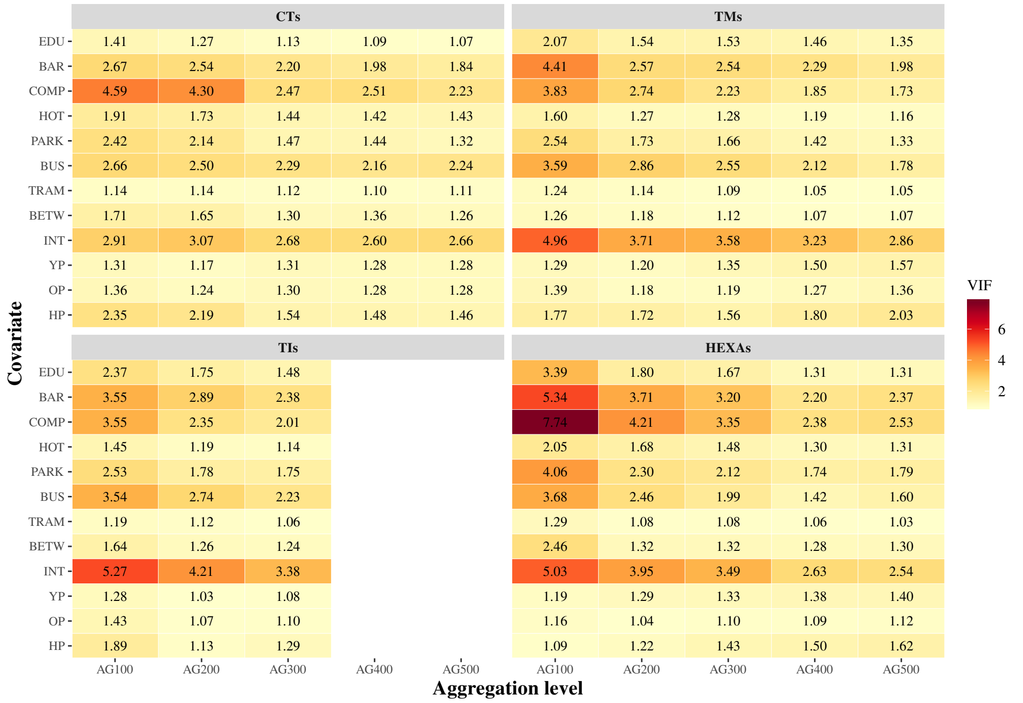


Figure 5.4: Assessment of multicollinearity among the covariates through variance inflation factor (VIF) for each BSU type and aggregation level tested

a shift in the scale. This is a positive result, since it indicates that MAUP effects were not the strongest possible across aggregation levels.

Regarding the GWR models, Figures 5.6-5.9 show that local parameter distributions may vary acutely after some changes on the aggregation level. It is hard to assess if the distribution of the local parameters tends to be more concentrated (leptokurtic) or flat (platykurtic) as the level of aggregation increases/decreases, as this seems strongly dependent on the covariate and the BSU. Furthermore, the contradictory presence of local parameters of opposite signs that takes place for most of the covariates is a well-known issue that often arises in GWR models (Hadayeghi et al., 2010; Xu and Huang, 2015; Amoh-Gyimah et al., 2017). Figure 5.10 shows the behaviour of the local parameter estimates obtained from the GWR models through the signs of 5th, 10th, 20th, 80th, 90th and 95th percentiles. Hence, a negative value for 80th, 90th or 95th percentiles indicates a high agreement among the local GWR coefficients and a negative association of the covariate with crash counts. Analogously, a positive 5th, 10th or 20th percentiles means the same but for a positive association.

In contrast to CAR models, for which no covariate showed a significant change of effect after a variation in scale, the GWR models experimented this issue for the covariates representing the number of companies (COMP), parking zones (PARK) and intersection density (INT) when using CTs (considering the percentile-based criteria that has been employed for assessing the association between crash counts and covariates with the GWR models). This lack of coherence was specifically serious for the number of companies, a covariate that showed a strong autocorrelation accord-

BSU	AG	CV	Kurtosis	Zeros (%)
CT	AG100	0.71	1.77	0.00
CT	AG200	0.87	3.41	0.00
CT	AG300	0.88	4.71	0.00
CT	AG400	0.92	10.86	0.25
CT	AG500	1.04	16.07	0.40
TM	AG100	0.54	1.76	0.00
TM	AG200	0.72	0.59	0.00
TM	AG300	0.76	2.11	0.33
TM	AG400	0.79	7.07	0.50
TM	AG500	0.88	12.03	0.60
TI	AG100	0.69	1.65	0.00
TI	AG200	0.77	1.49	0.00
TI	AG300	0.81	6.77	0.00
HEXA	AG100	0.85	2.32	2.68
HEXA	AG200	0.83	-0.15	6.50
HEXA	AG300	0.88	1.63	7.00
HEXA	AG400	0.84	3.75	5.75
HEXA	AG500	1.01	4.70	10.00

Table 5.3: Basic distributional properties of the response variable (crash counts) corresponding to each BSU type and aggregation level, where CV means coefficient of variation (standard deviation to mean ratio)

ing to Moran's I at this BSU system (Figure 5.3). The exploration of crash counts and the number of companies (COMP) at AG100, AG300 and AG500 (Figure 5.12) unveils some singular patterns around the city center (some surrounding Districts are highlighted in blue in Figure 5.12). Thus, whereas at AG100 most of the BSUs presenting a high number of companies were located within these Districts, the use of a more disaggregated configuration provided greater variation in the number of companies across the whole city, with many more BSUs in the periphery of Valencia presenting high values. With regard to intersection density (INT), this covariate also presented high Moran's I values (Figure 5.3), but visual inspection was far less clear than in the case of the number of companies, becoming challenging to figure out how the effect of intersections changed from AG300 to AG400 and again from AG400 to AG500 (for CTs and the GWR models). On the other hand, it is remarkable how the number of educational centers (EDU), which presents the more coherent behaviour across BSUs and aggregation levels in the case of the GWR models, is one of the covariates that showed a lower range of values for Moran's I statistic. Similarly, this result also agrees with that provided by the NNI (Table 5.8), as some of the covariates based on point patterns lying over the city presenting a high level of clustering (companies, NNI=0.26) display a more sensitive to MAUP behaviour than other that, albeit clustered, show a more regular pattern (educational centers, NNI=0.86). A similar level of consistence to that found for the educational centers was also obtained for the number of tram stops (TRAM) in the case of the GWR models, although this covariate resulted non-significant for almost all combinations

	AG100		AG200		AG300		AG400		AG500	
Covariate	**Est.**	**SD**	**Est.**	**SD**	**Est.**	**SD**	**Est.**	**SD**	**Est.**	**SD**
Intercept	-13.2477*	2.6119	-11.4619*	1.8909	-8.8597*	1.6753	-8.0273*	1.1308	-8.1816*	1.0403
EDU	-0.0327	0.0261	-0.0661*	0.0260	-0.0797*	0.0285	-0.0778*	0.0303	-0.0656*	0.0308
BAR	0.0025	0.0027	0.0030	0.0027	0.0021	0.0033	-0.0017	0.0035	0.0004	0.0042
COMP	-0.0005	0.0006	-0.0005	0.0006	0.0000	0.0009	0.0002	0.0009	0.0002	0.0009
HOT	0.0002	0.0004	0.0002	0.0003	0.0003	0.0004	0.0001	0.0004	0.0001	0.0004
PARK	-0.0230	0.0657	-0.0970	0.0607	-0.0187	0.0665	0.0102	0.0659	-0.0249	0.0661
BUS	0.0006	0.0134	0.0224*	0.0121	0.0186	0.0141	0.0156	0.0150	0.0222	0.0155
TRAM	-0.0436	0.0949	0.0424	0.0763	-0.0065	0.0924	-0.0711	0.0906	-0.0586	0.0911
BETW	0.0001	0.0001	0.0001	0.0001	0.0001	0.0000	0.0001	0.0000	0.0001*	0.0000
INT	-0.0130	0.0187	-0.0053	0.0122	-0.0010	0.0093	0.0095	0.0076	0.0051	0.0065
YP	-0.0265	0.0523	-0.0019	0.0313	0.0260	0.0245	0.0122	0.0206	0.0170	0.0176
OP	0.0322*	0.0151	0.0156	0.0096	0.0161*	0.0085	0.0120*	0.0072	0.0131*	0.0063
HP	0.7517*	0.2087	0.6004*	0.1487	0.3605*	0.1355	0.3111*	0.0921	0.3169*	0.0842
ψ	3.0831*	0.4127	8.8277*	2.9744	6.5329*	1.8981	5.4961*	1.3313	5.2406*	1.0798

Table 5.4: Estimates with standard deviation (SD) for the parameters involved in the CAR model, considering CTs as BSUs. An asterisk (*) indicates the significance of a parameter with a 90% credibility

of aggregation level and BSU for the CAR models. The NNI of the point pattern formed by the tram stops was 1.30 (Table 5.8), clearly indicating the dispersed configuration of these stops across Valencia.

5.4.3 Parameter variations across BSU types

Contrary to scale, MAUP effects from zoning variations were by far more severe in this case study. Indeed, some covariates showed a significant and opposite effect depending on the BSU system being considered, including old population percentage, the number of bus stops and intersection density (Figure 5.5). The cases of both the number of bus stops and intersection density were specifically related to the differential behaviour exhibited by the highest aggregation level, AG100. This aggregation level was clearly the least coherent among all the levels tested, possibly indicating its unsuitability to capture some micro-variations present in the data. On the other hand, the percentage of old population (OP) appeared as a highly sensitive-to-MAUP covariate, standing out from all the ones supplied to the models. Thus, whereas this covariate showed a clear positive association with traffic crashes for CTs and, to a lesser extent, for TIs, it associated with a decrease in crash counts with HEXA units at the three most aggregated levels. This contradictory result was investigated through the cartographic representation of crash counts (CRASH) and the old population covariate for CTs and HEXAs at AG100, AG200 and AG300 (Figure 5.13). In all the maps available in Figure 5.13, the border of a census tract located in the South of Valencia (which is the largest of the city) is highlighted in blue. This census tract constitutes a wide area of low population density and a high percentage of residents with 65 or more years of age (OP). In addition, the area is not dense in road network, which naturally reduces the number of traffic crashes. Hence, the use of HEXAs led to a covering of this census tract with several hexagons of very high percentage of old population and very low number of crashes, which surely affected the estimation of the parameter related to this covariate. On

	AG100		**AG200**		**AG300**		**AG400**		**AG500**	
Covariate	**Est.**	**SD**	**Est.**	**SD**	**Est.**	**SD**	**Est.**	**SD**	**Est.**	**SD**
Intercept	-8.3199*	1.2364	-7.7404*	1.0021	-7.2087*	0.7263	-6.7423*	0.6376	-6.0589*	0.5407
EDU	-0.0475*	0.0270	-0.0515*	0.0279	-0.0732*	0.0316	-0.0963*	0.0325	-0.0967*	0.0324
BAR	0.0019	0.0026	0.0029	0.0028	-0.0002	0.0034	-0.0016	0.0035	-0.0031	0.0038
COMP	-0.0002	0.0006	0.0001	0.0009	-0.0002	0.0010	0.0002	0.0011	-0.0001	0.0013
HOT	-0.0001	0.0004	0.0007	0.0005	0.0009	0.0006	0.0008	0.0006	0.0010*	0.0006
PARK	-0.0126	0.0594	-0.0397	0.0625	-0.0185	0.0800	-0.0870	0.0799	-0.0357	0.0788
BUS	-0.0380*	0.0118	-0.0010	0.0140	-0.0039	0.0163	-0.0086	0.0193	0.0079	0.0208
TRAM	-0.1556*	0.0792	0.0021	0.0988	-0.0631	0.1141	-0.1625	0.1170	-0.1163	0.1138
BETW	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
INT	-0.0033	0.0164	0.0081	0.0107	0.0115	0.0098	0.0000	0.0075	-0.0006	0.0066
YP	0.0082	0.0398	0.0365	0.0268	0.0042	0.0195	0.0084	0.0165	0.0061	0.0143
OP	0.0339*	0.0122	-0.0039	0.0083	0.0026	0.0059	0.0038	0.0052	0.0021	0.0047
HP	0.3635*	0.0993	0.3028*	0.0755	0.2849*	0.0568	0.2565*	0.0501	0.1996*	0.0428
ψ	3.8533*	0.5287	8.4472*	1.7932	3.1448*	0.5569	2.7796*	0.3599	2.9332*	0.3634

Table 5.5: Estimates with standard deviation (SD) for the parameters involved in the CAR model, considering TMs as BSUs. An asterisk (*) indicates the significance of a parameter with a 90% credibility

the other hand, the use of CTs summarizes this part of the city in only one area presenting a high percentage of old population and moderate value of crash counts, which can barely alter model estimations. In conclusion, the use of covariates that depend on possibly sparse population should be used with special care, as the choice of the wrong BSU type in such cases could lead to artefactual associations between the covariate and crash rates observed.

The differences in local parameter estimates for the GWR models across the four BSU types is rather evident (Figure 5.10). Several covariates presented a controversial behaviour with a strong dependence to the BSU system. These discrepancies are more obvious than with the CAR models given the higher number of covariates that are highlighted with the specified percentile criteria. Leaving apart the AG100 aggregation level because it has overall produced more disparate results (reducing its reliability), there were still some covariates showing inconsistent associations with crash counts, confirming the consequences of MAUP in this case study. Furthermore, a global high level of coincidence between two distributions of local parameter estimates derived from GWR (at two different scales and/or zonings) does not guarantee, at all, that the local estimates vary similarly across space, which is clear in view of Figure 5.11.

5.4.4 Model performance comparisons

Table 5.9 provides information with regard to model fitting for all the BSU systems and aggregation levels employed in the study. It is appreciable that CAR models performance improved gradually (decrease in DIC) as the aggregation increased (reaching the minimums at AG100 for all the BSU systems). Although this is an issue already pointed out by Fotheringham and Wong (1991), that does not always represent a real improvement in model quality and interpretation, in this case may be also the consequence of a weak multicollinearity among the covariates at AG100

	AG100		AG200		AG300	
Covariate	**Est.**	**SD**	**Est.**	**SD**	**Est.**	**SD**
Intercept	-6.7825*	1.5880	-6.8832*	0.9019	-6.2381*	0.6777
EDU	-0.0399	0.0281	-0.0633*	0.0330	-0.0580*	0.0298
BAR	0.0028	0.0028	0.0022	0.0034	0.0039	0.0035
COMP	0.0007	0.0007	0.0012	0.0009	-0.0004	0.0009
HOT	-0.0002	0.0005	0.0002	0.0006	0.0003	0.0005
PARK	-0.0133	0.0632	-0.0286	0.0724	-0.0143	0.0638
BUS	-0.0182	0.0128	-0.0123	0.0167	0.0002	0.0161
TRAM	-0.0962	0.0754	-0.2074*	0.1079	-0.1248	0.1088
BETW	-0.0001	0.0001	0.0000	0.0001	0.0000	0.0000
INT	-0.0126	0.0178	-0.0052	0.0125	-0.0126	0.0085
YP	-0.0452	0.0485	0.0329*	0.0193	0.0226	0.0142
OP	0.0207	0.0145	0.0138*	0.0074	0.0010	0.0043
HP	0.2806*	0.1284	0.2234*	0.0604	0.1994*	0.0471
ψ	3.2962*	0.4602	2.3784*	0.2386	7.0513*	2.1572

Table 5.6: Estimates with standard deviation (SD) for the parameters involved in the CAR model, considering TIs as BSUs. An asterisk (*) indicates the significance of a parameter with a 90% credibility

(specially for HEXAs).

On the other hand, for any fixed scale with the exception of AG100, HEXAs appeared as the optimal choice for the CAR models. Hence, the use of hexagonal grids would be a reasonable recommendation, although care must be taken with areas of low population density if population-related covariates are being used, as shown in the previous subsection. For the latter, CTs or other administrative division should be more convenient.

One positive conclusion is the substantial level of agreement shown by the CAR and GWR models for each BSU system and level of aggregation, as it can be observed from the comparison of Figure 5.5 and Figure 5.10. However, GWR models showed superior values for MAD, SAD and PMAD for most of the combinations of BSU types and aggregation levels, an opposite result to that found by other authors (Xu and Huang, 2015; Amoh-Gyimah et al., 2017). The use of semiparametric GWR models or an adaptive version of their kernel's bandwidth may have led to more close performance results for some combinations of scale and zoning, but this possibility was discarded to guarantee a fair comparison between models in the context of this analysis, which is more focused on model parameter estimations rather than on model performances.

5.4.5 Parameter interpretations

Despite MAUP effects, several associations between crash counts and some of the covariates were observed at several combinations of aggregation level and BSU type, deserving a deeper analysis. In particular, the number of undergraduate educational

	AG100		AG200		AG300		AG400		AG500	
Covariate	**Est.**	**SD**	**Est.**	**SD**	**Est.**	**SD**	**Est.**	**SD**	**Est.**	**SD**
Intercept	-8.8001*	1.3441	-4.8880*	0.7877	-5.2212*	0.6348	-4.3829*	0.4955	-5.1725*	0.3578
EDU	-0.0457	0.0396	-0.0470	0.0299	-0.0759*	0.0311	-0.0814*	0.0339	-0.0435	0.0334
BAR	0.0001	0.0035	-0.0026	0.0030	-0.0010	0.0037	-0.0008	0.0038	0.0012	0.0036
COMP	-0.0011	0.0010	-0.0002	0.0009	-0.0001	0.0009	0.0011	0.0011	0.0011	0.0010
HOT	0.0004	0.0006	0.0004	0.0007	0.0002	0.0006	0.0006	0.0006	0.0005	0.0006
PARK	0.1677	0.1075	0.0371	0.0781	0.0741	0.0796	0.0406	0.0805	0.0068	0.0766
BUS	0.0267	0.0231	0.0816*	0.0204	0.0902*	0.0212	0.0934*	0.0229	0.1336*	0.0222
TRAM	-0.1348	0.1352	0.1095	0.1633	-0.0518	0.1356	-0.0057	0.1233	0.0244	0.1290
BETW	0.0001	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0000
INT	0.0908*	0.0409	0.0151	0.0199	0.0097	0.0153	0.0021	0.0102	-0.0012	0.0090
YP	0.0455	0.0397	0.0036	0.0261	0.0079	0.0126	0.0175	0.0121	0.0021	0.0095
OP	-0.0199*	0.0103	-0.0127*	0.0075	-0.0096*	0.0056	-0.0038	0.0047	-0.0033	0.0037
HP	0.3438*	0.0971	0.0395	0.0520	0.0703	0.0464	-0.0055	0.0347	0.0480*	0.0274
ψ	1.4053*	0.1862	1133.6054*	9922.7680	205.1350*	579.7774	79.1317*	100.2495	103.7747*	162.6996

Table 5.7: Estimates with standard deviation (SD) for the parameters involved in the CAR model, considering HEXAs as BSUs. An asterisk (*) indicates the significance of a parameter with a 90% credibility

	EDU	**BAR**	**COMP**	**HOT**	**PARK**	**BUS**	**TRAM**
NNI	0.86	0.51	0.26	0.35	1.08	0.70	1.30
***p*-value**	0.00*	0.00*	0.00*	0.00*	0.15	0.00*	0.00*

Table 5.8: Nearest neighbour indexes (NNI) and p-value associated with each index. An asterisk (*) indicates the statistical significance of the index ($p < 0.05$), which indicates clustering (NNI<1) or dispersion (NNI>1)

centers showed a consistent negative correlation with crash counts, whereas average horsepower of the cars in the BSU generally associated with more traffic crashes. Other covariates, such as the number of bus stops in the BSU, the average betweenness of its road segments or the percentage of population with 65 or more years living in the BSU also suggested the presence of a positive relationship, but uncertainties from MAUP issues were stronger for them. The knowledge of the city being analyzed arouses the suspicion that some of these correlations could be related to a hidden (not included in the models) factor as it is the distance to the city center of Valencia (case of average horsepower and old population), but this question would require a specific research.

5.5 Conclusions

This study is, to the best of my knowledge, the first one that provides a simultaneous investigation of scale and zoning effects regarding the modifiable areal unit problem in the context of traffic safety analysis. Furthermore, another capital objective was to specifically assess how a change in the aggregation level or BSU type may affect the basic characteristics of both the response variable being considered (crash counts) and the set of covariates included in the models. The consequences of MAUP for the data analyzed were notorious from the perspective of both scale and (specially) zoning alterations. Some of the effects of MAUP were understandable from visual inspection of the data, as shown through some exemplifications, but it is really tough sometimes to explain certain model parameter disagreements that arise from

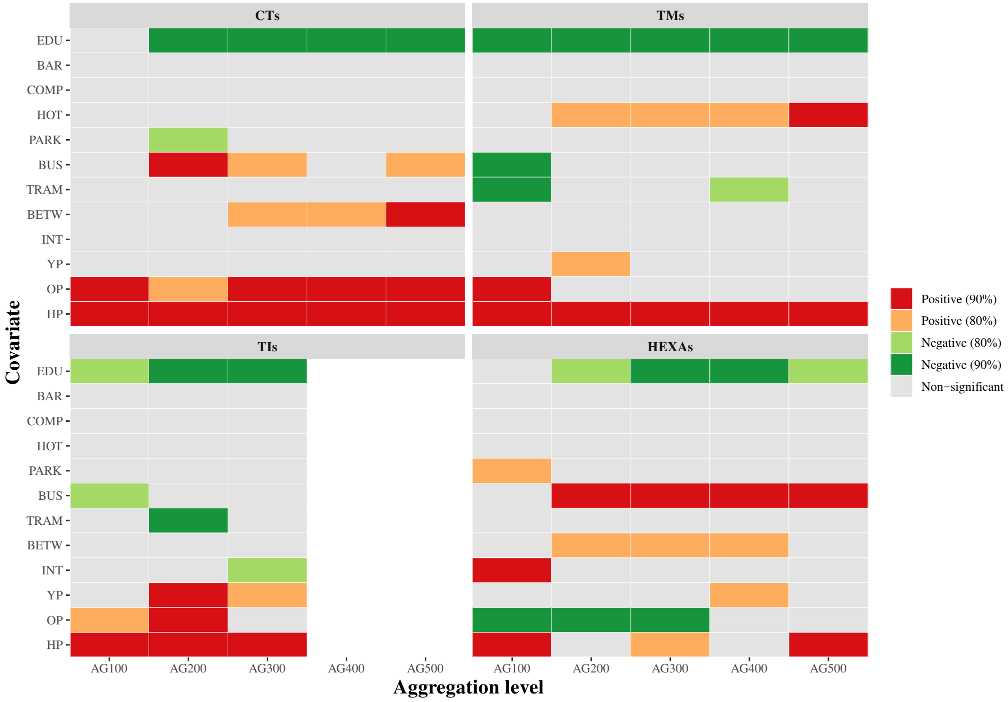


Figure 5.5: Summary of the results obtained for the CAR models, considering the four types of BSUs and the levels of aggregation that were applied

a change in aggregation level or BSU configuration.

The comparative analysis yielded that CAR models using hexagonal gridded units (HEXAs) were the best choice according to the performance measurements adopted. The employment of BSU types based on the road network being analyzed (TMs and TIs), which naturally avoided boundary effects (although HEXAs even outperformed them in this aspect), did not lead to better model performances. Anyhow, model performance measures should not be the only instrument to select one combination of scale and zoning over others. Indeed, the use of CAR models and HEXAs also unveiled controversial behaviours for some model parameter estimates. These were found to be a consequence of the fact of using a BSU type (hexagonal unit) that may not be the best one to represent demographic characteristics of the area of investigation. Thus, even though the results for census tracts were more modest in terms of model performance, this kind of administrative unit is possibly the most suitable one to seek more robust conclusions if several demographic covariates are present.

The analysis of the MAUP presented in this study has also emphasized how the changes in scale or zoning alter the typology of the response and predictor variables that are eventually provided as the input to a statistical model. Specifically, higher aggregation levels associated with a reduction in overdispersion and kurtosis. This result suggests that the choice of a modelling approach once a change in scale or zoning has been produced should be well addressed, implying the reconsideration or even rejection of a previously selected approach. Furthermore, higher levels of

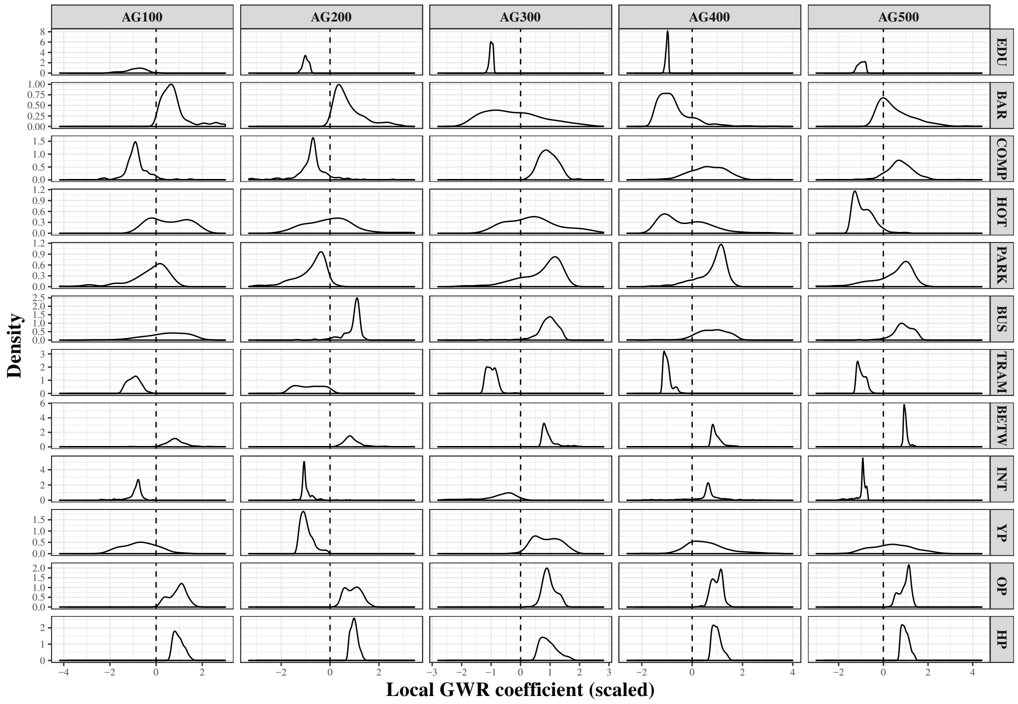


Figure 5.6: Combined graph showing the distributions of local parameter estimates, for the covariates used in the GWR models (in rows) and each level of spatial aggregation (in columns) tested for the CTs

spatial aggregation yielded an overall increase of variance inflation factors, a sign of multicollinearity risk that leads to the conclusion that an excessive aggregation of the data should be avoided, or at least properly checked. The spatial nature of some of the covariates has also provided some clues on their sensitivity towards MAUP. Indeed, covariates having low levels of spatial autocorrelation or generated from point patterns not extremely clustered have displayed a more coherent behaviour among scales and zonings. However, this issue requires a deeper investigation.

Moreover, some limitations of this study deserve some comment. First, it is worth noting that selecting a proper exposure measure is essential to avoid bias in parameter estimations. Indeed, the lack of consideration of a exposure measure could have a greater impact on statistical estimations than a variation in scale or zoning. Due to the unavailability of traffic volume, non-pedestrian road length was used as a proxy for exposure. Second, the choice of certain homogeneity criteria to carry out the regionalization process may be another factor, other than scale and zoning, that affects model fitting. In this analysis, homogeneity criteria were solely based on crash counts. Other factors, such as land use and socio-economic characteristics, should also be considered in future studies.

To conclude, this study has provided more evidence regarding the complications that the MAUP can create in the context of a spatial traffic safety analysis. The performance of sensitivity analyses (testing different BSUs at various levels of scale and zoning) suggested by Xu et al. (2018) considering model estimates for several

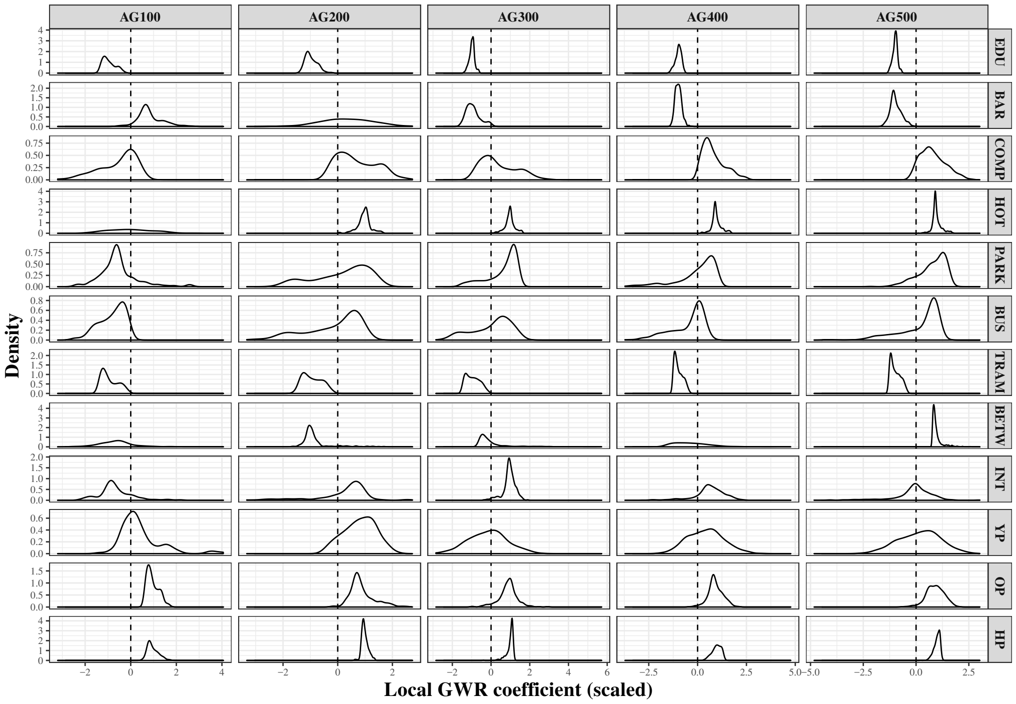


Figure 5.7: Combined graph showing the distributions of local parameter estimates, for the covariates used in the GWR models (in rows) and each level of spatial aggregation (in columns) tested for the TMs

scales or zonings (or both) seems unavoidable, but this kind of analysis should also include the investigation of the "intermediate" factors that affect statistical inference such as the modelling approach, the multicollinearity shown by the covariates and their spatial autocorrelation. The consideration of all of these factors should help researchers to achieve firmer conclusions, although one cannot forget that it is still likely that the MAUP will never be solved (Manley, 2014).

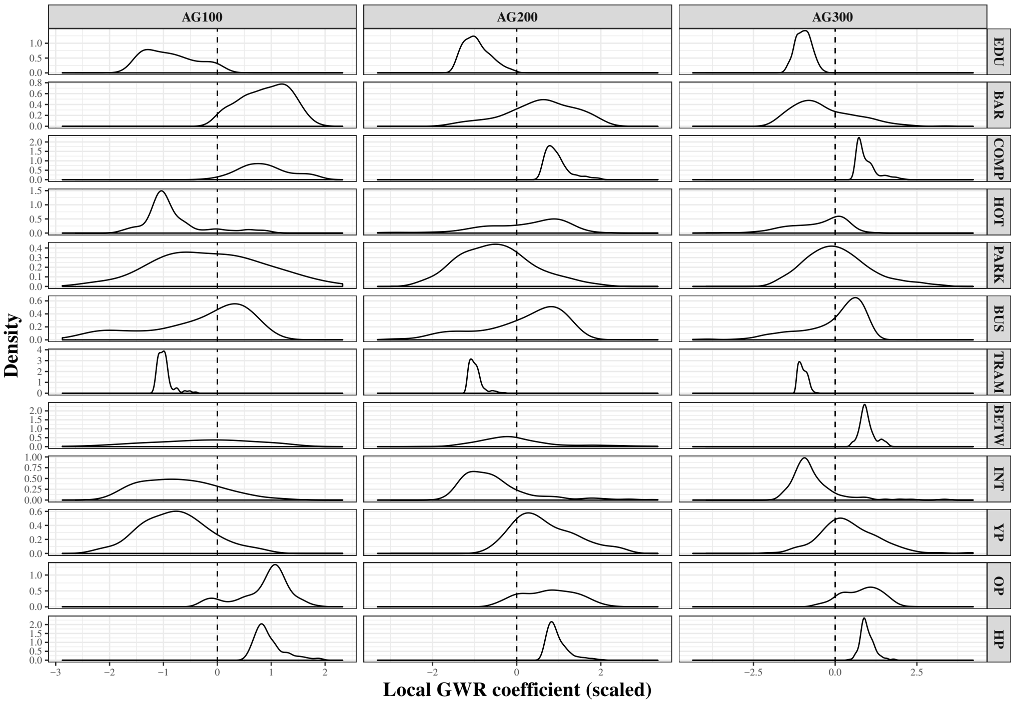


Figure 5.8: Combined graph showing the distributions of local parameter estimates, for the covariates used in the GWR models (in rows) and each level of spatial aggregation (in columns) tested for the TIs

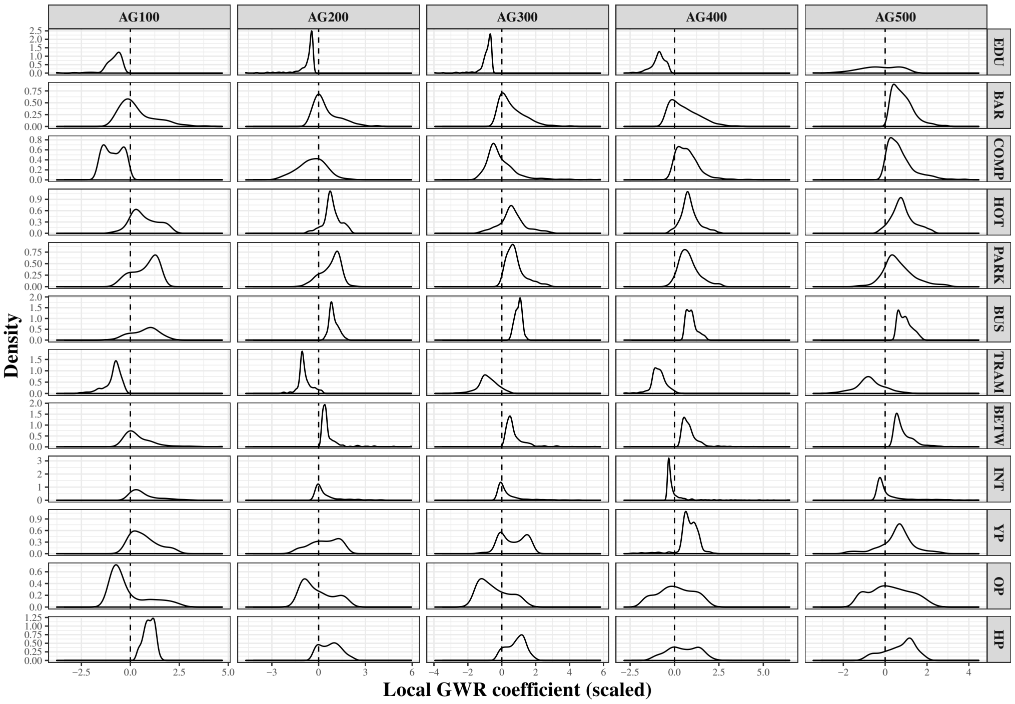


Figure 5.9: Combined graph showing the distributions of local parameter estimates, for the covariates used in the GWR models (in rows) and each level of spatial aggregation (in columns) tested for the HEXAs

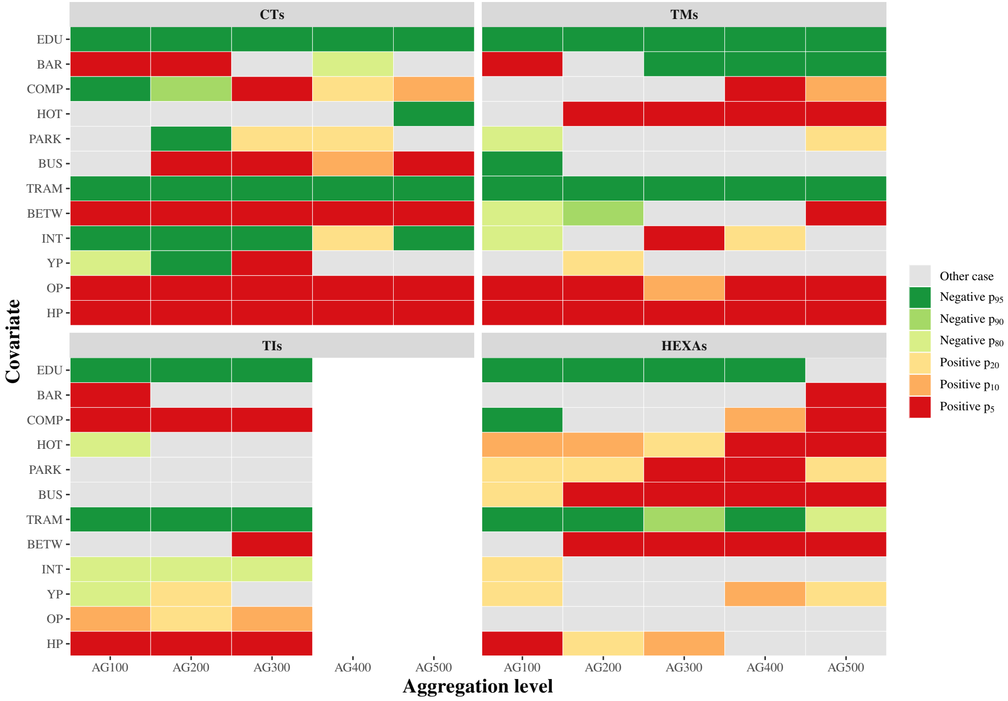
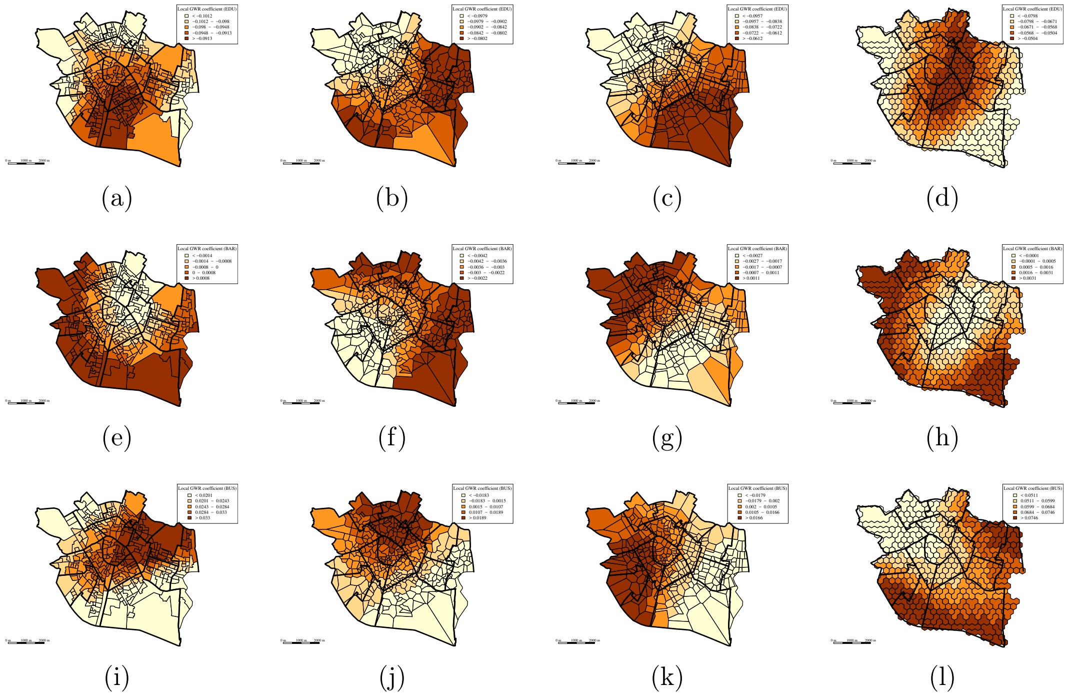


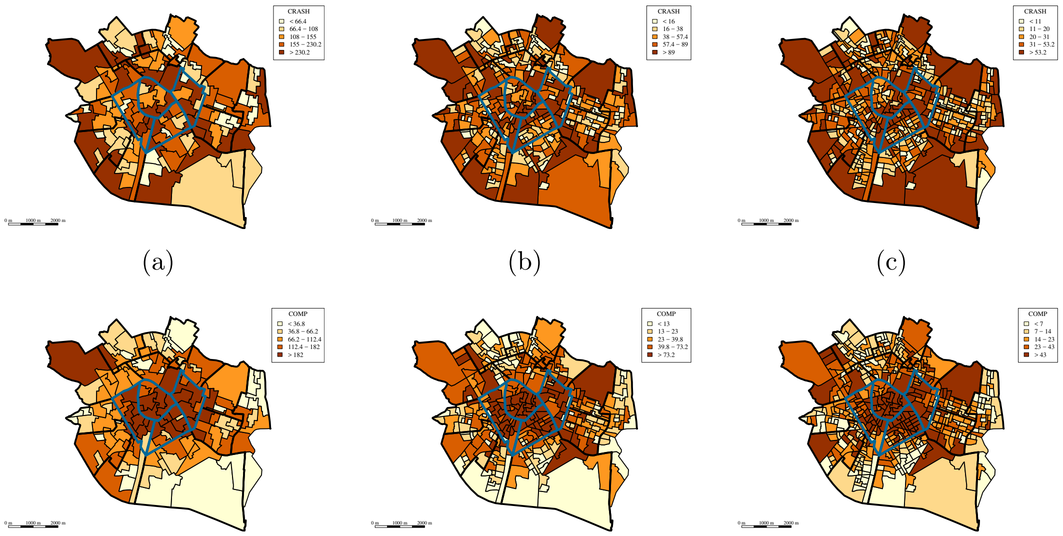
Figure 5.10: Summary of the results obtained for the GWR models, considering the four types of BSUs and the levels of aggregation that were applied


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

(e) (f) (g) (h)

(i) (j) (k) (l)

Figure 5.11: Estimated local parameters for the GWR model considering CTs, TMs, TIs and HEXAs for EDU (a-d), BAR (e-h) and BUS (i-l) at AG300. Districts of Valencia are overlayed (thicker lines, in black) for better readability and comparison



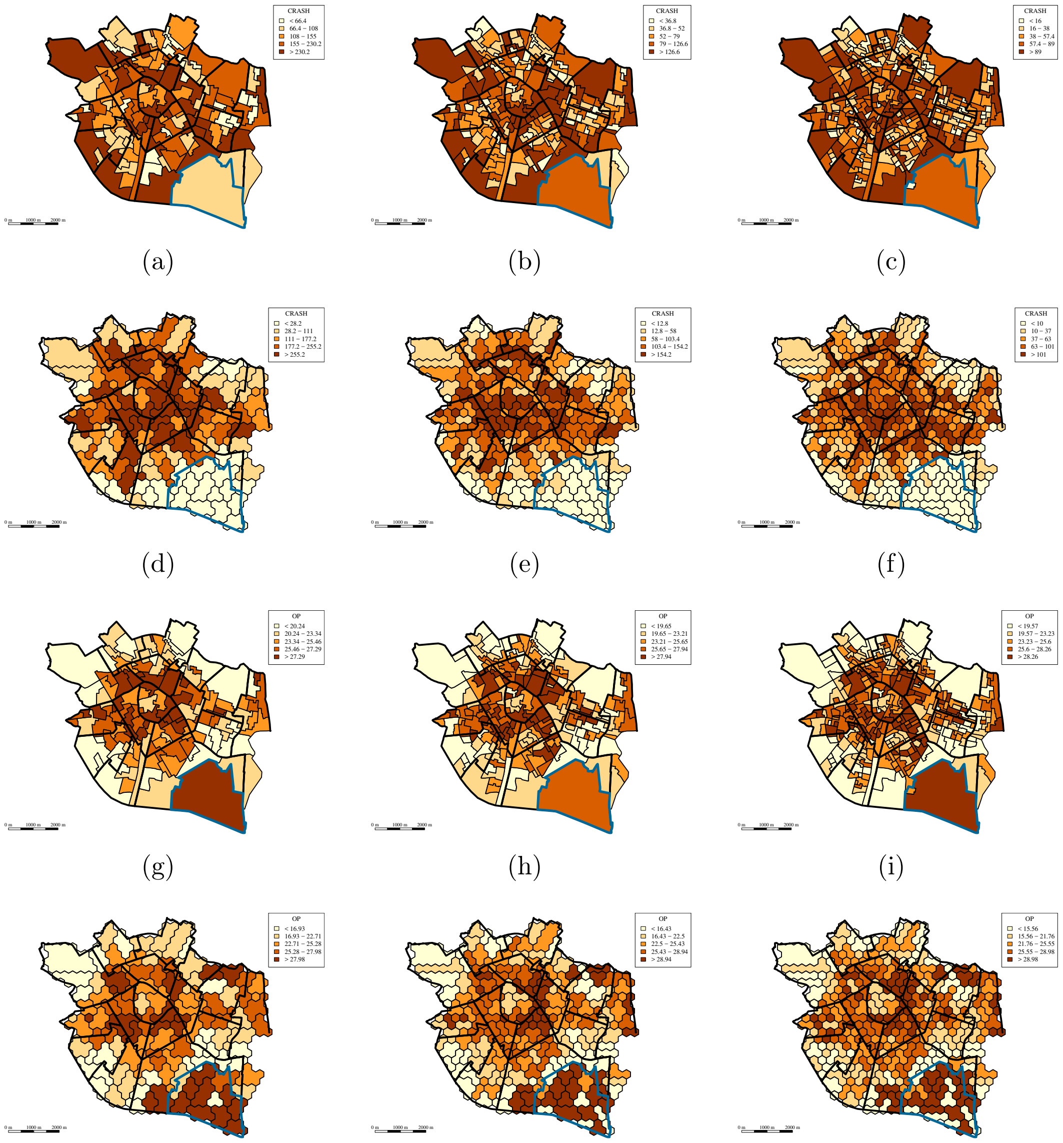
(a) (b) (c)

(d) (e) (f)

Figure 5.12: Crash counts (a-c) and COMP values (d-f) for CTs at AG100, AG300 and AG500 (in order of appearance at each row, from more to less aggregated). Some central Districts of Valencia are highlighted in blue

AG	BSU	CAR				GWR			
		DIC	MAD	SAD	PMAD	AIC	MAD	SAD	PMAD
AG100	CT	1317.37	60.55	6963.82	38.77	1348.66	54.29	6243.25	35.21
AG100	TM	1311.01	62.20	7029.14	39.13	1276.00	51.04	5767.66	32.53
AG100	TI	1243.94	72.99	7737.23	43.07	1241.80	55.19	5849.80	32.99
AG100	HEXA	1306.55	102.84	11517.79	64.95	1330.17	56.50	6327.53	35.68
AG200	CT	1920.69	18.34	3667.84	20.42	2185.88	37.25	7449.06	42.01
AG200	TM	1881.92	10.35	2101.65	11.70	2156.57	37.89	7691.15	43.37
AG200	TI	2116.80	42.06	8411.20	46.83	2169.72	34.07	6813.67	38.43
AG200	HEXA	1586.44	2.56	512.40	2.89	2164.42	37.73	7546.31	42.56
AG300	CT	2694.35	14.33	4299.53	23.94	3073.02	29.43	8828.77	49.79
AG300	TM	2906.43	22.33	6698.53	37.29	3017.90	31.04	9313.18	52.52
AG300	TI	2739.21	11.64	3490.79	19.43	3068.26	26.49	7947.48	44.82
AG300	HEXA	2311.54	3.57	1071.61	6.04	3050.20	27.52	8256.20	46.56
AG400	CT	3455.81	11.98	4792.86	26.68	3918.14	23.66	9463.93	53.37
AG400	TM	3700.36	20.20	8078.93	44.98	3868.05	25.05	10021.12	56.51
AG400	HEXA	2981.30	2.99	1194.31	6.74	3903.94	22.20	8879.82	50.08
AG500	CT	4103.36	10.73	5367.13	29.88	4772.90	19.95	9976.61	56.26
AG500	TM	4386.43	15.87	7934.65	44.17	4724.74	21.20	10597.96	59.77
AG500	HEXA	3275.34	1.99	997.13	5.62	4768.70	17.18	8590.50	48.45

Table 5.9: Models performance in terms of DIC and AIC (for CAR and GWR models, respectively), MAD, SAD and PMAD for all the BSU configurations and levels of aggregation tested



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Figure 5.13: Crash counts for CTs (a-c) and HEXAs (d-f) and OP values for CTs (g-i) and HEXAs (j-l) at AG100, AG200 and AG300 (in order of appearance at each row, from more to less aggregated). A CT in the South of Valencia is highlighted in blue

Chapter 6

Modelling risk in relation to multiple sources: The effect of school locations on traffic accidents

The aim of this Chapter is to present the use of several statistical techniques in order to estimate the effects that school locations and commuting to school may have on the incidence of traffic accidents in specific time windows. These techniques are applied to a dataset of traffic accidents recorded over two years in the city of Valencia (Spain). The objectives are both methodological and practical, as will be highlighted within the text.

6.1 Introduction

Guaranteeing safety near school locations is a fundamental objective for the experts in charge of traffic management, especially in order to protect children from traffic accidents. In Spain, high densities of vehicles are usually observed around school locations at starting and ending times on school days. This is a factor that certainly conditions traffic flow and dangerousness at these times, but few scientific research studies are available on this issue in Spain. Hence, the following paragraphs include a description of several studies conducted in the last fifteen years on the subject of traffic safety around schools, the factors involved and the implications. An emphasis is placed on the methodological approaches chosen for these studies.

Most of them have focused on areal zones centered around school locations (buffer zones). First, Abdel-Aty et al. (2007) employed a log-linear model and a set of categorical variables including driver characteristics, pedestrian/cyclist characteristics and other characteristics related to traffic, road and vehicle typologies. These authors showed that most of the traffic accidents involving school-aged children took place close to school locations, considering a buffer zone of 0.5 miles from each school. More specifically, they found higher accident rates for middle and high school children, which was associated with the fact that these schools are frequently situated in the vicinity of a multi-lane high-speed road. Clifton and Kreamer-Fults (2007) modelled five types of aggregated dependent measures (two related to accident occurrence and three to pedestrian exposure) through linear regression on the basis of 0.25 mile buffer zones around schools. Among the set of covariates considered by these authors, percentage of nonwhite residents and population density were associated with more traffic accidents, whereas transit access (percentage of households within 0.25 miles of a transit stop) correlated negatively with accident counts. Warsh et al. (2009) created buffer zones with a radius ranging from 150 to 450 m around schools and confirmed the higher proportion of child pedestrian-vehicular accidents at 7.00 - 10.00 and 15.00 - 17.00. They also observed a decrease in the risk for older students, specially those in the 15-17 age group. Yu and Zhu (2016) assessed the presence of modifiable areal unit problems (MAUP) (Amoh-Gyimah et al., 2017; Zhai et al., 2018b; Xu et al., 2018) in the specific context of school safety by defining 0.5, 1, 1.5, and 2 mile buffer zones around each school. It was found that highways and interstates, traffic-generating land uses and transit stops were associated with more traffic accidents. On the other hand, higher sidewalk coverage and local roads around schools correlated with fewer traffic accidents.

The use of buffer zones around school locations can be combined with a case-control study design. Indeed, Rothman et al. (2017b) performed a case-control study in Toronto (Canada) by distinguishing school attendance boundaries belonging to the highest quartile of pedestrian-vehicle accident rates (cases) from those in the lowest (controls). It was found through a multivariate logistic regression that some factors such as one way streets, crossing guards, traffic signal density and social disadvantaged areas were associated with a higher incidence of traffic accidents. Moreover, Rothman et al. (2017a) focused their research on traffic safety near schools around risky drop-off behaviours following the same case-control strategy as Rothman et al. (2017b). Observational covariates related to risky behaviours by both drivers and pedestrians in school proximities were obtained, and these were then investigated in combination with a collection of environmental covariates through logistic regressions. Several important findings for safety planning were found, including, for instance, the association between traffic congestion and risky driving and walking behaviours.

Furthermore, in the last few years there has been an increasing number of studies approaching the research from a road segment level perspective. For example, Hwang et al. (2017) considered street segments at a 0.25 mile distance from school locations and a buffer distance of 100 ft around these street segments as the focus of their analysis. Accident counts at the road segment level were recoded into a binary outcome and logistic regression was used for modelling purposes. These authors found a positive correlation between accident rates and block length, proportion of missing sidewalks, crosswalk density and commercial land use. Furthermore, the study revealed some factors specifically affecting students from disadvantaged neighbourhoods. Park et al. (2018) compared negative binomial and Poisson inverse Gaussian model approaches and found that the latter provided better results in terms of forecasting accuracy. They worked at the road segment level, considering roadways connected to a school building or to a nearby area over which school-related activities were taking place. Finally, Yu (2015) made use of a hierarchical logistic model at two spatial levels: neighbourhood and road segment. The findings were similar to those of Yu and Zhu (2016).

The Chapter is structured as follows. First, the data used in the analysis is described, including the dataset of traffic accidents, the collection of schools and typologies available in the city, and the set of covariates considered with explanatory objectives. This is followed by a methodological section containing an exploratory analysis that helps to determine the time windows that may be affected by school-related trips, and the explanation of four different statistical methods (observed vs expected ratios, spatial count models, case-control logistic regression and multiple source regression) that allow us to investigate the causal relationship between school locations and traffic accidents occurring during the time windows predefined through the exploratory analysis just mentioned. The results derived from each of the methods are then discussed and compared.

6.2 Data

6.2.1 Accident dataset

A total of 18037 traffic accidents recorded by the Local Police of Valencia (Spain) during the years 2014 and 2015 were used for the analysis. Geographical coordinates for the accidents and information about the date and time of occurrence were provided by the Local Police. The accidents were located with precision on the road network of the city. This road network has a total length of 840.3 km (with a diameter of nearly 11.6 km) and contains 6110 road intersections. Information regarding the type of traffic accident (vehicle-vehicle, vehicle-pedestrian, etc.), its severity (severe, non-severe, etc.) and the age of the people involved was unavailable, although no reporting bias in favour of a specific type of accident should be present.

6.2.2 School dataset

A total of 372 schools of various age levels located in Valencia were considered for the analysis (Figure 6.1). Four main levels of education in Spain can be distinguished (with approximate ages): Preschool I (0-3 years), Preschool II (4-5 years), Primary (6-11 years) and Secondary (12-17 years). For research purposes, the schools were classified into four categories, according to the educational levels offered: All-level (81 schools), Preschool (178 schools, which include centers that offer only Preschool I or II or both), Primary (81 schools) and Secondary (32 schools). Hence, one of the objectives was also to assess whether differential associations with traffic accidents arise depending on the school classification.

6.2.3 Covariate definition

Several covariates were considered in order to control for baseline effects that may be responsible for the higher incidence of traffic accidents near schools. These covariates were classified into three categories: traffic-related, environmental and socioeconomic/demographic. The values for these covariates were always computed over basic spatial units (BSU) of analysis, which varied depending on the analysis being performed. Table 6.1 contains a description of all the covariates treated during the analysis, which were conveniently standardized in order to facilitate parameter interpretation and comparison. In particular, the following lines describe two of the covariates included: betweenness and land use entropy.

Betweenness (BETW) is a measure of network connectivity that was used as a proxy for average vehicle miles travelled per road segment, which was already defined on Section 5.2.2. Land use entropy (LUE) was defined as in Rothman et al. (2017b), following the next expression:

$$\mathrm{LUE}(\mathrm{BSU}_i) = -\frac{\sum_j p_{ij}\log(p_{ij})}{\log(n)}$$

where j iterates over the indexes associated with the land uses that are present at BSU i (otherwise, the logarithmic expression is not computable), p_{ij} is the proportion of land use of type j at BSU i and n is the total number of land uses considered by the Land Occupancy Information System (SIOSE, by its acronym in Spanish). The LUE index lies in the [0,1] interval, with a value closer to 1 indicating higher land diversity and 0 meaning a unique land use.

Type	**Variable**
Accidents	No. of traffic accidents
Exposure	Non-pedestrian road length
School-related	No. of schools (all types) Distance to the closest All-level school Distance to the closest Preschool school Distance to the closest Primary school Distance to the closest Secondary school
Environmental	No. of education-related services per road km No. of services from various sectors (non-educational) per road km % of road length with parking spaces available No. of parking zones per road km No. of bus stops per road km Land use entropy
Traffic-related	Average betweenness per road segment Complex intersections (four-or-more-leg) per road km Main road length per road km Traffic lights per road km
Socioeconomic and demographic	No. of school-aged residents (0-18 years) per road km Percentage of high-end cars

Table 6.1: Description and classification of the covariates defined for the analysis

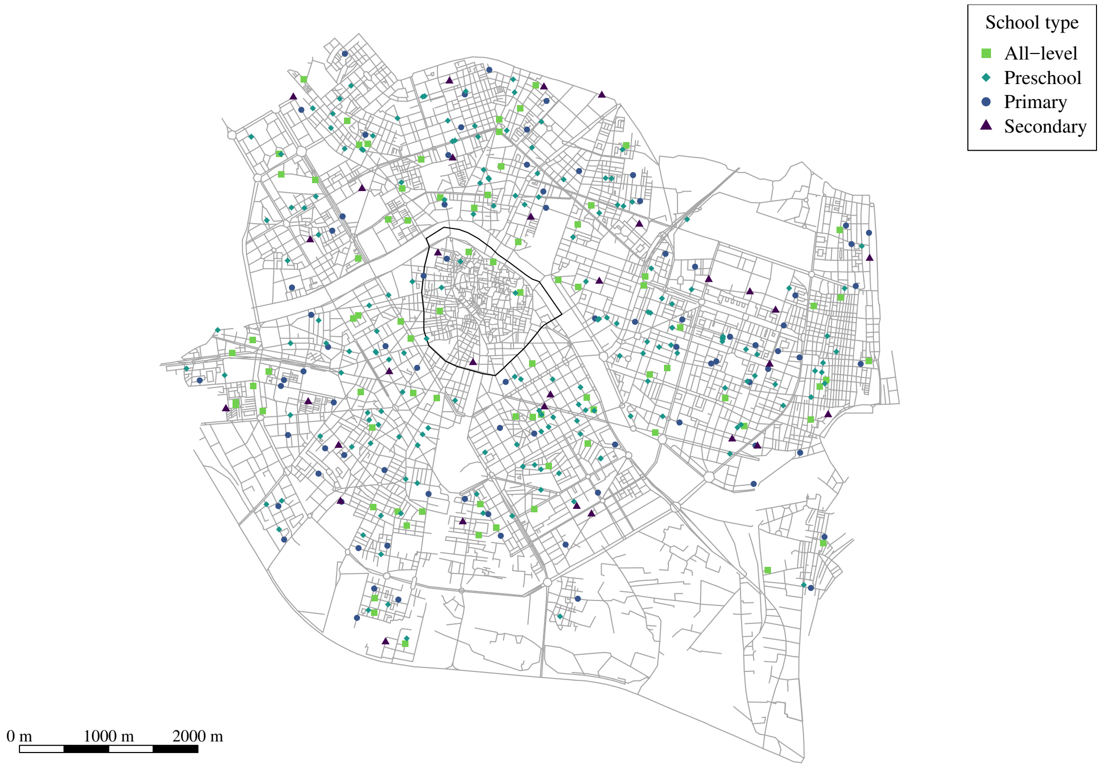


Figure 6.1: School locations in the city of Valencia distinguished by the academic level they offer. The central district of Valencia (city center) is highlighted with a thicker black line

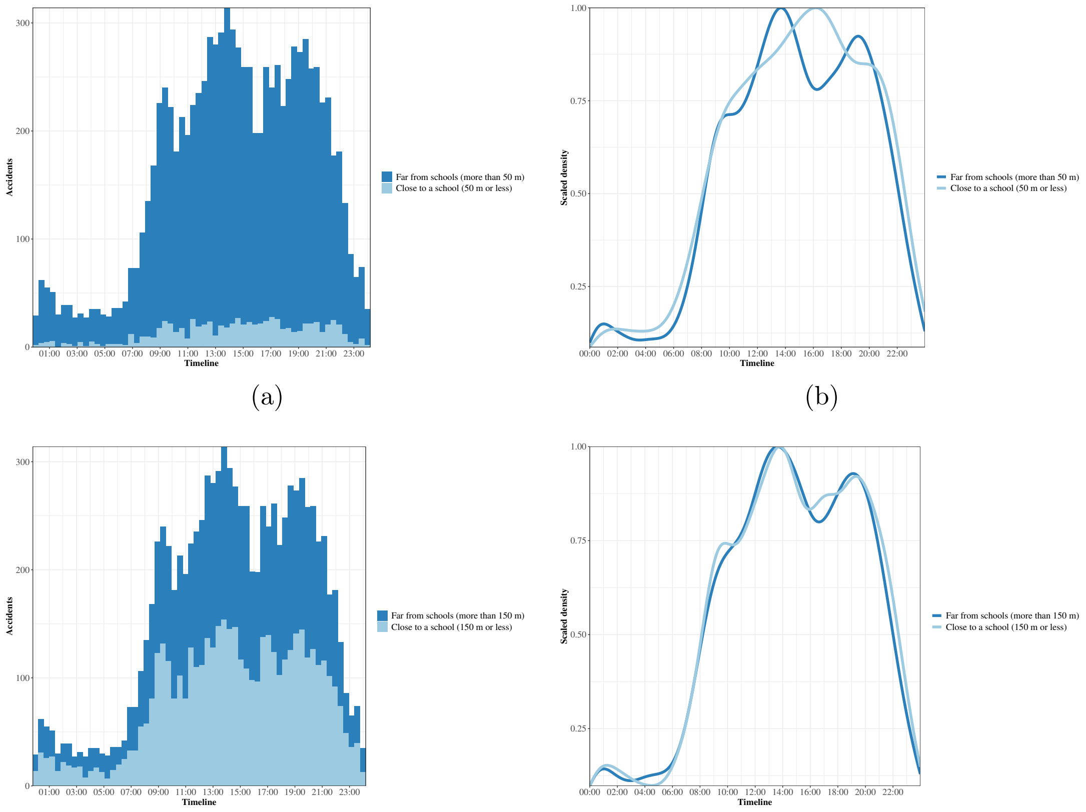
6.3 Methodology

6.3.1 Software

The R programming language (3.5.2 version, R Development Core Team, Vienna, Austria) (R Core Team, 2018) was used to obtain all the results presented in this work. The R packages DEoptim (Mullen et al., 2011), ggplot2 (Wickham, 2016), INLA (Rue et al., 2009; Martins et al., 2013; Lindgren and Rue, 2015), rgeos (Bivand and Rundel, 2018a), spatstat (Baddeley et al., 2015) and spded (Bivand and Piras, 2015) were specifically required to perform the complete analysis.

6.3.2 Time window of analysis

A preliminary question was to determine the hours within the day that may be affected by traffic dynamics generated as a consequence of arrivals at (or departures from) schools. School starting and ending times should be the axes of such a temporal interval, but this is hard to define because most schools in Valencia are free to determine their own schedules (subject to some common restrictions). Furthermore, it is quite normal for students in Spain to attend extracurricular activities during weekday evenings (immediately after school), which are usually carried out in the school facilities or in the surrounding area (extracurricular academies, institutions



(a) (b)

(c) (d)

Figure 6.2: Histograms and scaled densities along the timeline of school days (in 2014-2015) shown by traffic accidents that occurred close to a school and far from schools for a threshold distance of 50 m (a)-(b) and 150 m (c)-(d)

or centers are frequently located in the vicinity of a school).

Exploratory analyses were performed (Figure 6.2) and two time windows of interest were finally established: the one around school starting times and the period in the afternoon and evening that includes school ending times and the subsequent hours. Hence, the two time windows 7:30 - 9:30 and 15:00 - 19:00 were selected for further analysis. In view of Figure 6.2, it seems that school locations do not have an strong effect on the incidence of traffic accidents around starting times, but this period of the day was maintained for further investigation. The choice of the interval 7:30 - 9:30 was based on the fact that most schools in Valencia start their classes in the 8:00 - 9:00 window. In contrast, the period 15:00 - 19:00 shows a clear differential effect in terms of density of accidents close to and further away from schools, especially for short threshold distances such as 50 m, for example (Figure 6.2b). From now on, the time window 7:30 - 9:30 will be referred to as the Starting Time Window (STW) and 15:00 - 19:00 will be denoted as the Evening Time Window (ETW).

6.3.3 Observed vs. expected ratios

The first statistical analysis to identify the association between traffic accidents and school locations consisted in computing observed/expected accident ratios at different distance thresholds, σ, from all schools in the city. The sequence of values selected for σ were 25, 50, 75, 100, 150, 200, 250 and 300 m, which allowed analyzing the effect of interest at various spatial scales.

A Monte Carlo approach was taken to assess the statistical significance of the ratios considering the full set of schools and each school type separately for the two time windows established. This process consists in generating 999 datasets that preserve the locations observed for all the accidents recorded in 2014-2015, while permutating their corresponding dates and times of occurrence. Hence, the number of traffic accidents that lie below the threshold distance (σ) from a school location (for both STW and ETW) is kept for each simulation. The average of the 999 simulated values represents an expected value, which is then compared with the real number observed, providing an observed/expected ratio. The 2.5th, 5th, 95th and 97.5th percentiles of the set of simulated distributions of counts are kept to make it possible to assess a significance level for each observed/expected ratio (for a given σ).

6.3.4 Macroscopic modelling

A conditional autoregressive (CAR) model with negative binomial (NB) response was chosen to fit the accident counts recorded on school days in the STW or ETW (a total of 800 and 2884, respectively, for the period 2014-2015) over a hexagonal grid of 198 BSUs of side length slightly over 250 m and an area of around 0.175 km^2. The use of a hexagonal grid was preferred over the employment of administrative division units (such as boroughs or census tracts) because it provides the possibility of defining a spatial unit of analysis of intermediate scale. This makes it possible to perform an accurate analysis while keeping a good balance between the number of spatial units and the number of covariates being considered. Hexagonal units containing a minimal road structure (mostly located within a green area or in semirural zones along the periphery of the city) were removed from the grid in order to avoid a possible distortion of the results.

If $Y \sim \mathrm{NB}(\mu, \psi)$ (NB distribution of mean μ and shape ψ) then it holds that $E(Y) = \mu$, $V(Y) = \mu + \frac{\mu^2}{\psi}$ and $P(Y = x) = (\frac{x+\psi-1}{\psi-1})(\frac{\psi}{\mu+\psi})^{\psi}(\frac{\mu}{\mu+\psi})^{x}$. Then, assuming a NB distribution for the accident counts, the following spatial model was implemented:

$$\begin{aligned} Y_i &\sim \mathrm{NB}(\mu_i, \psi) \\ \log(\mu_i) &= \log(E_i) + \lambda_0 + \sum_{m=1}^{p} \lambda_m X_{im} + \phi_i \end{aligned} \tag{6.1}$$

where Y_i is the number of accidents observed at hexagonal unit i, μ_i and ψ are, respectively, the mean risk (for hexagonal unit i) and overdispersion (shape) values for the NB distribution, the natural logarithm acts as a link function for μ_i, E_i (exposure at hexagonal unit i) is the length of non-pedestrian road at hexagonal unit i (offset of the equation), X_{im} represents the value of the m-th covariate at hexagonal unit i, λ_m is the coefficient that controls the effect of the m-th covariate and ϕ_i represents a spatial effect for hexagonal unit i, which is derived from the neighbourhood structure formed by the hexagons that are part of the grid.

The spatial effect was modelled through the well-known CAR structure (Besag, 1974; Besag et al., 1991):

$$\phi_i \mid \phi_j, j \neq i \sim N(\alpha \sum_{j=1}^{n} w_{ij}\phi_j, \tau_i^{-1}) \tag{6.2}$$

where τ_i is a precision parameter that varies with spatial unit i and w_{ij} is an indicator parameter that is 1 if hexagonal units i and j are contiguous and 0 otherwise. The use of a CAR structure for the modelling of accident counts at the macroscopic level is a common practice in traffic safety analysis (Quddus, 2008; Huang et al., 2010).

6.3.5 Risk modelling in relation to several point sources

Diggle and Rowlingson (1994) proposed a class of regression models inspired by previous works focused on estimating disease risk around one hazardous point source. This kind of model has been successfully applied in many epidemiological studies involving pollution sources that negatively affect human health (Ramis et al., 2011; Reeve et al., 2013). Analogously, this approach could be useful to model the risk triggered by the close presence of a particular type of building within the road network, such as a school. As far I know, multiple source regression models have not been used before in the field of traffic safety analysis.

Diggle et al. (1997) adapted the work from Diggle and Rowlingson (1994) to enable this class of models to be used with aggregated data. This approach was chosen for the available dataset, considering the same hexagonal grid that was employed for the CAR modelling of accident counts. Hence, Equation 6.3 displays the mathematical expression of this multiple source regression technique, which accommodates four sources of risk that correspond to the four school types that exist in Valencia.

$$\begin{aligned} Y_i &\sim \text{Po}(\mu_i) \\ \mu_i &= \exp\left(\sum_{m=1}^{p} \lambda_m X_{im}\right) \prod_{j=1}^{4} f(d_{ij}) \\ f(d_{ij}) &= 1 + \alpha_j \exp(-(d_{ij}/\beta_j)^2) \end{aligned} \tag{6.3}$$

where j indicates school type (1 = All-level, 2 = Preschool, 3 = Primary, 4 = Secondary), α_j is the proportional increase (or decrease) in risk produced by school type j, d_{ij} is the distance in km between hexagonal unit i and type j (from the centroid of the hexagon to the closest element of school type j) and β_j measures the rate of decay in risk that occurs as distance from type j increases.

Diggle et al. (1997) approximated the log-likelihood function associated with the model in Equation 6.3 through the next expression:

$$L(\boldsymbol{\Lambda}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = -\sum_i \mu_i + \sum_i O_i \log(\mu_i) \tag{6.4}$$

where $\boldsymbol{\Lambda}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ represent the three vectors of coefficients that include, respectively, the λ_m's, α_j's and β_j's, μ_i follows Equation 6.3 and O_i is the number of accident counts observed in hexagonal unit i. The expression in Equation 6.4 was maximized with the aid of the R package DEoptim (Mullen et al., 2011). The method implemented in DEoptim relies on the theory of differential evolution algorithms for global optimization (Price et al., 2006).

6.3.6 Case-control study

The calculation of observed/expected ratios in buffer zones of various radii from each school location enabled detecting those with higher ratios. Thus, the same Monte Carlo procedure described in Section 3.3 was applied around each of the schools for STW and ETW, separately, to assess statistical significance. Schools with a significantly (at the 0.10 level) high (over 1) observed/expected ratio within a certain buffer zone built around them were considered as cases for establishing of a case-control study. On the other hand, schools presenting a ratio lower than 1 (not necessarily showing statistical significance for such a low ratio) were declared as control schools. Hence, a sample of cases and controls was obtained, constraining buffer zones around school locations to not overlap (avoiding an excess of multicollinearity). A distance of $\sigma = 100$ m was found optimal for defining the buffer zones according to the number of cases and controls provided (choosing this distance allowed having a case:control ratio close to 1:4) and the spatial accuracy achieved. Then, a multivariate logistic regression model was specified:

$$\log\left(\frac{pi}{1-p_i}\right) = \log(E_i) + \sum_{m=1}^{p} \lambda_m X_{im} + \phi_i + \text{Type}_i \tag{6.5}$$

where $\frac{pi}{1-p_i}$ represents the odds ratio for a school being a case. E_i, λ_m, X_{im} and ϕ_i are as in Equation 6.1, but now for the buffer zone associated with school i. Neighbourhood relationships for estimating ϕ_i were now established between schools (either cases or controls) if they were closer than 500 m. For each of the few isolated schools in the city (under a threshold distance of 500 m), the closest school from those available was defined as the unique neighbour. Finally, a factor representing school type was added to the model.

6.4 Results and discussion

This section contains one subsection for each of the statistical approaches chosen for this study. Each part begins by reporting the results obtained for the corresponding approach, indicating some technical issues, and then sets out the interpretations and implications in terms of traffic safety analysis that may be drawn from them.

6.4.1 Risk ratios by school type

The observed vs expected ratios strategy was used with buffer zones of various radii centred at all the school locations available in Valencia. Table 6.2 indicates the estimated risk ratios (observed/expected) that were overall achieved for each school type and for the complete set of schools, along with the statistical significance derived for them from 999 Monte Carlo simulations. No significant associations were found (at the 0.1 level) for the STW, although the estimated ratios are higher for Preschool and Secondary. On the other hand, several associations were determined to be significant for ETW, including the one considering all school typologies for the lowest value of σ. The lower sample of traffic accidents at STW probably reduced the statistical power of the test, and therefore the chances of observing statistically significant ratios.

Therefore, it can be concluded in the light of Table 6.2 that the contribution of school locations to traffic accidents in Valencia is not high in magnitude, although some of the associations found for ETW are quite notable. In this regard, Figure 6.2b already suggested the high presence of traffic accidents within 50 m of school locations from 15:00 to 19:00.

	σ (m)							
STW	25	50	75	100	150	200	250	300
All-level	-	0.92	0.91	0.91	1.10	1.05	1.00	1.00
Preschool	-	1.11	1.10	1.05	1.00	0.99	1.01	1.02
Primary	-	0.87	0.86	0.99	0.97	0.94	0.95	0.97
Secondary	-	0.95	1.23	1.12	1.05	0.97	0.90	0.93
All types	-	1.00	1.01	1.00	1.03	1.01	1.02	1.01
ETW	25	50	75	100	150	200	250	300
All-level	0.97	0.89	1.04	1.03	1.01	1.01	1.00	0.99
Preschool	**1.23**	**1.14**	1.03	1.03	0.99	1.02	1.01	1.00
Primary	1.25	**1.25**	1.10	1.08	1.06	1.03	1.03	**1.03**
Secondary	0.79	0.83	0.84	1.07	0.99	1.03	1.03	1.03
All types	**1.15**	1.05	1.03	1.04	1.00	1.01	1.00	1.00

Table 6.2: Risk ratios for each combination of distance threshold (σ), school type or combination of types and time window (STW or ETW). Risk ratios in bold were found significant at the 0.1 level. The ratios for $\sigma = 25$ m at STW were unreliable due to the small samples available and are not shown

6.4.2 Spatial count models

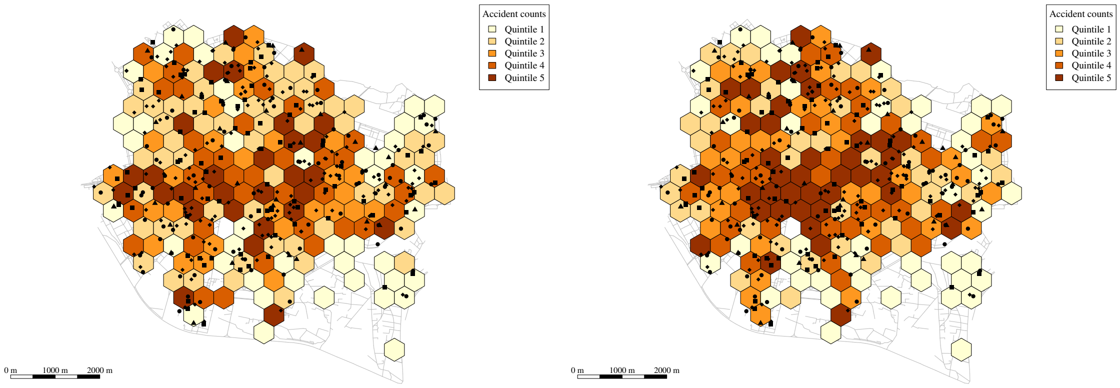
The analysis through CAR models revealed several associations between accident counts at the macroscopic level (hexagonal units) and the set of covariates, with little difference shown by the two time windows considered (see Figure 6.3 for a graphical summary of the data). When the distances to each school type were considered, the differences between STW and ETW increased. Hence, the CAR model for STW revealed a negative association for the Primary type, although for

ETW the same model indicated a positive association for the All-level type and a negative one for Preschool and Primary (Table 6.3). Here the signs of the coefficients estimated need to be understood differently given the nature of these distance-based covariates. Thus, higher distances to Preschool and Primary schools correlated with fewer traffic accidents, whereas the same situation was associated with more traffic accidents when considering All-level schools. In other words, a decrease in accident counts was attributed to the proximity of an All-level school (given these results), while an increase arose with the close presence of Preschools and Primary schools.

Main roads and higher bus stop density correlated with more traffic accidents for STW and ETW, as shown in Table 6.3. On the other hand, higher land use entropy was associated with fewer accidents (Table 6.3), possibly suggesting calming effects in traffic produced by the coexistence of different types of facilities. The positive association of traffic accidents with main roads and areas dense in bus stops is consistent with other studies (Yu, 2015; Yu and Zhu, 2016). However, the result for land use entropy is far more unexpected according to previous literature (Rothman et al., 2017b). Moreover, complex intersections were associated with more traffic accidents at STW (also showing a positive estimate for ETW), which seems entirely plausible given the natural increase in risk that these road entities produce (Miaou and Lord, 2003; Huang et al., 2017; Lee et al., 2017). The analysis of intersection characteristics and their relationship with actual and perceived risk for students is another topic of interest (Lee et al., 2016).

	STW					**ETW**				
Covariate	Est.	p_5	p_{10}	p_{90}	p_{95}	Est.	p_5	p_{10}	p_{90}	p_{95}
(Intercept)	**1.29**	1.19	1.21	1.36	1.38	**2.58**	2.51	2.53	2.63	2.65
No. of schools	-0.00	-0.11	-0.09	0.08	0.11	-0.06	-0.14	-0.13	0.01	0.03
Distance to closest All-level	0.13	-0.07	-0.03	0.28	0.33	**0.20**	0.06	0.09	0.31	0.34
Distance to closest Preschool	-0.10	-0.27	-0.23	0.03	0.07	**-0.17**	-0.29	-0.26	-0.07	-0.04
Distance to closest Primary	**-0.22**	-0.44	-0.39	-0.04	0.01	**-0.32**	-0.48	-0.44	-0.20	-0.16
Distance to closest Secondary	-0.04	-0.19	-0.16	0.08	0.11	-0.01	-0.12	-0.10	0.08	0.11
No. of educational services	-0.04	-0.20	-0.16	0.08	0.12	-0.01	-0.13	-0.10	0.09	0.11
No. of services (non-educational)	0.01	-0.18	-0.14	0.16	0.21	0.06	-0.09	-0.05	0.18	0.22
% of road with parking slots	0.01	-0.11	-0.08	0.10	0.12	0.06	-0.02	-0.01	0.13	0.15
No. of parking zones	0.01	-0.13	-0.10	0.12	0.16	0.04	-0.07	-0.05	0.12	0.14
No. of bus stops	**0.13**	0.01	0.04	0.22	0.25	**0.08**	-0.00	0.02	0.15	0.17
Land use entropy	**-0.11**	-0.22	-0.20	-0.03	-0.01	**-0.17**	-0.25	-0.23	-0.10	-0.09
Betweenness	-0.01	-0.12	-0.10	0.07	0.10	-0.01	-0.10	-0.08	0.05	0.07
No. of complex intersections	**0.13**	0.01	0.04	0.23	0.26	0.05	-0.05	-0.03	0.12	0.14
Main road length	**0.34**	0.20	0.23	0.45	0.48	**0.23**	0.12	0.15	0.31	0.33
No. of traffic lights	-0.04	-0.20	-0.16	0.08	0.12	-0.01	-0.13	-0.10	0.09	0.11
No. of school-aged residents	0.01	-0.12	-0.09	0.11	0.14	-0.07	-0.17	-0.15	0.00	0.03
Percentage of high-end cars	0.06	-0.06	-0.04	0.17	0.19	0.07	-0.03	-0.01	0.14	0.17
ψ	**3.34**	2.35	2.52	4.30	4.66	**4.15**	3.23	3.41	4.95	5.21

Table 6.3: Estimates (Est.) and 5th, 10th, 90th and 95th percentiles of the posterior distributions of the parameters involved in the CAR model for both time windows considered (STW and ETW). Estimates shown in bold are significant with 90% credibility



(a) (b)

Figure 6.3: Hexagonal grid coloured according to accident counts observed at STW (a) and ETW (b). The school locations are represented with black squares (All-level), circles (Primary), diamonds (Preschool) and triangles (Secondary)

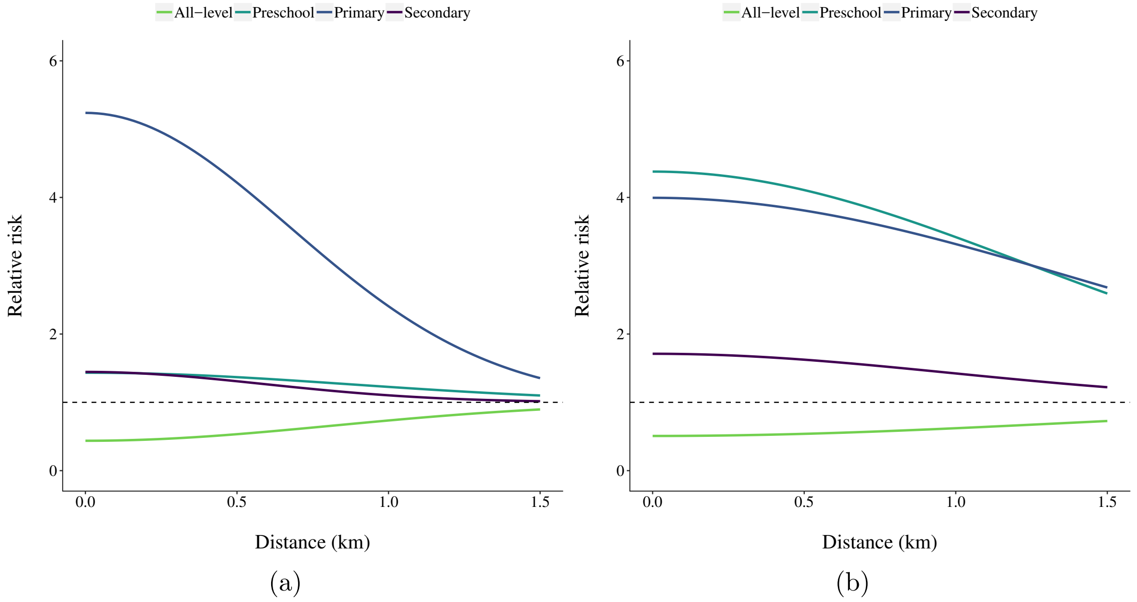
6.4.3 Multiple source regression

The model described in Equation 6.3 was fitted for the four types of schools, which were taking the role of sources of putative risk, and the set of covariates considered. The optimization of the log-likelihood function associated with this model (Equation 6.4) required choosing some constraints for the parameters involved to reach convergence. The addition of covariates to the model was done successively in order to avoid overfitted models and to guide the choice of the constraints. The baseline model with no covariates was also tested. Finally, the constraints defined were $-1 \leq \lambda_m \leq 1$, $-1 \leq \alpha_j \leq 5$ and $0 \leq \beta_j \leq 2$. Furthermore, only the four covariates that showed a greater effect through the CAR modelling of accident counts were used in the final model: the number of bus stops, land use entropy, the number of complex intersections and main road length. Hence, a parsimonious criterion was followed, as the inclusion of more covariates barely increased the value of the log-likelihood function. The parameter estimates obtained for the multiple source regression model are shown in Table 6.4. The relative risk curves (as a function of the distance to each school type) derived from the estimates of the α and β parameters are shown in Figure 6.4.

All the coefficients found for the multiple source regression are coherent overall with those obtained through the macroscopic CAR modelling. In particular, the differential effects produced by the two factors, school type and time window of analysis, appear again, as is evident from Figure 6.4. The use of this modelling technique enables the risk to be characterized easily as a function of the distance to the source. For instance, according to the results obtained, the relative risk starts at a value slightly over 4 in the immediate vicinity of a Primary school for the ETW, and is then progressively reduced to 2 as the distance from the source reaches a value of 1.5 km.

	λ_1	λ_2	λ_3	λ_4	α_1	α_2	α_3	α_4	β_1	β_2	β_3	β_4
STW	0.05	0.33	0.09	-0.13	-0.56	0.43	4.24	0.45	1.15	1.24	0.95	0.83
ETW	0.09	0.29	0.08	-0.18	-0.49	3.38	2.99	0.71	1.96	1.73	1.97	1.39

Table 6.4: Parameter estimates obtained for the multiple source regression models in the two time windows, STW and ETW. Parameters λ_1, λ_2, λ_3 and λ_4 were associated with the number of complex intersections, main road length, number of bus stops and land use entropy, respectively. The indexes for the α's and β's represent: $1 =$ All-level, $2 =$ Preschool, $3 =$ Primary, $4 =$ Secondary



(a) (b)

Figure 6.4: Relative risk curves derived from the source regression model with covariates for the four types of schools available in Valencia in STW (a) and ETW (b). The black dashed line indicates a relative risk of 1

6.4.4 Multivariate logistic regression

A total of 40 cases and 160 controls were established for the STW (Figure 6.5a), and 45 cases and 155 controls for the ETW (Figure 6.5b), following the Monte Carlo procedure explained in Section 3.3. Hence, a fair ratio of around 4 control schools for each school defined as a case were included in the multivariate logistic model. As indicated in Section 3.5, the choice of a radius $\sigma = 100$ m for the construction of the buffer zones was based on the availability of a controls/cases ratio that allows detecting the differences between the two conditions. The consideration of a higher value of σ made the non-overlapping condition between buffer zones too restrictive to obtain a sufficient sample of both cases and controls, whereas a smaller value complicated particularly the identification of cases given the loss of sample size and statistical power.

Table 6.5 shows the results for the multivariate logistic regression fitted to the samples of cases and controls (for STW and ETW) and covariates available in a radius of 100 m around them. Contrary to the CAR modelling of accident counts over the hexagonal grid, the results from the case-control approach presented multiple differences between STW and ETW. Indeed, the number of complex intersections, the percentage of road length with parking slots, the number of parking zones and the percentage of high-end cars decreased the probability of a school being a case in STW, whereas the number of non-educational services and the number of bus stops had the same effect for the ETW. On the other side, the number of traffic lights was the only covariate that increased the probability of case in the STW, while main road length, the number of educational services and the percentage of high-end cars showed this behaviour for the ETW.

These findings will now be discussed in relation to a school location being unsafe (case) or not (control). The facility of parking (represented by both the percentage of road providing parking slots and the number of parking zones) indicated a protective effect in the STW. This is a reasonable result, as a lack of parking opportunities usually generates complicated and competitive traffic movements that may be responsible for an increase in traffic accidents. On the other hand, the number of parking zones seems to increase the risk in the ETW. The proximity of most of the parking zones in Valencia to shopping areas may be producing this effect in the ETW. In any case, one should not overlook the fact that establishing a correlation between parking difficulties and the number of vehicles arriving at a school at starting and ending times (possibly causing complex traffic dynamics that may lead to more accidents) is a hard task. Indeed, the scarcity of parking facilities leads many students' parents to reduce the use of their private vehicles to commute to school, but such reductions can also be seen as an incentive for other parents, generating some kind of equilibrium situation (Black et al., 2001).

Regarding the increase in risk suggested by main road length and the number of educational services in the ETW, both were expected. The first, because it usually correlates with more traffic accidents, regardless of the context being studied. The second, because these activities attract high numbers of commuters in the ETW, although the tendency to locate educational services near schools may be increasing this effect (which may also be responsible for the positive association of educational activities with traffic accidents in the STW).

The rest of the associations revealed by the case-control analysis are harder to interpret and in many cases require the consideration of confounding factors that have not been included in the models due to their unavailability. For example, the number of complex intersections was associated with a decrease in the probability of being a risky school zone, the opposite result to that found with the CAR model. It is logical that users drive more carefully in a small area (100 m buffer zone) which is dense in complex intersections, a fact that may explain this result. However, when a higher level of spatial aggregation is considered, as in the macroscopic modelling described in Section 3.4, a high number of complex intersections is more likely to be associated with an increase in traffic accidents.

The number of bus stops, which showed a positive correlation with traffic accidents according to the macroscopic modelling, was associated with a decrease in risk at the ETW. The fact of being a school located in a zone of the city endowed with several bus stops should reduce the number of commuters by private vehicle and therefore traffic congestion, although a higher number of bus stops does not always imply a higher level of connectivity within the public bus network. As in the case of complex intersections, at a higher level of spatial aggregation, it is more plausible that bus stop density should be associated with an increase in traffic dangerousness.

The behaviour of other covariates, such as the number of traffic lights or the number of non-educational services, may be a consequence of the presence of confounding effects. The number of traffic lights in the STW is possibly absorbing the more expected effect of main road length, as these two covariates are moderately correlated. On the other hand, the number of non-educational services is highly concentrated in the southern zone of the city center and its contiguous district in this direction, which mostly presented control schools in the ETW, even though the two facts do not seem to be related.

Another covariate whose behaviour showed a strong dependence on the time window being considered was the percentage of high-end cars registered in the zone. This covariate was chosen as an approximation to socioeconomic status, given the unavailability of other sources of information (such as income level or housing price), yielding a positive association with traffic accidents only in ETW. Participating in extracurricular activities is much more frequent among students belonging to wealthier families (Leung et al., 2019), as most of these activities are not publicly financed. Consequently, more vehicles may be arriving at schools located in wealthier areas and afterwards transporting students to the location of the extracurricular activities. In contrast, students enrolled in schools situated in more economically depressed areas of the city are more likely to go home by their own means (under the plausible assumption that students' homes are on average closer to their school than extracurricular activities that do not take place at the same school). Nevertheless, these findings would require a deeper and more specific investigation to confirm this effect.

Finally, the categorical covariate adding the effect of school type presented different behaviour between the STW and the ETW, considering that in Table 6.5 the estimates are computed in relation to the All-level type (which is hidden because it is the reference category). In the case of the STW, Primary and Secondary types showed a significant increase in the odds ratios for being a case, in agreement with the other models tested (especially in the case of Primary schools). However, all the estimates for the ETW were not significant with 90% credibility, a surprising result in view of the previous models.

6.5 Conclusions

In this Chapter, several approaches have been followed to measure the effect of school locations on traffic accidents in Valencia. From a methodological perspective, the main conclusion is the desirability of applying different methods to obtain consistent knowledge about the phenomenon of interest. All the statistical techniques chosen

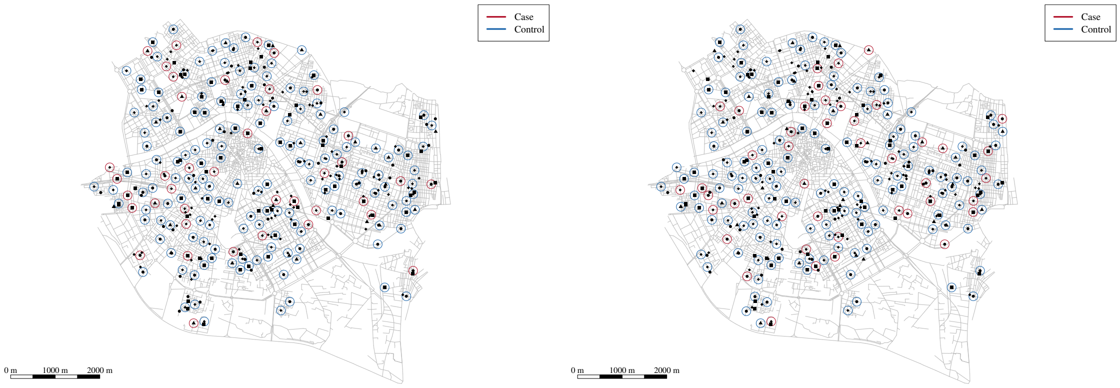
	STW					**ETW**				
Covariate	Est.	p_5	p_{10}	p_{90}	p_{95}	Est.	p_5	p_{10}	p_{90}	p_{95}
(Intercept)	**-2.13**	-2.90	-2.71	-1.57	-1.43	**-0.95**	-1.64	-1.47	-0.44	-0.30
No. of educational services	**0.30**	-0.04	0.03	0.57	0.65	**0.66**	0.24	0.33	0.99	1.09
No. of services (non-educational)	0.16	-0.51	-0.35	0.66	0.79	**-1.04**	-1.80	-1.61	-0.51	-0.38
% of road with parking slots	**-0.36**	-0.75	-0.66	-0.08	-0.00	0.07	-0.29	-0.21	0.36	0.44
No. of parking zones	**-0.47**	-0.98	-0.85	-0.12	-0.05	**0.30**	-0.01	0.05	0.56	0.64
No. of bus stops	0.14	-0.20	-0.12	0.40	0.48	-0.23	-0.61	-0.52	0.05	0.12
Land use entropy	**-0.40**	-0.81	-0.71	-0.09	-0.00	0.01	-0.38	-0.29	0.32	0.40
Betweenness	0.05	-0.30	-0.22	0.31	0.37	0.08	-0.31	-0.21	0.37	0.44
No. of complex intersections	**-0.49**	-0.92	-0.82	-0.17	-0.09	0.09	-0.24	-0.17	0.34	0.41
Main road length	0.02	-0.39	-0.29	0.33	0.42	**0.35**	-0.04	0.05	0.65	0.74
No. of traffic lights	**0.58**	0.17	0.25	0.92	1.01	0.25	-0.16	-0.07	0.57	0.66
No. of school-aged residents	0.09	-0.32	-0.23	0.40	0.48	0.11	-0.33	-0.24	0.45	0.54
Percentage of high-end cars	-0.16	-0.59	-0.48	0.14	0.22	0.25	-0.11	-0.02	0.52	0.59
School type (Preschool)	0.17	-0.71	-0.52	0.87	1.08	-0.51	-1.36	-1.17	0.17	0.36
School type (Primary)	**1.11**	0.15	0.36	1.87	2.10	-0.48	-1.40	-1.20	0.23	0.43
School type (Secondary)	**1.14**	-0.01	0.24	2.05	2.31	-0.63	-1.94	-1.63	0.36	0.62

Table 6.5: Estimates (Est.) and 5th, 10th, 90th and 95th percentiles of the posterior distributions of the parameters involved in the logistic model for both time windows considered (STW and ETW). Estimates shown in bold are significant with 90% credibility

for accomplishing the research purposes established have provided information about the question of interest from different perspectives, providing several points of agreement between them. However, each technique has offered a particular view of the phenomenon, highlighting the necessity of carefully choosing a specific approach and even the advisability of using more than one in order to strengthen the findings.

First, observed vs expected ratios of traffic accidents were computed at a range of distances from school locations. Although this approach has the disadvantage of not allowing the use of any auxiliary information (covariates), it serves to get an overall idea of the magnitude of the effect being analyzed. In this study, it was a starting point which suggested that school locations have an impact on traffic accidents, albeit only a moderate impact. Furthermore, the observed vs expected ratios also made it possible to define case and control schools for the subsequent performance of a case-control study design. In the absence of external information (if no particular school zone has been declared dangerous by traffic experts), this simple method can fill the gap.

Besides observed vs expected ratios, three other statistical models were used in the analysis: macroscopic CAR modelling, multiple source regression and a logistic regression under a case-control design. The macroscopic modelling of accident counts indicated that some environmental and traffic-related factors, such as the number of bus stops, land use entropy, main road length and the number of complex intersections, have an effect on the incidence of traffic accidents in the two time windows investigated within school-days. Furthermore, this model enabled confirming that in the city of Valencia the type of school entails a specific risk with respect to traffic accident occurrence. The last statement was also confirmed through the multiple source regression model. According to both modelling approaches, the All-level



(a)

(b)

Figure 6.5: Case and control buffer zones (100 m radius) defined around school locations in STW (a) and in ETW (b). The school locations are represented by black squares (All-level), circles (Primary), diamonds (Preschool) and triangles (Secondary)

school type shows an overall protective effect against traffic accidents in the city of Valencia. On the other hand, proximity to Primary (in both time windows studied) and Preschool schools (in the ETW) is subject to more risk.

Regarding the case-control study, the logistic models showed substantially different results for the STW and the ETW. For instance, this modelling approach does not detect any significant effect from school types for the ETW, a result that is inconsistent with the other methods employed. It is worth of noting that the case-control study design may be too sensitive to the choice of cases and controls and the results should be interpreted with caution, especially if the methodology for selecting both cases and controls is not supported by other evidence apart from the not large sample of traffic accidents available. Furthermore, it needs to be remembered that each model type relies on a different type of spatial unit of analysis (arbitrary hexagonal units covering the city vs buffer zones around schools), a fact that is likely to give rise to MAUP effects. Hence, the scope and extent of each approach is different and direct comparisons are not completely suitable. In any case, the results suggest that this approach could be powerful for detecting small effects that a macroscopic modelling of accident counts may otherwise miss. One drawback of this approach, in comparison to the former, is the likely possibility of dealing with small samples of school locations that can make difficult, or even impossible, the achievement of reliable parameter estimates.

From a practical perspective, within the context of traffic safety around school locations, two important results are now discussed: the need to distinguish between coexisting school types and the consideration of time windows other than those that strictly correspond to school starting and ending times. With regard to school types, and getting back to the fact that All-level schools have been shown to be a safer context for vehicle traffic, it is worth noting that the differences between school types coexisting in Valencia go beyond the age levels. Indeed, most of the All-level schools in Valencia are semiprivate, and parents of students enrolled in this kind of school have, on average, higher income levels and educational attainment than parents of students registered at public institutions (Llera and Pérez, 2012), a category that includes the vast majority of Primary and Secondary schools in Valencia. Furthermore, it needs to be remarked that it is a very challenging task to disentangle the origin of the protective effect found for All-level schools, meaning that it is unclear whether commuting time is particularly dangerous for certain school surroundings given the travel mode choices or travel behaviours of their students (or of the adults in charge) or, alternatively, whether some schools (or school types) are located in intrinsically riskier/safer areas for driving. Although it is known that All-level schools in Valencia correlate with higher socioeconomic status of students, it is not possible to extract more conclusions in this regard, in the absence of more precise information on commuters' travel mode choices and attitudes towards safe driving, walking or cycling. It can only be guessed that more educated parents will be more predisposed towards safety (Murray, 1998). The place of residence and socioeconomic characteristics of students belonging to each school, which are sometimes used to make approximations about travel modes (Wilson et al., 2010; Kelly and Fu, 2014), were not available either. In any case, even if this information was available, the inference of travel modes from students' characteristics could have been highly uncertain, as Valencia is quite different from the cities where these studies were developed.

On the other hand, it is remarkable that the evening time window is often overlooked in the study of traffic safety around school locations. Many of the research papers available focus on the first hours in the morning, but traffic generated in the evening as a consequence of collecting students and taking them to leisure activities deserves more attention. Indeed, several papers that have compared school and non-school trips (those related to extracurricular or leisure activities) have coincided in pointing out that the use of private vehicles is much more likely in non-school trips (Hjorthol and Fyhri, 2009; Fyhri and Hjorthol, 2009; Park et al., 2018), probably augmenting traffic volume. The initial exploratory analysis and the computation of risk ratios near schools already suggested this situation, which was confirmed via the macroscopic modelling and the multiple source regression with the detection of significant effects for most of the school types.

Finally, it is worth mentioning the main limitations of this study, which are a consequence of data unavailability. Besides the absence of information regarding travel mode choices, there was also a lack of data regarding the trip destination of vehicles involved in the traffic accidents employed for the analysis and of the age of the people travelling in them. In other words, an unproved causal relationship had to be established: traffic accidents that occurred in proximity to schools in the STW or ETW were assumed to be a consequence of the traffic dynamics generated by people commuting to school. Of course, these accidents could involve the commuters themselves or other traffic users with a non-school destination not influenced by school commuters. Furthermore, one cannot overlook the possibility that traffic accidents temporally associated with school starting and ending times do not occur as close to school locations as one might expect, a possibility suggested by Kingham et al. (2011), which makes this kind of research very challenging.

Chapter 7

Geocoding quality and missing data: Reestimating a minimum acceptable hit rate

The main goal of this Chapter was to revise the minimum acceptable hit rate determined by Ratcliffe (2004a), while accounting for some of the factors that may have an influence on such estimation. Indeed, the two characteristics that essentially characterize a point pattern of events, namely intensity and clustering level, and the level of aggregation presented by the set of spatial units covering the whole space where the pattern lies are explicitly analyzed. Some alternatives to the Mann-Whitney test, or the specification of non-uniform geocoding error distributions in the context of Ratcliffe's procedure, are also considered. The investigation is performed through several crime datasets collected by the Local Police of Valencia (Spain) and simulated datasets that allowed controlling for each of the factors under consideration.

7.1 Introduction

The process of geocoding is generally automated, although its success strongly depends on both the quality of the textual description representing the place to be geocoded and the quality of the base street files that are employed for matching the results. Indeed, Zandbergen (2009) identified three main errors that derive from the process of geocoding: incompleteness (non-geocoded locations), positional error (distance between the geocoded location and the real location), and erroneous assignment to some underlying areal unit.

The existence of geocoding errors or inaccuracies can lead researchers to perform biased statistical analyses and hence extract mistaken conclusions from them. There are multiple studies available in the literature that have focused on this issue. Oliver et al. (2005) compared the distribution and clustering patterns of cancer cases at counties (∼100% geocoding success) to those presented at census tracts (∼74%

geocoding success), observing notorious disparities. Zimmerman et al. (2008) showed that the statistical power for detecting spatial disease clustering gets reduced if both the disease risk and the geocoding failure distributions are spatially associated. DeLuca and Kanaroglou (2008) checked that the choice of a spatial geocoding tool can have an impact on the subsequent data analysis. Concretely, although the bivariate K-function (Diggle, 2013) and to a lesser extent the ratio of kernel density estimates did not show remarkable disparities for the two geocoding tools that they tested, the scan statistic for cluster detection (Kulldorff, 1997) displayed a greater sensitivity to the choice of a specific geocoding tool.

In the present Chapter, the focus is put on the first of the three geocoding error types identified by Zandbergen (2009): data incompleteness that arises as a consequence of having non-geocoded locations. In this regard, Harries et al. (1999) identified several common errors that usually lead to completely fail to geocode a location described by a text, such as misspelling the street name, entering an incorrect street type or introducing a house number that does not correspond to the street provided. Given a dataset of locations to be geocoded, the hit rate (or match rate) is the percentage of the locations that are geocoded with a certain level of accuracy (prespecified according to research conditions or purposes). Therefore, a natural question that arises in this context is: what is the minimum acceptable hit rate that one needs to achieve to rely on the subsequent analysis of the data? Although this question has received little attention in literature, Ratcliffe (2004a) designed a methodology (that will be described later) to provide an answer to it. Then, the methodology was applied to five crime datasets collected at different areas of New South Wales (Australia), and a hit rate of 85% was determined to be a first estimate of a minimum acceptable geocoding hit rate.

It is worth noting that the original minimum acceptable hit rate of Ratcliffe (2004a) still remains, to the best of my knowledge, as the only estimation available in literature. Indeed, this paper has been cited repeatedly since its publication by many researchers performing a spatial analysis (near 300 citations according to Google Scholar as of August 2019), who usually declare that a geocoding hit rate above the threshold of 85% has been achieved (as a guarantee of having geocoded a large enough percentage of the events of interest). Furthermore, the minimum acceptable hit rate of 85% has also been considered for the construction of a nonparametric test of similarity for area-based spatial patterns (Andresen, 2009, 2016), which is largely used in quantitative criminology, among other fields of research.

One limitation of the investigation conducted by Ratcliffe (2004a) (as the author himself declares) is the fact that it assumes that geocoding errors are uniformly distributed across space. There is enough evidence, in view of existing literature, to firmly believe that this assumption may not hold in reality: recently constructed houses may not be included in the base street file being used for geocoding purposes (Ratcliffe, 2004a); rural zones and street segment length (for urban addresses) associate with higher positional errors, whereas road intersection density correlates with a lower incidence of error (Zimmerman and Li, 2010); geocoding failure rates associates with certain socio-demographic covariates (Oliver et al., 2005).

In addition, it is worth mentioning that the estimate proposed by Ratcliffe (2004a) is based on the use of the Mann-Whitney nonparametric test for comparing two count distributions. However, there exist several area-based (Alba-Fernández et al., 2016; Andresen, 2009, 2016; Wheeler et al., 2018) and distance-based (Hahn, 2012) tests in literature that could be employed instead to compare a fully geocoded dataset with an incomplete version of it. Other methodologies inspired on kernel smoothing that do not rely on areal data could be also considered (Borrajo et al., 2019; Fuentes-Santos et al., 2017). Furthermore, certain statistical measures that are usually employed in spatial analyses, such as some popular indices to assess global (Moran, 1950a,b) or local spatial association (Anselin, 1995; Getis and Ord, 1992), may also be particularly affected by lower than desired geocoding hit rates. Hence, although the Mann-Whitney test is a good option for the purposes established in Ratcliffe's methodology, exploring some of the alternative approaches available seems rather convenient.

Hence, the rest of this Chapter is structured as follows. The Data section includes a description of the crime datasets and spatial representations of Valencia (Spain) that were used to reestimate and investigate Ratcliffe's first estimate of a minimum acceptable hit rate. The Methodology section contains multiple subsections. First, the procedure proposed by Ratcliffe (2004a) is depicted. Then, other statistical measures that could be chosen for determining this kind of estimation (instead of the Mann-Whitney test) and a strategy to simulate a non-random removal of the points within Ratcliffe's procedure are also explained. The last subsection of the Methodology is dedicated to point processes and point pattern generation, which are essential for the performance of simulation studies. The Results section has a subsection corresponding to each of the analyses that were executed, which involved the crime datasets of Valencia mentioned above and several simulation studies. A final Conclusions section summarizes the findings and presents some discussions.

7.2 Data

7.2.1 Crime events from Valencia

A dataset of 11768 crime incidents recorded by the Local Police of Valencia during the years 2014 and 2015 was considered for the analysis. For each incident, geographical coordinates had been obtained by the 092 Call Center of Valencia managed by the Local Police. This service geocodes the origin of citizen calls in order to offer a quick response to the incident. Thus, it is assumed that a dataset coming from the 092 Call Center is geocoded with a 100% hit rate.

The incidents that the 092 Call Center receives are codified according to caller's explanations. The dataset was constructed from the calls codified as robbery (5332 events), burglary (1198 events), theft (2465 events), car theft (706 events) and vandalism (2067 events). Hence, the whole crime dataset was splitted into five smaller datasets, allowing the investigation of various levels of intensity and spatial clustering depending on the type of crime.

7.2.2 Spatial representations of Valencia

Four different spatial structures representing the city of Valencia were chosen for the analysis, which are displayed in Figure 7.1. Three of them are areal subdivisions of the city: boroughs, Thiessen tessellations constructed around main road intersections and census tracts. The fourth structure considered was the road network of Valencia.

Boroughs and census tracts are two of the three administrative divisions coexisting in Valencia (districts is the other one). A total of 70 boroughs and 566 census tracts of the city (as of January of 2015) were used for the analysis. Thus, the choice of both administrative units allowed approaching the research at two very different spatial scales.

A division of the city through a Thiessen tessellation, which was already described on Section 5.2.3, was also performed. Concretely, the polygons were set around the main road intersections of the city, meaning the physical intersections between any combination of two main roads of Valencia. Hence, each of the Thiessen polygons defined represents the neighbourhood of a main intersection of the city. This procedure generated a total of 378 Thiessen polygons, which constitutes an intermediate spatial scale between boroughs and census tracts.

Finally, a representation of the road network of Valencia made of 9099 vertices and 12886 road segments (spatial units where events are located) was also available for the analysis. The choice of this structure may imply losing a part of the sample, as some events take place in zones of the city that are not close to the road network and hence cannot be properly located in the structure. In order to reduce network's complexity and hence favour a less sparse distribution of the event counts at the road segment level, the road network was simplified using the SpNetPrep R package (Briz-Redón, 2019). Therefore, the road network finally used had 7986 vertices and 11773 segments, which approximately represents a 9% reduction in the number of segments with respect to the original one.

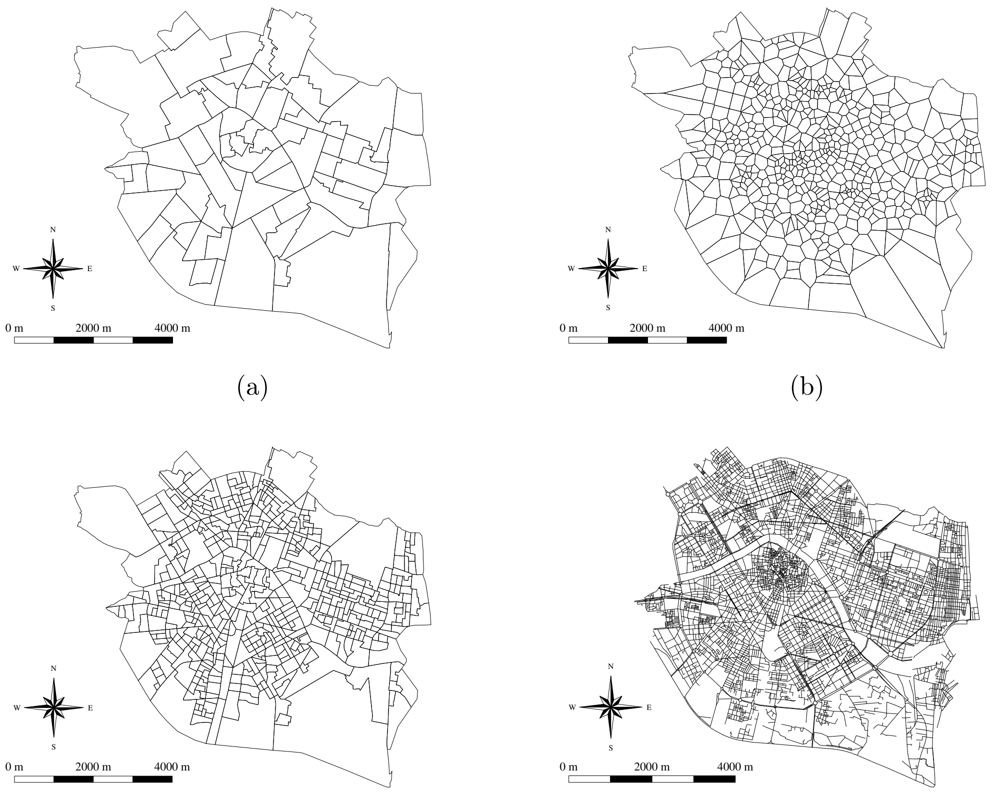
7.3 Methodology

7.3.1 Software

The R programming language (3.5.2 version, R Development Core Team, Vienna, Austria) (R Core Team, 2018) was used to obtain all the results presented in this study. The R packages ggplot2 (Wickham, 2016), rgeos (Bivand and Rundel, 2018a), spatstat (Baddeley et al., 2015), spded (Bivand and Piras, 2015) and SpNetPrep (Briz-Redón, 2019) were specifically required for performing some parts of the study.

7.3.2 Estimation of a minimum acceptable hit rate

The method proposed by Ratcliffe (2004a) was used to determine a minimum acceptable hit rate for each spatial structure and point pattern under investigation.



(a)

(b)

(c)

(d)

Figure 7.1: The four spatial configurations of Valencia that were employed for the analysis: boroughs (a), Thiessen tessellations based on the main road intersections of the city (b), census tracts (c) and road network (d)

This method follows a Monte Carlo approach, which, given a dataset of events, allows the generation of an empirical distribution for the minimum hit rate through an iterative process. The steps that need to be carried out are the following (for each iteration):

1. Start with a point pattern lying on a spatial structure, assuming a 100% hit rate for the pattern (no data is missing). Compute the original distribution of the points across the spatial units forming the spatial structure (this means counting the number of events lying on each spatial unit).

2. Randomly select 1% of the points of the pattern and remove them.

3. Compute the new distribution of the points across the spatial structure.

4. Assess if the new and the original distributions are statistically different (through the Mann-Whitney test). If they are statistically different, record the percentage of points removed and start again at (1) (new iteration). If they are not statistically different, go back to step (2) and repeat the subsequent steps.

The replication of these steps for a certain number of times enables generating an empirical distribution of the minimum acceptable hit rate, which is the variable under analysis. The mean value of the empirical distribution can be taken as the estimate of the minimum acceptable hit rate.

A workflow diagram that may be helpful to better understand this procedure is available in Ratcliffe (2004a).

7.3.3 Beyond the Mann-Whitney test

Two methods other than the Mann-Whitney test have been considered to estimate a minimum acceptable geocoding hit rate. First, a parametric approach has been tested consisting in the specification of a generalized linear model that assesses if there is a significant dissimilarity between two count distributions. Second, a cluster and outlier detection procedure based on local indicators of spatial association (Anselin, 1995) has also been implemented. The estimations are obtained for each crime dataset available from Valencia.

Comparing two distributions through generalized linear models

A generalized linear model (GLM) (Faraway, 2016) provides a more flexible modelling framework than ordinary linear regression by allowing model residuals to be non-normally distributed. Concretely, in a GLM, the relationship between a response variable, $Y = \{Y_i\}_{i=1}^n$, following an exponential family distribution with mean $\{\eta_i\}_{i=1}^n$, and a set of regressors, $X_1 = \{X_{1i}\}_{i=1}^n$, ..., $X_k = \{X_{ki}\}_{i=1}^n$, is expressed by means of a link function, $g(\cdot)$ (a natural logarithm is a common choice):

$$g(\eta_i) = \gamma_0 + \gamma_1 X_{1i} + ... + \gamma_k X_{ki}$$

where η_i satisfies $E[Y_i|X_{1i}, ..., X_{ki}] = \eta_i$, γ_0 is a constant term, and γ_j is the coefficient associated to the effect of regressor X_j. The choice of a certain distribution type from the exponential family (Poisson, binomial, negative binomial, etc.) should be based on the properties of Y.

The GLM scheme can be used to assess if two count distributions are significantly different. Indeed, if $Y_1 = \{Y_{1i}\}_{i=1}^n$ and $Y_2 = \{Y_{2i}\}_{i=1}^n$ denote the two distributions, one can consider the union of Y_1 and Y_2, $Y = \{Y_k\}_{k=1}^{2n} = \{Y_{11}, ..., Y_{1n}, Y_{21}, ..., Y_{2n}\}$, and a binary regressor $X = \{X_k\}_{k=1}^{2n}$ such that $X_k = 0$, if $k \leq n$, and 1 otherwise. Then, the following univariate GLM can be fitted for the purpose of comparing the distributions of Y_1 and Y_2:

$$g(\eta_i) = \gamma_0 + \gamma X_i \tag{7.1}$$

Therefore, if the fitting of a GLM model following Equation 7.1 yields that γ is significantly different from 0 ($p < 0.01$), one would conclude that the distributions of Y_1 and Y_2 are significantly different. If one chooses a Gaussian distribution for Y, the test just described is exactly equivalent to the classical Student's t-test for comparing two samples. However, a discrete distribution would be more suitable if Y_1 and Y_2 represent counts of events.

Cluster and outlier detection with LISA

As it has been stated in previous Chapters, Moran's I (Moran, 1950a,b) is a measure of global spatial autocorrelation for area-based spatial patterns. A higher Moran's I value indicates a higher tendency of the variable to show strongly associated values for neighbouring spatial units. The local version of Moran's I index called LISA (Anselin, 1995) can be used for performing cluster/outlier detection, as shown in Chapter 2 (hotspots and coldspots are two types of clusters). If I_i denotes the local indicator of spatial association for spatial unit i, then spatial units showing a significant value $I_i \neq 0$ ($p < 0.01$) can be assigned to one of the following four categories: high-high (unit i has a high value and it is surrounded by high-valued units), high-low (unit i has a high value, but it is surrounded by low-valued units), low-high (unit i has a low value, but it is surrounded by high-valued units) and low-low (unit i has a low value and it is surrounded by low-valued units). The high-high and low-low zones are usually referred to as clusters, whereas high-low and low-high zones are commonly called outliers.

In the context of Ratcliffe's procedure, it is considered that two datasets (one complete and the other one missing a certain percentage of points) provide the same information if their corresponding categorizations of spatial units according to LISA indices are exactly the same. Hence, once the percentage of points removed led to an erroneous categorization of (at least) one spatial unit, that percentage was annotated for the corresponding iteration.

7.3.4 Introduction of non-uniform geocoding error rates in Ratcliffe's procedure

The consideration of non-uniform (in space) geocoding error rates and its implementation within Ratcliffe's procedure was performed as follows. First, events were classified according to a given condition (dependent on some spatial characteristic) into two groups representing two different geocoding error rates: "baseline" and "increased". To this end, a numeric parameter denoted by β was introduced, allowing the distinction between those events belonging to the "baseline" group and those belonging to the "increased" group. More specifically, events belonging to the "baseline" group were assigned a weight of 1, whereas events corresponding to group "increased" were assigned a weight $1 + \beta$. Then, the probability of removing one event from a dataset (in the context of the procedure described in Section 7.3.2) is set to be proportional to the weight assigned to the event (1 or $1 + \beta$). In other words, this selection procedure assumes that an event of the group "increased" is more likely to be non-geocoded by a factor $1 + \beta$.

It is important to remark that the method just described involving parameter β is insufficient to represent all kind of non-uniform geocoding error rates that may arise for a given dataset. However, the strategy proposed in this section is enough to measure how the presence of certain non-uniform geocoding rates across space may affect the estimation of a minimum acceptable hit rate. Specifically, several real situations that may originate non-uniform geocoding error rates can be modelled through this method, where a higher value of β represents a greater disparity from uniformity.

7.3.5 Remarks on point process theory

An homogeneous planar Poisson point process (Diggle, 2013) (from now on, referred to as an homogeneous Poisson process) was considered for the analysis of the effect of the intensity of the pattern on the minimum acceptable hit rate. A Matérn cluster process (Matérn, 1986) (from now on, referred to as a Matérn process) was selected for simulating point patterns of varying levels of intensity and clustering. In the following lines both processes are briefly described.

An homogeneous Poisson process of intensity $\lambda > 0$ satisfies the following three properties (Diggle, 2013):

1. Given an areal region of the plane, A, the number of points that lie on A, $N(A)$, follows a Poisson distribution with mean $\lambda|A|$.

2. If $N(A) = n$, the n points are distributed uniformly on A.

3. If A and B are two disjoint regions of the plane, $N(A)$ and $N(B)$ are independent variables.

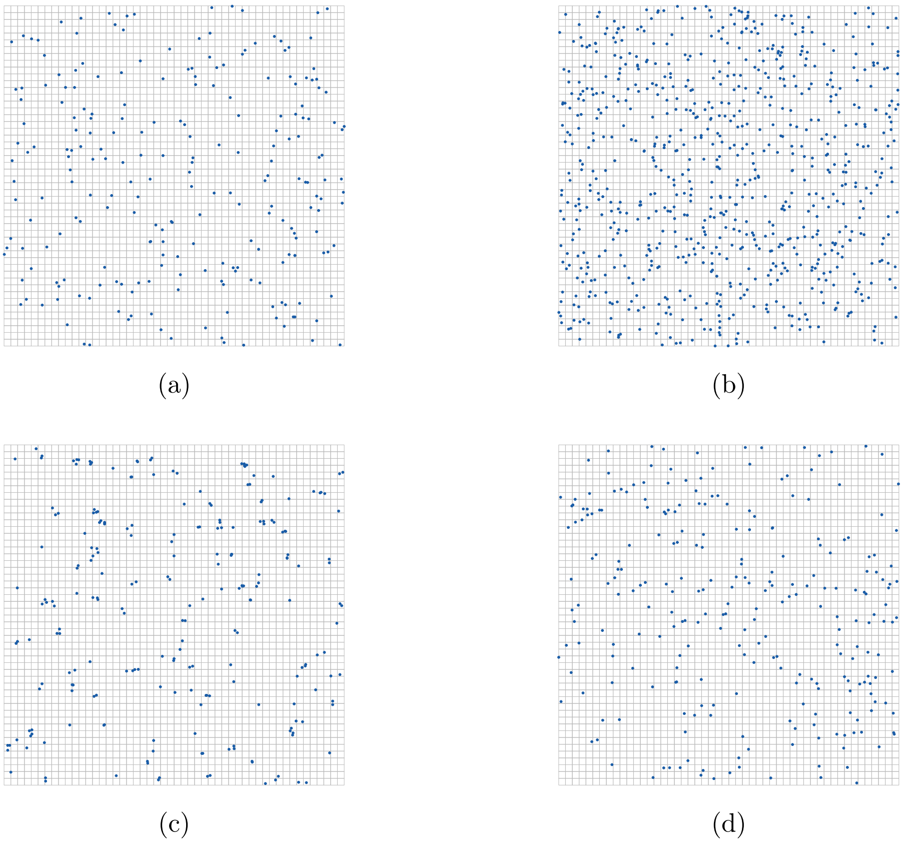
On the other hand, a Matérn process of parameters $\lambda > 0$, $s > 0$ (scale) and $\mu > 0$ can be generated through the following two steps:

1. An homogeneous Poisson process of intensity λ generates a set of “parent points”.

2. A cluster of “offspring points" is generated (uniformly) inside a disc of radius s centred on each “parent point”. The number of “offspring points" lying around each “parent point” follows a Poisson distribution of parameter μ.

The Matérn process is a special case of the more general class of Neyman-Scott processes (Neyman and Scott, 1958). The intensity of a Matérn process of parameters λ, s and μ is $\lambda \cdot \mu$. Increasing the value of s implies that the points forming each cluster are progressively less concentrated around their corresponding “parent points”, leading to relatively lowly-clustered point patterns. Contrarily, very small values of s are expected to generate highly-clustered point patterns.

The generation of point patterns following the properties of both Poisson and Matérn processes was performed with the spatstat R package (Baddeley et al., 2015). As an illustration, Figure 7.2 shows several examples of point patterns generated with this R package for different choices of the parameters (in a squared area representing 2500 m^2). From (a) to (b) the intensity of the underlying homogeneous Poisson process is increased by a factor of 3. On the other hand, the two patterns in (c) and (d) were created following a Matérn process, imposing a higher clustering level to the one in (c) in comparison to the one in (d) ($s = 0.5$ vs $s = 4$).



(a)

(b)

(c)

(d)

Figure 7.2: Four examples of simulated point patterns over a squared area of 2500 m^2: homogeneous Poisson process with $\lambda = 0.1$ (a), homogeneous Poisson process with $\lambda = 0.3$ (b), Matérn process with $(\lambda, s, \mu) = (0.1, 0.5, 1)$ (c) and Matérn process with $(\lambda, s, \mu) = (0.1, 4, 1)$ (d)

7.4 Results

7.4.1 Crime datasets from Valencia

Mann-Whitney test

The first step was to replicate the procedure depicted in Ratcliffe (2004a) with the five crime datasets recorded in Valencia for the period 2014-2015. In this case, however, each crime dataset was investigated considering four different spatial structures covering the city: boroughs, Thiessen polygons generated around main road intersections in the city, census tracts and road network. It was found sufficient to carry out the Monte Carlo process 250 times (as it was done in Ratcliffe (2004a)) to guarantee a reliable estimation of a minimum acceptable hit rate in each case.

The estimations of a minimum acceptable hit rate for the crime datasets of Valencia are shown in Table 7.1, where some basic characteristics of the point patterns corresponding to each of the crime types investigated such as the intensity (number of points per unit area or length, denoted by $\hat{\lambda}$) or the nearest neighbour index (NNI) (defined in Section 5.3.5) are also shown.

First, one can observe in Table 7.1 how the minimum acceptable hit rate (MAHR) varies largely depending on the crime or structure type that is selected for its estimation. It is notorious that, for each crime type, the minimum acceptable hit rate gets higher as the number of units that constitute the spatial representation of Valencia increases. Indeed, in the case of census tracts the minimum acceptable hit rate gets very close to 90% for some of the crime types, whereas for the road network the hit rate attains a maximum value of 99% (for all the datasets). Furthermore, among the five crime types, it is also evident that the minimum acceptable hit rates estimated for the robbery, theft, and vandalism datasets are higher than those found for the burglary and car theft ones. Interestingly, the higher the number of records available in the dataset (the higher the intensity of the point pattern), the higher the minimum acceptable hit rate becomes. On the other hand, it is hard to find a relationship between the clustering level (NNI) of each point pattern and the minimum acceptable hit rate obtained.

Crime	Structure	Units	Events	$\hat{\lambda}$	NNI	MAHR (%)	LO 95% CI	UP 95% CI
Robbery	Boroughs	70	5332	$1.08 \cdot 10^{-4}$	0.58	74.50	73.13	75.87
	Thiessen	566	5332	$1.08 \cdot 10^{-4}$	0.58	83.70	82.39	85.01
	Census tracts	378	5332	$1.08 \cdot 10^{-4}$	0.58	88.74	87.62	89.87
	Network	11773	4858	$6.36 \cdot 10^{-3}$	0.58	99.00	99.00	99.00
Burglary	Boroughs	70	1198	$2.43 \cdot 10^{-5}$	0.62	69.57	67.83	71.32
	Thiessen	566	1198	$2.43 \cdot 10^{-5}$	0.62	72.79	69.15	76.43
	Census tracts	378	1198	$2.43 \cdot 10^{-5}$	0.62	79.52	77.00	82.04
	Network	11773	1116	$1.43 \cdot 10^{-3}$	0.76	99.00	99.00	99.00
Theft	Boroughs	70	2465	$4.99 \cdot 10^{-5}$	0.63	71.70	69.90	73.51
	Thiessen	566	2465	$4.99 \cdot 10^{-5}$	0.63	80.66	79.08	82.24
	Census tracts	378	2465	$4.99 \cdot 10^{-5}$	0.63	86.75	85.19	88.31
	Network	11773	2270	$2.94 \cdot 10^{-3}$	0.68	99.00	99.00	99.00
Car theft	Boroughs	70	706	$1.43 \cdot 10^{-5}$	0.80	71.51	68.19	74.83
	Thiessen	566	706	$1.43 \cdot 10^{-5}$	0.80	77.34	74.30	80.37
	Census tracts	378	706	$1.43 \cdot 10^{-5}$	0.80	80.70	77.86	83.55
	Network	11773	619	$8.42 \cdot 10^{-4}$	0.83	99.00	99.00	99.00
Vandalism	Boroughs	70	2067	$4.19 \cdot 10^{-5}$	0.70	75.97	74.32	77.61
	Thiessen	566	2067	$4.19 \cdot 10^{-5}$	0.70	81.16	79.40	82.92
	Census tracts	378	2067	$4.19 \cdot 10^{-5}$	0.70	86.27	85.01	87.53
	Network	11773	1894	$2.46 \cdot 10^{-3}$	0.69	99.00	99.00	99.00

Table 7.1: Minimum acceptable hit rates (MAHR) estimated for the five crime datasets available for Valencia, along with the lower and upper bounds of the empirical 95% tolerance interval obtained for each estimation. The point patterns corresponding to each of the datasets are described according to their size (number of events), level of clustering (NNI) and intensity ($\hat{\lambda}$). The number of spatial units (areal or linear) of each spatial structure considered is also indicated. It is worth noting that $\hat{\lambda}$ is measured in $1/\text{m}^2$ in the case of boroughs, Thiessen polygons and census tracts, but in 1/m in the case of the road network structure

GLM approach and LISA indices

Table 7.2 shows the minimum acceptable hit rates that were obtained considering the GLM approach and the identification of clusters/outliers through LISA indices.

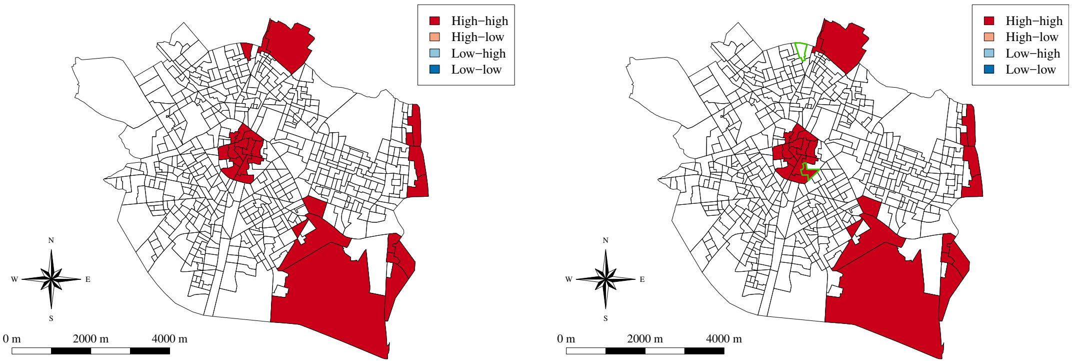
A negative binomial distribution and a natural logarithm as the link function were chosen for setting the GLM approach explained in Section 7.3.3, with the response $Y = \{Y_i\}_{i=1}^{n}$ representing crime counts per spatial unit. The negative binomial was selected with the aim of allowing the model to account for data overdispersion. The estimations provided by this method were very similar to those yielded by the Mann-Whitney test.

Remarkably, the estimations obtained for the road network are again very high, as the removal of 1% of the points forming the dataset alters the distribution of the counts at the road segment level. One might have thought that the high minimum acceptable hit rate obtained for the road network through the Mann-Whitney test could be a consequence of the highly skewed distribution that event counts present at the road segment level (McElduff et al., 2010), but the GLM approach based on a negative binomial probability distribution led to analogous results.

Contrarily, the categorization of the spatial units according to LISA indices showed to be much more sensitive to the existence of non-geocoded data. In this case, minimum acceptable hit rates raised dramatically except for the spatial structure based on city's boroughs. This fact suggests that the estimation of a minimum acceptable hit rate is strongly dependent on the methodology that is chosen to determine if the complete and incomplete datasets provide the same interpretation of the data. As an illustration, Figure 7.3 shows how the results that LISA indices provide can be affected by removing a small percentage of the points of the dataset.

		GLM			**LISA**		
Crime	**Structure**	**MAHR (%)**	**LO 95% CI**	**UP 95% CI**	**MAHR (%)**	**LO 95% CI**	**UP 95% CI**
Robbery	Boroughs	76.00	75.83	76.16	55.64	26.56	84.73
	Thiessen	84.00	84.00	84.00	93.46	80.96	100.00
	Census tracts	88.32	87.12	89.53	98.02	94.20	100.00
	Network	99.00	99.00	99.00	99.00	99.00	99.00
Burglary	Boroughs	65.44	64.14	66.75	53.02	25.93	80.12
	Thiessen	75.48	74.19	76.77	97.28	91.83	100.00
	Census tracts	81.22	80.15	82.30	98.92	98.24	99.61
	Network	99.00	99.00	99.00	99.00	99.00	99.00
Theft	Boroughs	72.29	71.12	73.46	83.56	42.57	100.00
	Thiessen	82.00	82.00	82.00	98.77	97.06	100.00
	Census tracts	85.00	85.00	85.00	96.71	87.04	100.00
	Network	99.00	99.00	99.00	99.00	99.00	99.00
Car theft	Boroughs	72.33	71.03	73.62	85.49	53.42	100.00
	Thiessen	78.74	77.60	79.87	98.96	98.22	99.71
	Census tracts	81.48	80.19	82.77	97.98	94.35	100.00
	Network	99.00	99.00	99.00	99.00	99.00	99.00
Vandalism	Boroughs	75.62	74.36	76.87	74.57	32.68	100.00
	Thiessen	83.00	83.00	83.00	98.46	95.52	100.00
	Census tracts	87.00	87.00	87.00	98.90	98.04	99.76
	Network	99.00	99.00	99.00	99.00	99.00	99.00

Table 7.2: Minimum acceptable hit rates (MAHR) estimated for the five crime datasets available for Valencia, along with the lower and upper bounds of the empirical 95% tolerance interval obtained for each estimation, considering the GLM approach and the clusters and outliers derived from LISA indices



(a) (b)

Figure 7.3: Classification of the census tracts of Valencia according to the LISA indices obtained for robberies. Results from the complete dataset (a) and from a simulated 99% hit rate dataset (b). The census tracts whose border is coloured in green in (b) are affected by the removal of a sample of 106 events randomly selected (2% of the whole robbery dataset). The one in the city center is suddenly assigned to the high-high type, whereas the one in the northern area of the city is not detected anymore as a high-high zone

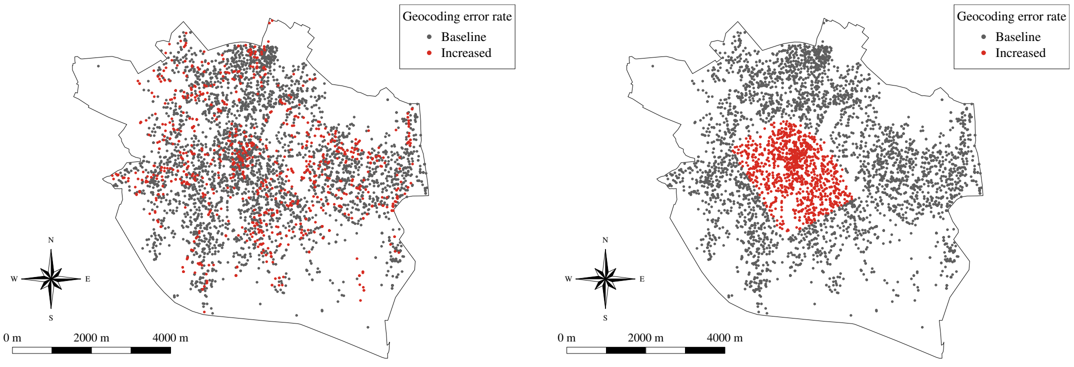
Effect of non-uniform geocoding errors

The crime datasets available for Valencia were reanalyzed considering that the presence of non-geocoded events may not be randomly distributed across the city. Thus, two experiments were performed pretending to represent, according to background experience, two sources of error that can arise during a manual data collection process, leading to lower geocoding hit rates at particular locations of the city (Figure 7.4 displays both experimental settings in the case of the robbery dataset).

First, it is usually observed that the events that take place in certain street types are more susceptible of not being geocoded. Indeed, avenues and squares are especially conflicting street types for geocoding purposes. According to my experience, the occurrence of events in somewhere on an avenue or a square is more likely to produce incomplete descriptions of the exact place where the event took place. A first experimental setting, denoted by Experiment I, was defined to account for this issue in some of the crime datasets studied.

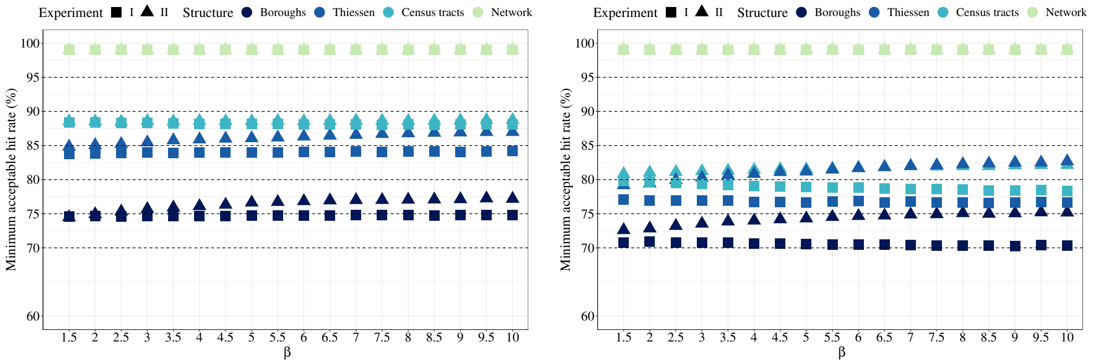
Similarly, a second experiment, referred to as Experiment II, was also carried out. In this case, the bias effect was incorporated to the central district of Valencia (city center) and its two adjacent district. It is also observed that geocoding inaccuracies can arise from the fact that most Police officers are assigned to one district or borough of the city. Hence, some inappropriate practices when collecting the data, that can increase the chances of producing geocoding failures, can be associated to the people carrying out the process. Furthermore, in the context of the city center of Valencia, this source of error could be also a consequence of the higher complexity of this area of the city in terms of road network structure.

Figure 7.5 shows the minimum acceptable hit rates that were reestimated for robberies and car thefts as the β parameter controlling the strength of the non-uniform geocoding error rate was increased from 1.5 to 10 (in steps of 0.5 units). The consequences of increasing β depended on the type of non-uniform geocoding error rate that has been introduced. Thus, whereas the non-uniform geocoding error included in Experiment II increased the estimations obtained for the minimum acceptable hit rate, the effect produced by Experiment I was almost negligible.



(a) (b)

Figure 7.4: Locations of the robberies recorded in Valencia in 2014 and 2015. For experimental purposes, a higher geocoding error rate was assumed for the robberies occurred in avenues and squares (Experiment I) (a) and also for the robberies that took place in the city center and adjacent districts (Experiment II) (b)



(a) (b)

Figure 7.5: Estimations of a minimum acceptable hit rate shown as a function of β (magnitude of the increased error) for the Experiment I and Experiment II, considering the four spatial structures available for Valencia for robberies (a) and car thefts (b)

7.4.2 Simulation studies

Motivated by the initial analysis of the crime datasets available for Valencia, several simulation studies were carried out to better account for some of the factors that may have an impact on the establishment of a minimum acceptable hit rate. Specifically, the intensity and spatial autocorrelation of the point pattern, and the aggregation level of the spatial structure, were the three examined factors. It is worth noting that all these simulation studies were performed considering the Mann-Whitney test as the tool that determines if a complete and an incomplete dataset are equivalent. Thus, according to the results obtained in previous sections, the estimations of minimum acceptable hit rates that were obtained through the simulation studies may be notably low for the execution of certain techniques (cluster detection). However, it was found suitable to limit the analysis to the Mann-Whitney test for the purpose of investigating the effect of the intensity, clustering and aggregation levels on such estimations.

Squared grids consisting of $n \times n$ cells defined over a square representing an area of 64 km^2 were used for all the simulation studies performed. The area of the square was chosen to be similar to the area of Valencia (49.4 km^2). The number of iterations of Ratcliffe's procedure was reduced to 100 for all the computations involved in the simulation studies in order to decrease the cost (the variability observed in the estimations was very low, a fact that guided this decision).

Study I: Effect of the intensity level on the minimum acceptable hit rate

Point patterns of intensity $\lambda \in \{5{\cdot}10^{-6}, 10^{-5}, 5{\cdot}10^{-5}, 10^{-4}, 5{\cdot}10^{-4}, 10^{-3}, 5{\cdot}10^{-3}, 10^{-2}\}$ (measured in 1/m^2) following an homogeneous Poisson process were generated over a 25×25 grid (625 cells) covering the square of 64 km^2 used for all the simulation studies. The use of lower values of λ was discarded because they lead to point patterns with too few events that make the analysis more unreliable (too dependent on the simulated point pattern). Indeed, using a value of, say, $\lambda = 10^{-6}$ implies the construction of a dataset of only 64 points, on average. Hence, the purpose was to estimate a minimum acceptable hit rate for each intensity level following Ratcliffe's methodology.

The minimum acceptable hit rates estimated for each intensity level are shown in Figure 7.6. It is clear that the minimum acceptable hit rate becomes progressively higher as the intensity of the point pattern increases. For an intensity value $\lambda = 10^{-5}$ the estimation is very similar to that provided by Ratcliffe (2004a). However, for orders of magnitude higher than 10^{-5}, the minimum acceptable hit rate needs to be raised, according to this experimental study.

Nevertheless, it is worth noting that many of the real datasets that are analyzed by researchers present intensity levels notably lower than the highest values of λ used in this simulation study. For instance, the five crime datasets from Valencia have an intensity below $1.1 \cdot 10^{-4}$ 1/m^2 when they are considered over the areal representation of the city, and below $7 \cdot 10^{-3}$ 1/m when they are projected into the linear network.

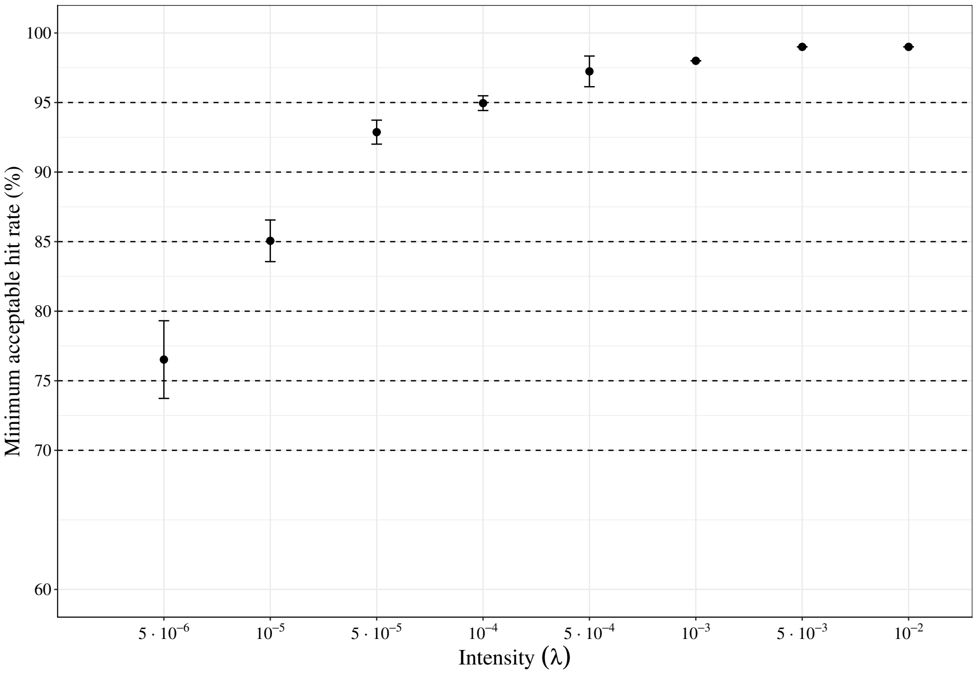


Figure 7.6: Estimations of a minimum acceptable hit rate shown as a function of the intensity (λ) defined for each point pattern generated through an homogeneous Poisson process

Study II: Effect of the clustering level on the minimum acceptable hit rate

Matérn processes were simulated (over the same 25×25 grid used in the previous simulation study) for enabling the investigation of the effect of the clustering level of a point pattern on the determination of a minimum acceptable hit rate. The value of μ involved in the definition of a Matérn process was set to 1, forcing the global intensity of the process ($\lambda \cdot \mu$) to coincide with λ (intensity of the "parent" process). Thus, the intensity values tested (λ) were the same considered for the previous simulation study. Finally, parameter s was varied in the range $]0, 160]$ in order to make possible the generation of point patterns displaying different levels of clustering.

Figure 7.7 shows the results corresponding to this simulation study. It can be observed that the presence of some level of clustering in a point pattern usually leads to lower minimum acceptable hit rates than in the case of patterns lacking a clustered structure (patterns following an homogeneous Poisson process). Indeed, Figure 7.7 suggests that higher clustering levels (lower values of s) are usually associated with lower minimum acceptable hit rates. However, this finding presents a greater uncertainty for point patterns of low intensity ($\lambda = 5 \cdot 10^{-6}$), which make the analysis more dependent on the simulated pattern for which a minimum acceptable hit rate is estimated, rather than on the parameter chosen for its generation (the scale of the Matérn process, s, in this case). On the other hand, for the highest values of λ that were tested (0.005 and 0.01), the effect of the clustering level is completely

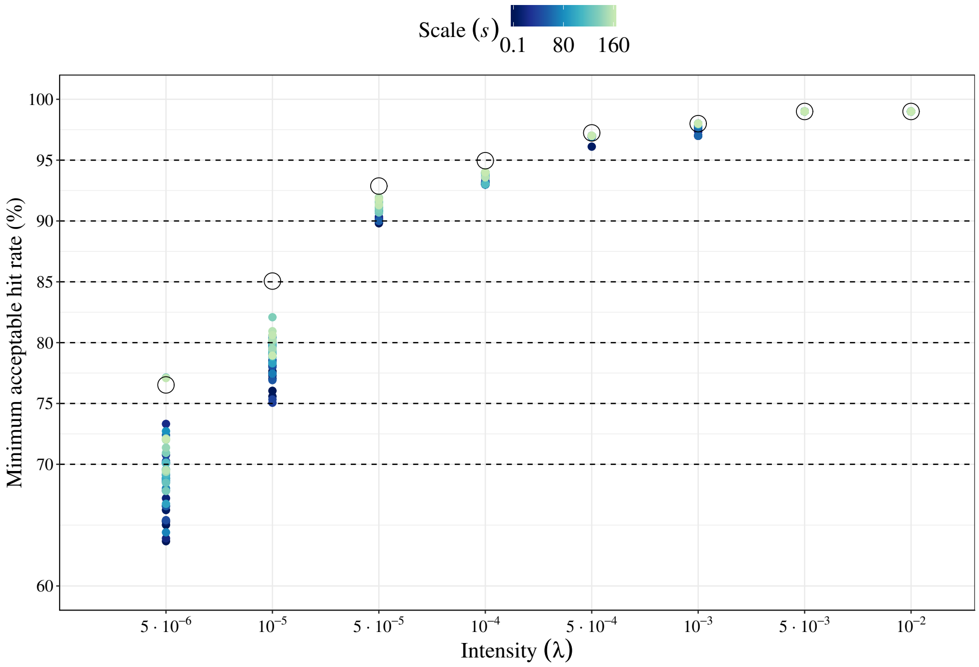


Figure 7.7: Estimations of a minimum acceptable hit rate shown as a function of the intensity (λ) and clusters' radius (s) defined for each point pattern generated through a Matérn process. Empty circles correspond to the values obtained in the Simulation Study I for homogeneous Poisson processes

absorbed by the one produced by the intensity (the minimum acceptable hit rate is not reduced).

In conclusion, the existence of clustering in a point pattern appears to reduce the minimum acceptable hit rate. Hence, achieving the minimum acceptable hit rates that were estimated in the Simulation Study I for totally random patterns should be sufficient, regardless of the presence of an attraction dynamic between the points of the pattern. In other words, the fact that many real spatial patterns show a certain (or even a high) level of clustering can be helpful to reduce, to some extent, minimum acceptable hit rates.

Study III: Effect of the aggregation level of the underlying structure on the minimum acceptable hit rate

The aggregation level of a spatial structure where a point pattern lies is given by the number of units that form the structure. Concretely, reducing the number of spatial units leads to the configuration of more aggregated structures, whereas increasing the number of units implies the opposite. Changing the aggregation level of a spatial domain usually carries some effects on statistical estimations and properties, as shown in Chapter 5.

Hence, the value of n that defines the $n \times n$ squared grids that have been considered for the simulation studies described before (for $n = 25$) was varied from 5 to 50 in steps of 5 units. A point pattern was simulated following an homogeneous Poisson

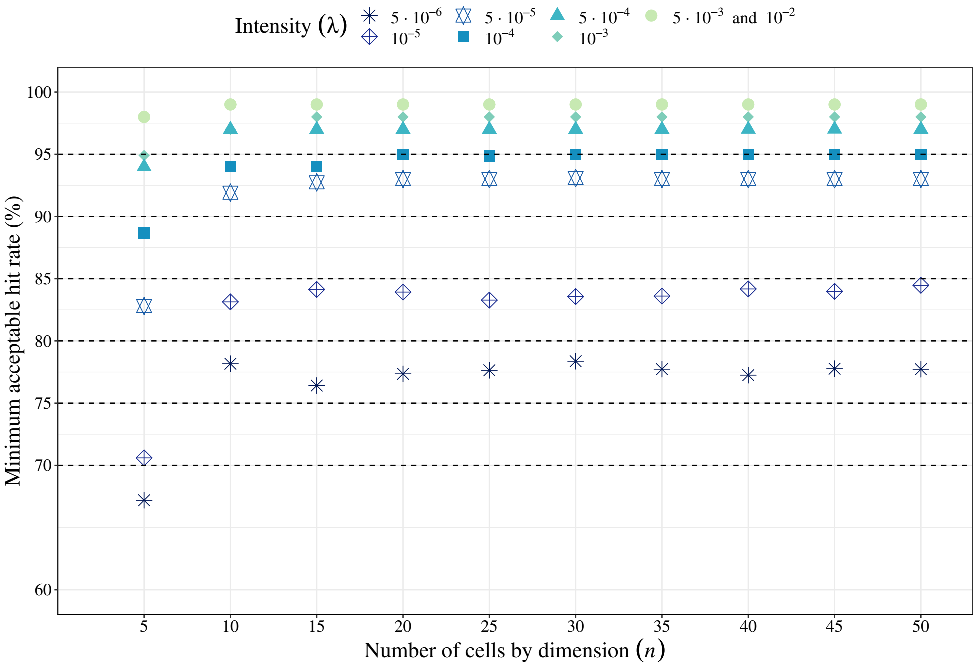


Figure 7.8: Estimations of a minimum acceptable hit rate shown as a function of the number of cells (n) and intensity (λ) defined for each point pattern generated through an homogeneous Poisson process. The values $\lambda = 5 \cdot 10^{-3}$ and $\lambda = 10^{-2}$ are represented with the same colour and shape because all the estimations were coincident. The points corresponding to $\lambda = 5 \cdot 10^{-4}$ and $\lambda = 10^{-3}$ for $n = 10$ also overlap

process for each of the intensities already tested in the previous simulation studies. Then, the methodology for estimating a minimum acceptable hit rate was used considering the $n \times n$ squared grid corresponding to each value of n.

Figure 7.8 indicates that increasing the value of n (that is, creating more disaggregated grids) increases the minimum acceptable hit rate. Furthermore, it is worth noting that this effect is particularly notorious when one increases the value of n within the range $[5, 20]$. After that, the minimum acceptable hit rate seems to become rather stable and unaffected by the increments of n (for all the intensity values being considered).

Providing an estimation of a minimum acceptable hit rate given an intensity value

Finally, a logistic growth equation was used to model the minimum acceptable hit rate (MAHR) as a function of the logarithm (with base 10) of the intensity of the point pattern:

$$\text{MAHR}(\log_{10} \lambda) = \frac{\phi_1}{1 + e^{-(\phi_2 + \phi_3 \log_{10} \lambda)}} \tag{7.2}$$

The upper bounds of the 95% tolerance intervals empirically obtained for the minimum acceptable hit rates estimated in the Simulation Study I were fitted through

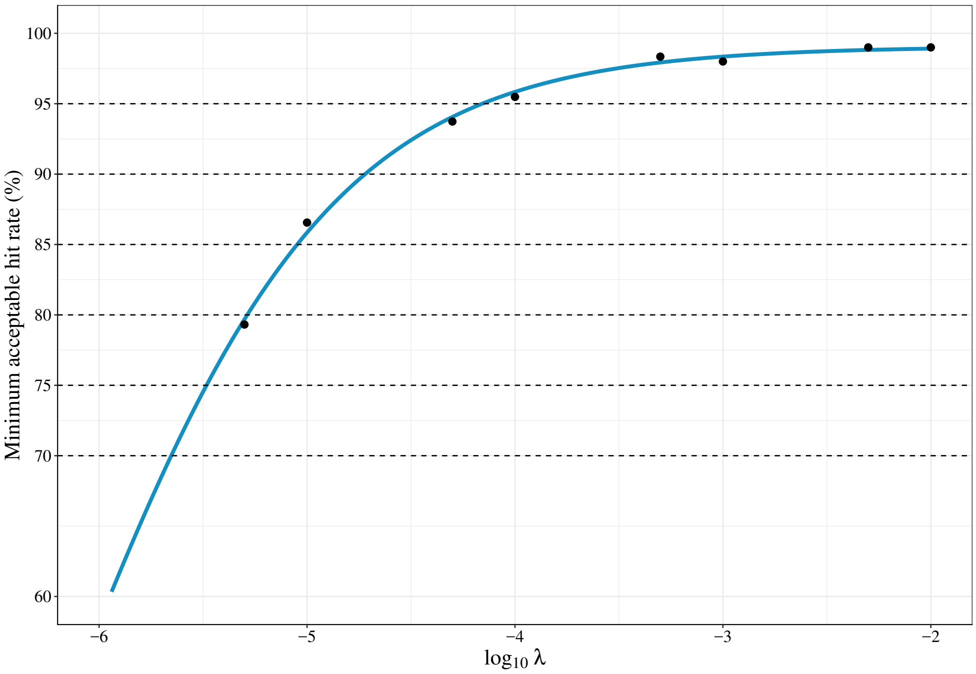


Figure 7.9: Logistic fit (blue curve) of the upper bounds of the 95% tolerance intervals obtained for the estimations of minimum acceptable hit rates obtained. The logistic curve corresponds to the parameters $\phi_1 = 99.07$, $\phi_2 = 9.46$ and $\phi_3 = 1.52$, according to Equation 7.2

Equation 7.2. These values were selected in order to obtain the most conservative approximation possible. The upper bounds corresponding to the 25×25 grid were found suitable enough as increasing the aggregation level from $n = 20$ barely altered the results (Figure 7.8). Table 7.3 includes the estimations that Equation 7.2 provides for the intensity levels used in the simulation studies.

Intensity ($\boldsymbol{\lambda}$)	$5 \cdot 10^{-6}$	10^{-5}	$5 \cdot 10^{-5}$	10^{-4}	$5 \cdot 10^{-4}$	10^{-3}	$5 \cdot 10^{-3}$	10^{-2}
MAHR (%)	79.68	85.85	94.06	95.84	97.93	98.35	98.82	98.91

Table 7.3: Proposals of minimum acceptable hit rates (MAHR) for planar point patterns of certain intensity levels according to the model fit shown in Figure 7.9

7.5 Conclusions

Achieving a high geocoding hit rate when one deals with point-referenced data is essential to perform a reliable statistical analysis. Multiple inadequate practices that may occur during the data collection stage can lead to the emergence of non-geocoded events that become unusable for the analysis. The existence of missing data always implies reducing the statistical power of any statistical test executed. Furthermore, and even more importantly, missing data can increase the risk of bias if the data that is not geocoded follows a non-random pattern. Therefore, it is important to determine a minimum acceptable hit rate that guides researchers, in the sense of making them aware of the geocoding hit rate they should achieve to guarantee, to some extent, a fair analysis of the data.

Overall, the results suggest that the well-accepted 85% hit rate may be insufficient under certain conditions. Indeed, analyzing large datasets or using highly-disaggregated spatial supports are two reasons that should oblige researchers to increase their geocoding hit rates, according to the simulation studies that have been carried out. In particular, road networks, which are becoming especially popular in the last years to conduct spatial analyses, have shown to be particularly sensitive to the presence of non-geocoded data, according to the crime datasets that were analyzed over this type of spatial structure. This result may have been a consequence of the fact that the intensity of the patterns increased when switching from an areal support to a linear one (the network), or the result from having many more spatial units and hence data sparsity. Furthermore, it has been observed that cluster/outlier detection based on LISA indices is more sensitive, by far, to the presence of non-geocoded data. Hence, the minimum acceptable hit rates may need to be raised dramatically when applying some methodologies of this nature. In this regard, it is important noting that providing a “universal” minimum acceptable hit rate seems rather unlikely.

This investigation also presents some limitations and research gaps that could be filled in the future. First, performing a more computationally intensive study would be convenient to provide more accurate estimations of a minimum acceptable hit rate as a function of intensity, clustering or aggregation levels. It is worth noting that the effect of the intensity level on minimum acceptable hit rates has been investigated by considering one simulated dataset for each intensity value tested. Although this strategy enables capturing a relationship between point pattern intensities and minimum acceptable hit rates, the generation of multiple datasets should be taken into account to also control for dataset-depending effects. In addition, the area of the underlying window where the events are located may also require a specific investigation. Second, in view of the remarkable raise that minimum acceptable hit rates experimented when performing cluster/outlier detection through LISA indices (in contrast to the Mann-Whitney test or the GLM approach, which simply indicate if the count distributions of a complete point-referenced dataset and an incomplete version of it are statistically different), a specific investigation on how the presence of non-geocoded data affects the estimation of covariate effects through statistical models may be of special interest. Third, the implementation of simulation studies where non-uniform geocoding errors are systematically introduced is another possible line of research for the future. Finally, it is worth mentioning that point patterns that are characterized by the presence of a regular or a repulsion tendency between the points (in contrast to clustered ones) have not been considered. It is true that these point patterns are more unusual than clustered ones in some fields such as criminology or epidemiology, but they frequently arise in situations of competition between animal or plant species (Stoll and Bergius, 2005).

In conclusion, it is important to highlight again that establishing a minimum acceptable geocoding hit rate heavily depends on the research purposes being pursued or, put more simply, on the statistical techniques that are going to be used. Therefore, it is necessary to provide the same advice that Ratcliffe (2004a) gave: “we should attempt to achieve a hit rate of 100% every time”.

Chapter 8

Software development

Carrying out complex statistical analyses requires the use of advanced statistical software that has already been developed by other researchers. In this regard, the R programming language (R Core Team, 2018) has been used to perform all the studies included in this thesis. At the same time, the singularities of your analysis or data often lead you to design new software. Thus, this Chapter presents two R packages that have been created during the development of this thesis.

8.1 SpNetPrep: Overview of the package

The interest in working at the road level renders some technical difficulties due to the high complexity of these structures, specially in terms of manipulation and rectification. The R Shiny app SpNetPrep, which is available online and via an R package named the same way, has the goal of providing certain functionalities that could be useful for a user which is interested in performing an spatial analysis over a road network structure. The SpNetPrep package does not deal with statistics, but with the previous steps that can be required in order to perform a spatial statistical analysis of a point pattern that lies on a linear network representing a road structure. In this regard, the name chosen for the package summarizes its main goal of "Spatial Network Preprocessing" (SpNetPrep). The main feature provided by the SpNetPrep package is an interactive application that allows to carry out the complete preprocessing of a linear network that comes from a road structure. First, the user needs to install the package via CRAN or via GitHub. Then, the execution of the function `runAppSpNetPrep()` in the R console launches the application allowing its full use, which is also possible to be done online following the link `https://albriz.shinyapps.io/spnetprep/`. If the application is run from the R console, it is necessary to click the option "Open in browser" when it shows, or define "Run external" for the opening of Shiny applications in order to be able to download the modifications performed on the objects uploaded to it.

According to the technical difficulties that the development of a spatial analysis over a linear network implies, the SpNetPrep takes advantage of the R packages leaflet

(Graul, 2016) and shiny (Chang et al., 2018) to provide an intuitive application that helps to reduce such difficulties.

Specifically, the SpNetPrep package focuses on the following parts of the preprocessing process that could be required prior to any spatial analysis over a linear network: network creation and edition, network direction endowment and point pattern revision and modification. Now, the complications that can be derived from each of these steps are discussed.

Users can obtain a road network of their interest via the OpenStreetMaps (OSM) platform (Haklay and Weber, 2008; OpenStreetMap contributors, 2017) or from other public or private sources. In addition, based on OpenStreetMaps, R users can employ the R packages osmar (Eugster and Schlesinger, 2013) and osmdata (Padgham et al., 2017) to get from console the network they wish. However, the obtention of a complete and detailed road network in a properly digitallized format that enables its use in the R framework can become much harder than the finding of administrative divisions in a suitable format, which are usually the base of many spatial statistics studies that deal with the commonly known as lattice data.

When the user is in possession of a road network in a right R format (these formats will be described later), the SpNetPrep application includes a "Network Edition" section that usually would constitute the starting point of the preprocessing phase. At this part of the application, users can introduce their networks in order to delete edges, join vertex to form new edges and create new points that are connected to the preexisting vertex or directly between them. Of course, users that had previously created their road networks with SpNetPrep can use this edition section to make changes on them.

The manual edition (or curation) of a linear network representing a road structure is an important step that must be taken in order to correct possible mistakes (not updated road configurations), remove some undesired parts (pedestrian or secondary roads, depending on the application) and also to simplify some zones of the network whose complexity could obscure the analysis being performed (which is sometimes very notorious in round-abouts or complex intersections).

Furthermore, in view of the difficulties that sometimes can arise when trying to obtain a road network structure, the "Network Edition" section of the SpNetPrep application could also be employed to create a road network from scratch (the user would need to upload a dummy road network of at least one segment within the region of interest). Of course, this would not be a good option if the aim of the user is the creation of a complex road network made of hundreds of kilometers, but it can be a cost-effective option for creating small road networks within an urban area, or even for a long network representing highways or rural roads given its (usually) greater simplicity.

Another important question to take into consideration when working with a linear network structure is its directionality. Depending on the kind of dataset being treated, network direction could be of no interest, but this should not be the case when analyzing traffic-related data. In fact, traffic flow could be dramatically influential for some classical spatial analysis that arise from this kind of data. For example, in order to fit a spatial model to a collection of accident counts at the road segment level (for instance, with the spdep package from Bivand and Piras (2015)), the provision of a directionality to the linear network would become essential to define a realistic neighbourhood structure. In a similar way, if a geostatistical approach was established (see the gstat package, from Pebesma (2004)) to predict a quantitative measure along a road network which is likely affected by traffic flow (as the level of pollution), the lack of consideration of the directions that can be taken by the vehicles that use the network could lead to meaningless results.

Again, it is not easy, at all, to find the information required to endow a linear network based on a road structure with a directionality. The network structures available in OSM contain some information regarding the direction of the streets and some cartographic platforms include the direction of traffic (measured in angles) at some points of the structure, but, in general, it can be really hard to obtain such information for a road network of your interest. For this reason, the "Network Direction" section of the SpNetPrep application attempts to facilitate the enhancement of a network with this valuable information.

Once the network structure is properly curated and endowed with a direction (if necessary), a point pattern can be formed along the network structure from a dataset containing geocoded information. In the case the information on the location of each event is in the form of a postal address, the R package ggmap (Kahle and Wickham, 2013b) can be very useful by providing an R interface that allows geocoding via the Google Maps API.

Then, when the coordinates of the events of interest are already available, regardless of the way they have been obtained, it is time to project them into the linear network. This step can be achieved straight by using the (shortest) orthogonal projection of each pair of coordinates into the linear network, for example with the `project2segment` function of the R package spatstat (Baddeley et al., 2015). However, depending on the level of accuracy of the coordinates, this process can lead to wrongly locate some of the events on the linear network. If the coordinates have been derived from postal addresses, one could find the points of the network whose projection distance (from their imputed coordinates to their projected) exceeds some threshold that could be indicating that something went wrong (short projection distances should be admitted, as the linear network is only a representation of the real space where the events of interest take place). Therefore, the "Edit a Point Pattern" section of the application allows the user to investigate this issue. While providing a whole picture of the distribution of the point pattern in the road network being studied, this section of the application offers the user information on the obtained projection distances remaining to the decision of the user if further revision of the geocoding process is required, or if there is a part of the road structure that is not well represented by the linear network in use.

As a summary, Figure 8.1 contains a workflow diagram that indicates the four steps that should be considered before performing an spatial analysis of a point pattern on a linear network.

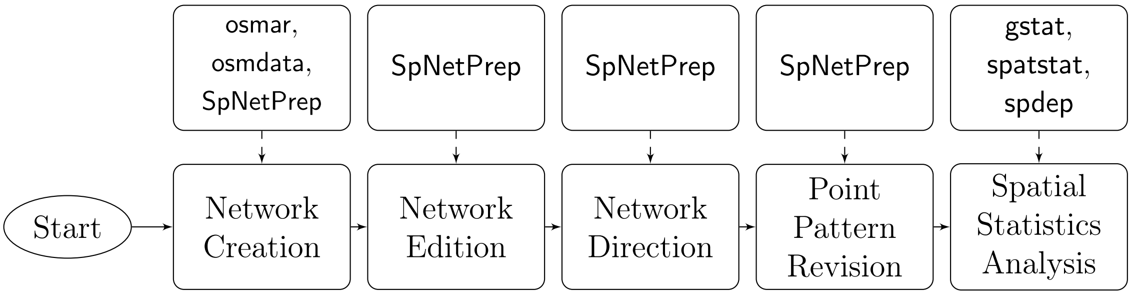


Figure 8.1: Workflow that describes all the steps that could be carried out in order to perform a spatial statistics analysis on a point pattern that lies on a linear network. Some of these steps which lead to the final statistical analysis may be skipped but, at least, all of them should be considered. The blocks pointing the steps of the process include some of the R packages that would allow to successfully achieve each of them

8.1.1 Technical issues

Linear networks in R

There are two main classes coexisting in R that represent what a linear network is: `SpatialLines` and `linnet`. The class `SpatialLines` belongs to the sp package (Pebesma and Bivand, 2005), whose capabilities are fully described in Bivand et al. (2013). In case it is needed to attach some information to the edges that form the linear network, the class `SpatialLinesDataFrame` from the same package should be used instead. On the other hand, the class `linnet` is part of the spatstat package, focused on spatial statistics analysis and modelling. Hence, both object classes can be used effectively to manage spatial linear networks in R, although it is required to choose the `linnet` format in order to get the capability of using all the statistical methods for linear networks implemented in the spatstat package.

There are several functions in R that facilitate the conversion between `SpatialLines` and `linnet` objects. Specifically, `as.linnet.SpatialLines` from the maptools (Bivand and Lewin-Koh, 2017) R package converts `SpatialLines` into `linnet` objects, whereas the double application (in this order) of the `as.psp` and `as.SpatialLines.psp` functions of the spatstat and maptools packages, respectively, enable the conversion from a `linnet` object into a `SpatialLines` one.

Geographical projections

Coordinate Reference Systems (CRS) are essential to locate entities in space. Concretely, each CRS defines a specific map projection that unambiguously determines the location of every point on the Earth, which logically makes impossible to work with two geographic objects that are described in a different projection.

The usual longitude and latitude coordinates, which range from -180° to 180° and -90° to 90°, respectively, correspond to the WGS84 (World Geodetic System 1984) geographical projection. One important characteristic of the WGS84 projection system is that it considers the whole world as a unique zone, that is, a pair longitude-latitude in this CRS system determines only one point of the Earth. However, this situation does not hold for the Universal Transverse Mercator (UTM) projection system, another well known CRS that divides the world into 60 zones whose coordinates are denoted easting and northing in analogy with longitude and latitude (respectively). The use of the UTM system is more convenient for performing statistical analysis given its higher level of accuracy (specially when working with small areas) and also because the coordinates it provides are expressed in meters, which renders very easy to compute distances.

The sp package allows to deal with projection systems by means of the CRS class and the proj4string method. The following lines exemplify how to proceed, assuming that wgs84object and utmobject are two sp objects expressed in WGS84 and UTM (zone 30) coordinates, respectively, whose projections had not been established yet in the R environment. Basically, the proj4string assigns a projection system to an sp object whereas the spTransform function changes an sp object's projection from one system to the other.

```
> CRS_wgs84 <- CRS("+proj=longlat +datum=WGS84
                   +ellps=WGS84 +towgs84=0,0,0")
> CRS_utm <- CRS("+proj=utm +zone=30 ellps=WGS84")
> proj4string(wgs84object) <- CRS_wgs84
> proj4string(utmobject) <- CRS_utm
> wgs84object_transform <- spTransform(wgs84object, CRS_utm)
> utmobject_transform <- spTransform(utmobject, CRS_wgs84)
```

Working with files in the SpNetPrep application

Format .RDS has been chosen for all the files possibly involved during the use of the SpNetPrep application, which means that inputs and outputs will always be in this R file format. Functions readRDS and saveRDS allow to read and create, respectively, a .RDS file for its use in the application or in the usual R console.

On the CRS system, the SpNetPrep application is only ready to accept input files from that are expressed in UTM coordinates. These coordinates are then internally converted into the WGS84 system in order to be usable by the leaflet functions that are employed for making the application work. Consequently, the output files that can be downloaded after the use of any of the sections of the application are also in UTM coordinates, allowing its direct use for statistical analysis if no more preprocessing steps are required.

Furthermore, for the "Network Edition" and "Network Direction" parts of the application the inputs are required to belong to the sp package, whereas for "Point Pattern Revision" it is needed to upload an object that has been created with the lpp function of spatstat (more details later). The UTM zone needs to be specified by the user with the `proj4string` method in the case of the `sp` objects and typing it on the corresponding text input for the "Point Pattern Revision" section of the application as otherwise the application will yield an error message. In case of trouble during the construction of the input files, the data objects `SampleNetwork`, `SampleDirectedNetwork` and `SamplePointPattern` available in the package can serve as a reference for the sections "Network Edition", "Network Direction" and "SamplePointPattern", respectively, although the first of these data objects also works for the "Network Direction" part.

Finally, it is convenient to remark that, even though the application has been subject to the usual debugging tests, the raise of an error could break the application and make users lose their work. For this reason, it is highly recommended to execute and download the changes being performed in the road network or point pattern being used regularly.

8.1.2 Network edition

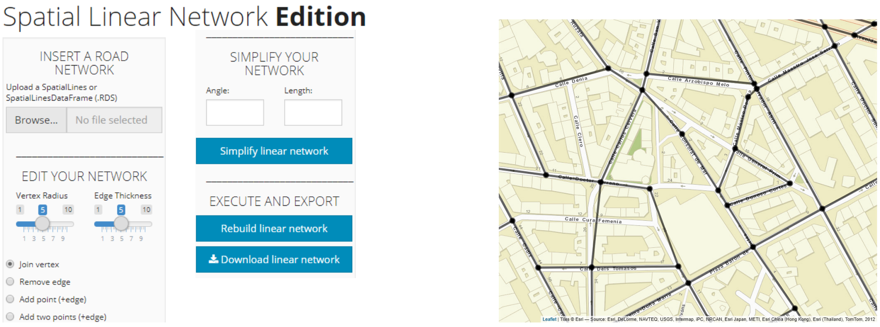
Manual edition of the geometry of a linear network is one of the main purposes of the SpNetPrep application. This process includes the manual rectification of the network, which basically consists of performing edge addition/deletion and vertex addition/deletion. Furthermore, the application provides an algorithm of automatic simplification that reduces network's complexity while accounting for its basic geometric structure, which will be later described. First, the use of the "Network Edition" section of the application is explained.

Application's use

There are four basic actions that can be performed for editing the linear network manually: "Join vertex", "Remove edge", "Add point (+edge)" and "Add two points (+edge)". The user only has to select the more convenient option and proceed intuitively. If "Remove edge" is selected, the click on an edge of the network (anywhere all along its length) serves to mark the edge in red, indicating a removal state. Oppositely, by choosing any of the options "Join vertex", "Add point (+edge)" or "Add two points (+edge)" the user needs to click on two points of the map accordingly to the option being selected. For the "Join vertex" option, two vertex must be clicked, whereas for the "Add two points (+edge)" two points of the map (that are not vertex) have to be clicked. Finally, the "Add point (+edge)" requires that the user clicks on a point and on a vertex of the network (in this order). All these three options that imply the addition of edges (and maybe vertex) to the road network are marked in green. The click of the button "Rebuild linear network" makes this manual editions effective and when the map refreshes the new (edited) road network is available for the user (which can be downloaded by clicking on the button available at the bottom of the application).

Now, let's see and example of use of the "Network edition" section of the application (Figure 8.2a includes a picture of the main elements of the application, although this does not correspond to the real distribution on-screen of these elements).

First, the user should introduce a road network (in the form of a `SpatialLines` or `SpatialLinesDataFrame` object) whose segments are expressed in UTM coordinates. Then, zoom can be increased to focus on a little part of the network, involving a little number of roads (Figure 8.2b).



Spatial Linear Network **Edition**

INSERT A ROAD NETWORK

Upload a SpatialLines or SpatialLinesDataFrame (.RDS)

No file selected

EDIT YOUR NETWORK

Vertex Radius: 5 (1–10)

Edge Thickness: 5 (1–10)

● Join vertex
○ Remove edge
○ Add point (+edge)
○ Add two points (+edge)

SIMPLIFY YOUR NETWORK

Angle:

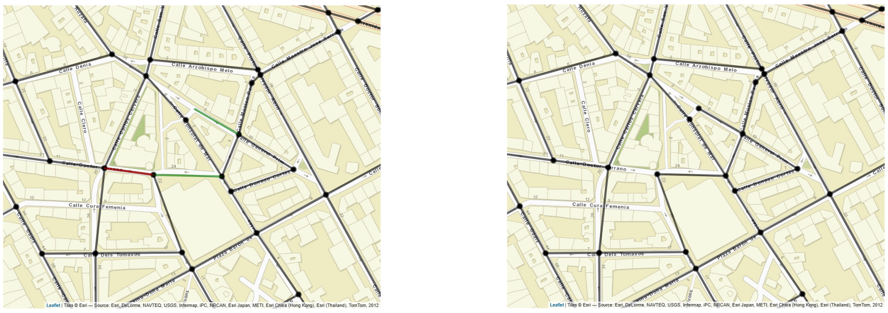
Length:

EXECUTE AND EXPORT

(a) (b)

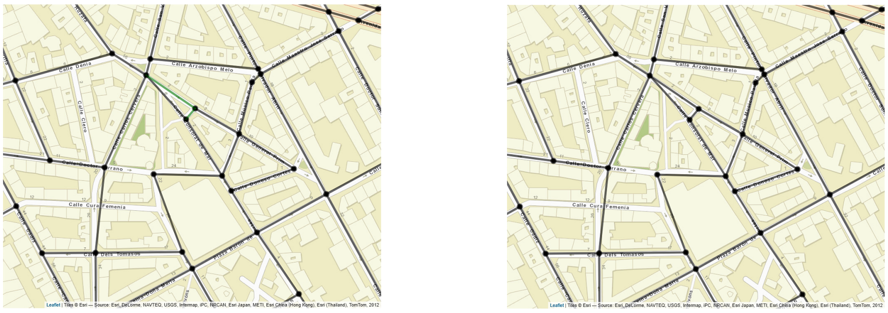
Figure 8.2: Overview of the "Network Edition" section of the SpNetPrep application (a) and example of a road network uploaded into the application (b)

At this point, the user could find something to rectify or change, which could finally derive into the road structure displayed in Figure 8.4a (differences from Figure 8.2b can be appreciated at the central part of the image). To arrive to this new network structure, the user would have to perform several steps with the SpNetPrep application, including edge removal, vertex connection and the addition of a new point that connects to a preexisting vertex (Figure 8.3a). Then, network should be rebuilt (Figure 8.3b) allowing the definition of two new vertex connections (Figure 8.4a) that finally lead to the network in Figure 8.4b. It has to be noted that the execution of the connections in Figure 8.4a require the rebuilt of the network as in Figure 8.3b, otherwise the new point "drawn" in Figure 8.3a would not be available to create these two new connections.



(a) (b)

Figure 8.3: Use of the "Join vertex" (in green), "Remove edge" (in red) and "Add point" options (in green) in the SpNetPrep application (a), and resulting network after pressing "Rebuild linear network" (b)



(a) (b)

Figure 8.4: Another use of the "Join vertex" (in green) option of the "Network Edition" section (a) that leads to the final corrected road structure in (b)

Automatic network simplification

The `SpNetPrep` package includes a function called `SimplifyLinearNetwork`, which is also provided in the sidebar panel of the "Network Edition" section of the application, that could be very helpful during the network preprocessing process. This function (which accepts and produces `linnet` objects) consists in the execution of an algorithm that attempts to automatically reduce network's complexity without altering its basic geometric configuration. The main objective of the algorithm is to merge the pairs of edges of the network that are connected by a second-degree vertex (with only two incident edges) into only one edge. Equivalently, this action means to join two vertex of the network whose path of connection only passes through another vertex of the network.

Two are the parameters that control the extent to which this algorithm simplifies the linear network: edge `Length` and `Angle` between edges. The tuning of these two parameters allows the user to test several simplifications of the network that imply different levels of conservation of its geometric structure. Both parameters work in the same direction: merging between two edges only produces if their lengths (of both) are lower than `Length` and if the angle they form is below the value of `Angle`. The continuous increase of `Length` and `Angle` can derive into a very simplified network (with a minimal number of vertex and edges), but this process has the cost of producing a geometric structure much more dissimilar to the original one.

More specifically, an analysis on the choice of `Length` and `Angle` was performed by using a road network from the city of València (Spain). The `Angle` parameter was varied from 0° to 90°, whereas `Length` made it from 0 m to 500 m. The level of simplification achieved with every combination of the parameters was measured in terms of the percentage of second-degree vertex that were removed by merging their two incident edges, which is the objective of the algorithm. The Hausdorff distance (Huttenlocher et al., 1993) was then used to measure the geometric dissimilarity existing between the original road network and each of its simplifications, which was achieved with the aid of the `gDistance` function of the `rgeos` package (Bivand and Rundel, 2018b). Hence, the ratio between the level of simplification obtained and the geometric dissimilarity produced was computed for every combination of parameters as a measure of efficiency for the algorithm. As a guide, a choice of `Angle` between 15° and 30°, and a `Length` from 50 m to 80 m suggested optimality (the optimal was attained for an `Angle` of 22.5° and a `Length` of 65 m) for the network tested, although these parameters could be less convenient for other road networks. As an illustration, Figure 8.5 shows the application of the `SimplifyLinearNetwork` function of the package with an `Angle` of 25° and a `Length` of 65 m. A reduction in the number of two-degree vertex of the network can be clearly appreciated from Figure 8.5a to Figure 8.5b.

Furthermore, for users that fail to determine a value for `Angle` and `Length` that work properly in all kind of roads of their network of interest, the algorithm allows the specification of several values for the `Angle` parameter dependent on the value of `Length`. For instance, one could wish to simplify extremely short edges of the network connected by a two-degree vertex, even though the angle they form is quite abrupt. In this case, the user can establish a threshold of 40° if both edges are shorter than 10 m, and a threshold of 25° otherwise. This information can then be passed to the algorithm function with a 2×2 matrix as the following, which would be created in the R console with the basic instruction `M <- matrix(c(10,60,40,25),nrow=2)`.

$$\begin{array}{rcl} & \begin{array}{cc}\text{Length} & \text{Angle}\end{array} & \\ M = & \left(\begin{array}{cc} 10 & 40 \\ 60 & 25 \end{array}\right) & \begin{array}{l}\text{Condition 1}\\ \text{Condition 2}\end{array}\end{array}$$

For practical reasons, the use of a combined condition for the `Angle` and `Length` parameters is only available in the `SimplifyLinearNetwork` of the package, which can be used from the R console. The "Network Edition" only includes an option to perform the simplification procedure with a global value for `Angle` and `Length`. At this section of the application, one can alter the values of these parameters and explore the results that produce, but the deeper employment of the algorithm (possibly including the use of the `gDistance` function to measure geometric dissimilarity) requires to be in the R console.

The following lines include an application of the `SimplifyLinearNetwork` function to the `SampleNetwork` available in the package including both, a unique value for `Angle` and `Length` and a combined effect of these parameters. The parameters of the `SimplifyLinearNetwork` function are (in this order): `network`, `Angle`, `Length` and `M`, which are not always specified in the next code for preserving the margins. As it can be seen, the direct use of `Angle` and `Length` leads to a superior simplification of the network (less edges), but some users could be particularly interested in the simplification of pairs of very short edges that meet in a two-degree vertex, which is accounted if the `M` matrix is used.

```
> network <- as.linnet(SampleNetwork)
> network
Linear network with 1664 vertices and 2513 lines
```

```
> simplified_network_1 <- SimplifyLinearNetwork(SampleNetwork,25,65)
> simplified_network_1
Linear network with 1598 vertices and 2447 lines
> M <- matrix(c(10,60,40,25),nrow=2)
> simplified_network_2 <- SimplifyLinearNetwork(SampleNetwork,M=M)
> simplified_network_2
Linear network with 1639 vertices and 2488 lines
```

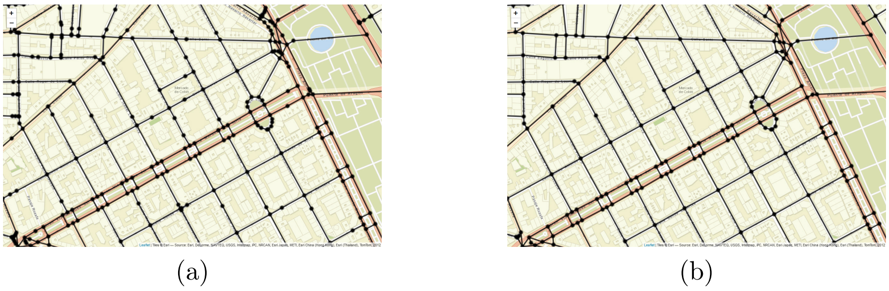

(a) (b)

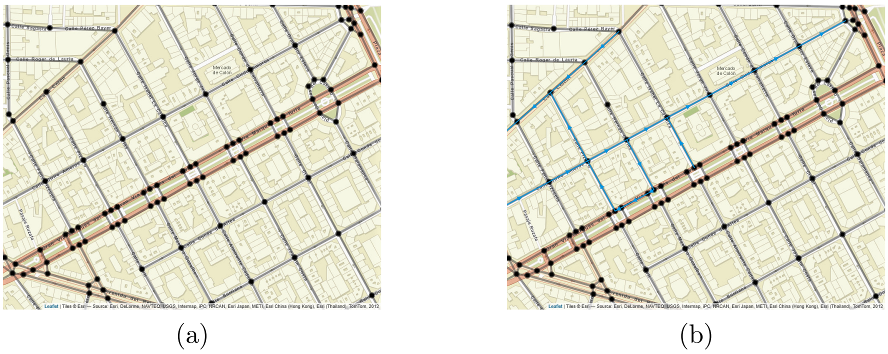
Figure 8.5: A road network introduced as input (a) and its simplified version after the application of the `SimplifyLinearNetwork` function with `Angle` = 25 and `Length` = 65 (b)

8.1.3 Network direction

Application's use

The "Network Direction" section of the application allows the user to endow the network with a direction according to traffic flow, which is facilitated by the presence of arrows indicating this information in the OSM layers. The option "Add flow" enable users to define a flow along the network by simply clicking on the two connected vertex that form the edge they want to give direction to (first click on the origin, second on the end, according to traffic flow). Analogously, "Remove flow" performs the opposite action by removing a direction previously defined, which requires to select the two vertex that form the road segment whose direction is being eliminated (the order of the selections is not important). The function `addFlows` of the leaflet package overlays a blue arrow on the map when a direction is set, and also erases it when the user decides to undo the defined direction (Figure 8.6).

Even though the "Add flow" and "Remove flow" options are sufficient to give direction to the whole network, "Add long flow" and "Remove long flow" attempt to save some time to the user. These functions take advantage of the `shortespath` function that was used to generate Chapter 17 of Baddeley et al. (2015) and is available in `http://book.spatstat.org/chapter-code/R/`. This function finds the vertex that have to be passed through to travel the shortest path that joins two given vertex of the road network, considering that all possible paths along the road network can be taken (without accounting for traffic flow). Therefore, the "Add long flow" option of this section of the application allows to define directionality, at once, for



(a)

(b)

Figure 8.6: A zone of a road network introduced as an input in the "Network Direction" section of the SpNetPrep application (a) and manual addition of traffic flow to the network by using the options "Add flow" and "Add long flow" (b)

all the edges that form the shortest path between two vertex that are not directly connected by an edge. This option can then be very helpful to provide direction to a set of road segments that follow a specific direction that are, simultaneously, the shortest path that joins two vertex of the road network. For example, long avenues can be properly endowed with a direction at once with this option of the application, but it can easily lead to mistakes if it is tested between vertex of the network that require to take more intricate paths to make them connected. Anyway, the "Remove long flow" option allows to undo a wrong direction assignment produced by "Add long flow".

Direction information storage

The directionality of the linear network is stored in the form of a `data.frame` with three columns named V1, V2 and Dir. For each edge of the network, i, V1 and V2 contain the indexes of the vertex of the network that define edge i (origin and end, according to the way the network was defined or created, which can be meaningless in terms of traffic flow). This `data.frame` is then attached to the `SpatialLines` introduced by the user, or added to the existing `data.frame` if the input is a `SpatialLinesDataFrame` object. Obviously, users that have already used the "Network Direction" section of the application with a specific road network only have to upload it again in order to make editions to its directionality, and the V1, V2 and Dir columns of the `data.frame` will be modified accordingly.

There are four possible values for the Dir column: 0, -1, 1 and 2. A value of 0 indicates the absence of a direction for an edge, 2 means double way direction, 1 that direction exists from vertex in column V1 to vertex in column V2 and -1 just the opposite (from vertex in V2 to vertex in V1). For example, the following lines include an example of such a `data.frame`, which describes the minimal linear network (with five edges and six vertex) represented in Figure 8.7.

```
V1 V2 Dir
 1  2   2
 1  3  -1
 2  4   1
 1  5   1
 2  6  -1
```

Taking this information into account, users can establish neighbouring relationships between the road segments of their networks that respect traffic flow (employing functions from the spdep package, for example) or compute distances between points that really represent the way vehicles move along the network.

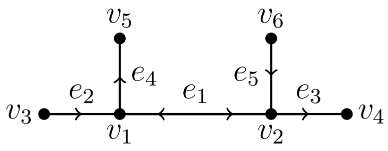


Figure 8.7: Example of a linear road network following usual notation for the edges (e_i) and vertex (v_i). Arrows represent the direction of traffic flow

8.1.4 Point Pattern Revision

Application's use

A point pattern that lies on a linear network can be created in R with the `lpp` function of the spatstat if a spatial point pattern (ppp class) and a linear network object (`linnet` class) are available. The use of the function `marks` from the same package allows to add several informative variables to each point of the pattern.

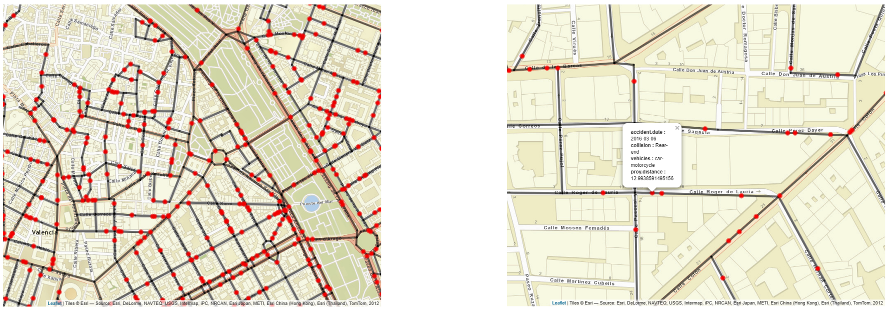
It is always useful to have the possibility of visualizing such a point pattern, which can be done with the "Point Pattern Revision" section of the SpNetPrep application keeping the default option "Explore pattern" (if the pattern is marked, the values of the first ten marks, following the definition of the object, are shown when clicking each event). First, visualization usually provides a better understanding of the point pattern which can condition the posterior statistical analysis, but also allows to check that the creation of the point pattern on the linear network produced correctly.

For illustrating this section of the application, Figure 8.8a shows a Poisson generated point pattern simulating traffic accidents occurred on the road network of the city of València (generated with `rpoislpp` from spatstat with an intensity value of 0.01). Several marks were created randomly and attached to the point pattern to make possible the emergence of a popup message when clicking on a point of the pattern (Figure 8.8b).

Edition of the point pattern

It has already been mentioned that the automatic creation of a point pattern that lies on a linear network in R implies the orthogonal projection of a collection of geocoded events into the network. This operation generally leads to an accurate representation of the observations, but it can produce some misplaced events along the road network.

As it is suggested in Figure 8.8b, one could compute the projection distance (by extracting the `d` component of the list that the `proj2segment` function returns) from each geocoded event to the linear network and include it as a mark of the pattern in order to facilitate the inspection of this operation (here, the projection distance mark was also generated randomly from a uniform distribution only for illustration, as the Poisson generated pattern lies exactly on the network). In addition, as another preventive measure, in case the coordinates of the events came from the geocoding of postal addresses, these could also be added as a string mark that allows the user to better check if this process went straight.


Valencia
Calle Don Juan de Austria
Calle Sagasta
Calle Roger de Lauria
Calle Mossen Femadés
Calle Martinez Cubells
Leaflet | Tiles © Esri — Source: Esri, DeLorme, NAVTEQ, USGS, Intermap, iPC, NRCAN, Esri Japan, METI, Esri China (Hong Kong), Esri (Thailand), TomTom, 2012

accident.date : 2016-03-06
collision : Rear-end
vehicles : car-motorcycle
proy.distance : 12.9938591495156

(a) (b)

Figure 8.8: An example of a point pattern that lies on a road network as it can be visualized in SpNetPrep (a), and information that is displayed (marks of the point pattern, if available) when an event is clicked

Given the fact that some points of the pattern could be projected into the wrong place after the simple application of the orthogonality criterion, the "Point Pattern Revision" part of the application includes an option "Move event" that allows users the change the location of the event within the network. A first click on the event to be moved and a second on the exact position of the network where it should be (so the user needs to click on the corresponding edge of the network and not on the underlying map) is enough to perform this task.

8.1.5 Summary

This Chapter has presented the main functionalities and purposes of the SpNetPrep R package. Mostly based on a shiny application that makes use of the leaflet library, SpNetPrep allows users to carry out the complete preprocessing of a linear network that represents a road structure, as a previous step to the execution of a spatial (or spatio-temporal) statistics analysis.

The use of linear networks is becoming popular in recent times to provide more realistic investigations of many events of interest that take place along road structures. However, dealing with linear networks can be quite more complicated than using other typical spatial structures, which in some extremes cases could even lead to discard its use.

The SpNetPrep application is then divided into three sections that attempt to reduce the difficulties that associate to the most common issues that arise when working with linear networks that represent road structures. First, the availability of road networks in the right format is sometimes scarce or not of enough depth to satisfy the necessities of the researchers. Second, linear networks that represent road structures can present both mistaken and excessively complex road segment configurations. The "Network Edition" section of the application provides several tools to try to overcome these two main difficulties, including an algorithm that automatically simplifies the network accounting for its geometric shape.

Another important step to perform before the execution of a spatial analysis over a linear network is the revision of the point pattern that is being employed. Point patterns on linear networks are commonly built by applying the orthogonal projection of a set of coordinates into the linear network. Even though this can work well most of the times, the excessive simplifications of the road structure or the inaccuracies derived from the geocoding of the events could cause serious alterations in the pattern. The section "Point Pattern Edition" allows users to inspect and correct a point pattern that lies on a network.

Finally, the SpNetPrep application includes a more specific part about "Network Direction". The tools available in this section of the application enable users to endow their whole road network according to traffic flow. This task could be really costly if the network is of considerable dimensions, but the value it can provide to some particular statistical analysis should make it worth it.

8.2 DRHotNet: Overview of the package

Hotspot detection basically consists in finding zones of a space where certain event is highly concentrated. There exists a wide variety of methods in literature that allow researchers to identify hotspots at a certain level of accuracy or spatial aggregation. Some of them have been massively used in the last decades, including certain local indicators of spatial association such as LISA (Anselin, 1995) or the Getis-Ord statistic (Getis and Ord, 1992), and the spatial scan statistic (Kulldorff, 1997). The first two of these methods are implemented in the R package spdep (Bivand et al., 2013), whereas the scan statistic is implemented in DCluster (Gómez-Rubio et al., 2005) (although there are other R packages that also provide an implementation of these methods). Furthermore, many new R packages focused on hotspot detection have been released in the last years, most of them being model-based and oriented to disease mapping studies that are carried out over administrative (areal) units (Allévius, 2018; Gómez-Rubio et al., 2019; Meyer et al., 2017).

However, the analysis of certain types of events requires detecting hotspots at a level of spatial accuracy greater than that provided by administrative or regular areal units. Indeed, many research studies of the fields of criminology (Andresen et al., 2017; Weisburd, 2015) and traffic safety (Briz-Redón et al., 2019e; Nie et al., 2015; Xie and Yan, 2013) that have been published in recent years were entirely carried out on road network structures rather than on administrative units. More specifically, some quantitative criminologists have estimated that around 60% of the total variability in crime incidence occurs at the street segment level (Schnell et al., 2017; Steenbeek and Weisburd, 2016), a fact that shows the essentiality of using road segments instead of areal structures to capture the spatial concentration of certain events more properly.

Fortunately, road network structures were introduced in the context of spatial statistics some years ago, providing the basis for analyzing events lying on such structures, which are usually referred to as linear networks. Indeed, the investigation of spatial patterns lying on linear networks is gaining attention in the last years. The design of new and more accurate/efficient kernel density estimators (McSwiggan et al., 2017; Moradi et al., 2018, 2019; Rakshit et al., 2019b), the introduction of graph-related intensity measures (Eckardt and Mateu, 2018), the construction of local indicators of spatial association (Eckardt and Mateu, 2017), or the estimation of relative risks (McSwiggan et al., 2019) are some topics that have recently started developing for linear networks.

Besides the theoretical requirements and advances, it is worth noting that using linear networks for carrying out a spatial or spatio-temporal analysis entails certain technical difficulties. In this regard, the R package spatstat (Baddeley et al., 2015) provides multiple specific functions that allow R users to carry out statistical analyses on linear networks. Furthermore, efforts are constantly being made to reduce the computational cost of adapting certain classical spatial techniques to the singular case of linear networks (Rakshit et al., 2019a).

Despite the existing necessity of analyzing point-referenced data coming from certain fields of research at the street level, there are not many software tools fully designed for hotspot detection on road networks. One relevant contribution in this regard is the KDE+ software (Bíl et al., 2016), although this is not integrated in R. The package DRHotNet, which has been created during the development of this thesis, is specifically prepared for allowing R users to detect differential risk hotspots (zones where a type of event is overrepresented) along a linear network. The methodology behind the functionalities of DRHotNet has been already presented in Chapter 3 of this thesis. Next section shows how this package can be used.

8.2.1 Using DRHotNet

This section shows the complete use of the DRHotNet package with a dataset of crime events recorded in Chicago (Illinois, US). The computation time is indicated for some of the steps (measured in real time for the user). An Intel(R) Core(TM) i7-6700HQ CPU @ 2.60GHz processor was used in this case.

First of all, the following R libraries have to be loaded to reproduce the example:

```
> library(DRHotNet)
> library(lubridate)
> library(maptools)
> library(raster)
> library(rgeos)
> library(sp)
> library(SpNetPrep)
> library(spatstat)
> library(tigris)
```

Downloading and preparing the linear network

The examples provided in this section fully employs open data available for Chicago. First, geographic data from Chicago was downloaded from the United States Census Bureau through package tigris (Walker, 2016). Specifically, census tracts and the road network of the state of Massachusetts were loaded into the R console. The function `intersect` from the package raster (Hijmans, 2019) can be used.

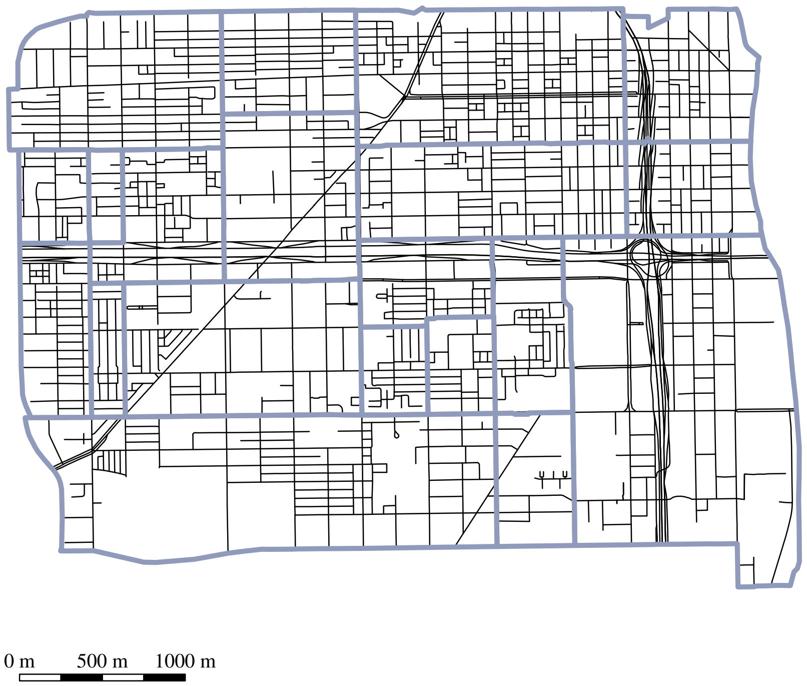


Figure 8.9: Road network (in black) corresponding to the Near West Side Community Area of Chicago. Census tracts of the area are overlayed in blue.

```
> cook.tracts <- tracts(state = "Illinois", county = "031")
> class(cook.tracts)
[1] "SpatialPolygonsDataFrame"
```

```
attr(,"package")
[1] "sp"
> cook.network <- roads(state = "Illinois", county = "031")
> class(cook.network)
[1] "SpatialLinesDataFrame"
attr(,"package")
[1] "sp"
```

The objects cook.tracts and cook.network are composed of 1319 polygons and 77698 lines, respectively (as of the end of September 2019). Now, both objects are used to construct a smaller road network that corresponds to the Near West Side Community Area of Chicago.

```
> names.tracts <- as.character(cook.tracts@data[,"NAME"])
> select.tracts <- c("8378","2804","8330","2801","2808","2809","8380",
                     "8381","8331","2819","2827","2828","8382","8329",
                     "2831","2832","8333","8419","8429","2838")
> cook.tracts.select <- cook.tracts[which(names.tracts%in%select.tracts),]
> chicago.SpLines <- intersect(cook.network, cook.tracts.select)
> length(chicago.SpLines)
[1] 1116
```

Object chicago.SpLine (SpatialLinesDataFrame) has 1116 lines. Then, this object's coordinates are converted into UTM (Chicago's UTM zone is 16):

```
chicago.SpLines <- spTransform(chicago.SpLines,
                               "+proj=utm +zone=16 ellps=WGS84")
```

Now the corresponding linnet object is created:

```
> chicago.linnet <- as.linnet(chicago.SpLines)
> chicago.linnet
Linear network with 9431 vertices and 10559 lines
Enclosing window:
    rectangle = [442563.5, 447320] x [4634170, 4637660] units
```

It is worth noting how the transformation of the network into a linnet object increases dramatically the number of line segments (from 1116 to 10559). This is a consequence of the fact that SpatialLines objects can handle curvilinear segments, made of multiple line segments, as a single line. However, linnet objects follow strictly the definition of linear network provided in the Introduction, which excludes this possibility.

It is required that the network is fully connected in order to allow the computation of a distance between any pair of points. This can be checked with the function connected.

```
> table(connected(chicago.linnet, what = "labels"))

   1    2
9429    2
```

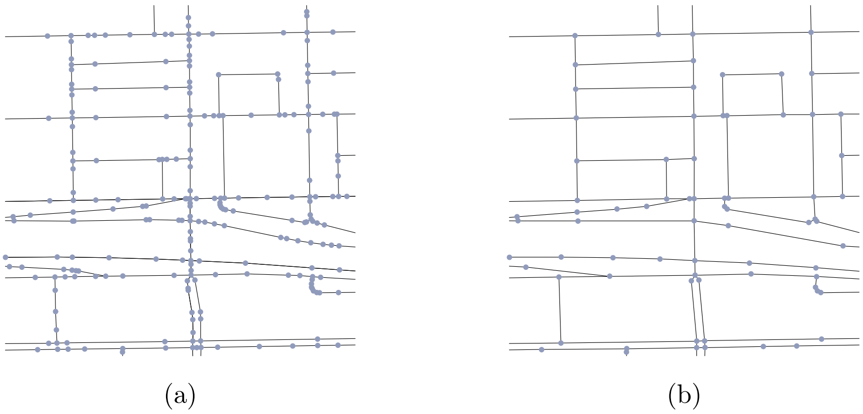
This output means that there is a connected component of 9429 vertices and a separate component of only two vertices. The use of `connected` with the option `what = "components"` enables us to extract the larger connected component for the analysis, discarding the other one.

```
> chicago.linnet.components <- connected(chicago.linnet,
                                          what = "components")
> chicago.linnet.components
[[1]]
Linear network with 9429 vertices and 10558 lines
Enclosing window:
rectangle = [442563.5, 447320] x [4634170, 4637660] units
[[2]]
Linear network with 2 vertices and 1 line
Enclosing window:
rectangle = [442563.5, 447320] x [4634170, 4637660] units
> chicago.linnet <- chicago.linnet.components[[1]]
```

At this point, it is worth considering the possibility of reducing network's complexity. The function `SimplifyLinearNetwork` of SpNetPrep can be used for this purpose. A reasonable choice of the parameters is `Angle = 20` and `Length = 50` (Briz-Redón, 2019). This choice of the parameters means that a pair of segments meeting at a second-degree vertex is merged into one single segment if the angle they form (measured from 0° to 90°) is lower than 20° and if the length of each of them is lower than 50 m. Hence, network's complexity is reduced (in terms of number of segments and lines) while its geometry is preserved. The following lines of code execute `SimplifyLinearNetwork` (computation time: 5.06 seconds) and redefine `chicago.SpLine` according to the structure of the final `linnet` object.

```
> chicago.linnet <- SimplifyLinearNetwork(chicago.linnet,
                                           Angle = 20, Length = 50)
> chicago.linnet
Linear network with 2564 vertices and 3693 lines
Enclosing window:
rectangle = [442563.5, 447320] x [4634170, 4637660] units
> chicago.SpLines <- as.SpatialLines.psp(as.psp(chicago.linnet))
```

Thus, the final road network from Chicago has 3693 lines and 2564 vertices. An example of how `SimplifyLinearNetwork` reduces network's complexity is shown in Figure 8.10, which corresponds to a squared zone of Chicago's network with a diameter of 600 m.



(a) (b)

Figure 8.10: Extracting a part of the road network structure analyzed from Chicago. Original structure extracted (left), made of 273 lines and 260 vertices, and simplified version of it (right) with 114 lines and 101 vertices.

Downloading and preparing crime data

Point-referenced crime datasets corresponding to several cities from the United States of America can be downloaded through the R package crimedata (Ashby, 2019). Concretely, crimedata currently provides (as of September 2019) crime open data recorded in Austin, Boston, Chicago, Detroit, Fort Worth, Kansas City, Los Angeles, Louisville, Mesa, New York, San Francisco, Tucson and Virginia Beach. Therefore, the function `get_crime_data` from this package can be used for downloading a dataset of crime events recorded in Chicago in the period 2007-2018.

```
> chicago.crimes <- get_crime_data(years = 2007:2018, cities = "Chicago")
> dim(chicago.crimes)
[1] 39151    12
```

The year, month and hour of occurrence of each crime can be extracted with the corresponding functions of the package lubridate (Grolemund and Wickham, 2011).

```
> chicago.crimes$year <- year(chicago.crimes$date_single)
> chicago.crimes$month <- month(chicago.crimes$date_single)
> chicago.crimes$hour <- hour(chicago.crimes$date_single)
```

Then, a marked point pattern lying on `chicago.linnet` can be created with function `lpp` to provide the framework required by the DRHotNet package. A `data.frame` is passed to `lpp` including the coordinates of the events (in UTM), the type of event according to the receiver of the offense (`offense_against`), and the year, month and hour of occurrence that have been just computed.

```
> chicago.crimes.coord <- data.frame(x = chicago.crimes$longitude,
                                     y = chicago.crimes$latitude)
> coordinates(chicago.crimes.coord) <-~ x + y
> lonlat_proj <- "+proj=longlat +datum=WGS84 +ellps=WGS84 +towgs84=0,0,0"
> utm_proj <- "+proj=utm +zone=16 ellps=WGS84"
> proj4string(chicago.crimes.coord) <- lonlat_proj
> chicago.crimes.coord <- spTransform(chicago.crimes.coord, utm_proj)
> X.chicago <- lpp(data.frame(x = chicago.crimes.coord@coords[,1],
                   y = chicago.crimes.coord@coords[,2],
                   offense_against = chicago.crimes$offense_against,
                   year = chicago.crimes$year,
                   month = chicago.crimes$month,
                   hour = chicago.crimes$hour),
                   chicago.linnet)
Warning message:
37825 points were rejected as lying outside the specified window
```

A total of 37825 points are rejected because they lie outside the road network. Hence, a marked point pattern of 1326 crimes that lie on `chicago.linnet` remains for the analysis. The four marks are categorical, presenting the following values and absolute frequencies:

```
> table(X.chicago$data$offense_against)

   other  persons property  society
      80      268      831      147
> table(X.chicago$data$year)

2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018
 138  130  133  129  116   97   94   90   85  111  105   98
> table(X.chicago$data$month)

  1   2   3   4   5   6   7   8   9  10  11  12
103  78 120 102 111 114 138 101 115 124 114 106
> table(X.chicago$data$hour)

 0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17
62 33 44 24 23 15 31 29 61 72 53 73 82 63 66 78 69 75
18  19  20  21  22  23
52  76  70  72  52  51
```

Hence, DRHotNet functionalities are now applicable to `X.chicago`. In the main example shown, the relative occurrence of crimes against persons along the road network included in `X.chicago` is analyzed. To this end, the functions of the package are used following the steps that have been established for the differential risk hotspot detection methodology previously described.

Estimating a relative probability surface

The function RelativeProbabilityNetwork of DRHotNet has to be used at first to estimate a relative probability surface over the linear network. As it has been explained, this implies estimating a relative probability in the middle point of each segment of the network.

```
> rel_probs_persons <- RelativeProbabilityNetwork(X = X.chicago,
                                                  lixel_length = 50,
                                                  sigma = 250,
                                                  mark = "offense_against",
                                                  category_mark = "persons")
```

In this example (computation time: 1.53 seconds), an upper bound of 50 m is chosen for lixel's length. This means that segments shorter than 50 m are not split, whereas those longer than 50 m are split into several of no more than 50 m length. This operation is performed internally by RelativeProbabilityNetwork with the function lixellate from spatstat. The bandwidth parameter, σ, is set to 250 m. The mark and category_mark parameters are used to specify the type of event that is under analysis.

The exploration of the object rel_probs_persons with function str allows the user to check that the choice of a 50 m threshold for lixel's length produces 7614 segments along the network. Thus, a relative probability is estimated in the middle point of each of these segments, which can be accessed typing $probs:

```
> str(rel_probs_persons)
List of 5
 $ probs        : num [1:7614] 0.131 0.162 0.137 0.115 0.108 ...
 $ lixel_length : num 50
 $ sigma        : num 250
 $ mark         : chr "offense_against"
 $ category_mark: chr "persons"
```

The function PlotRelativeProbabilities can then be used to obtain a map of the relative probability surface as the one shown in Figure 8.11b. Figure 8.11 also contains the relative probability surfaces corresponding to the choices of $\sigma = 125$ (Figure 8.11a), $\sigma = 500$ (Figure 8.11c) and $\sigma = 1000$ (Figure 8.11d). It can be observed that the choice of a larger value for σ smooths the relative probability surface, which in the case of $\sigma = 500$ or $\sigma = 1000$ leads to the configuration of a small number of clearly distinguishable zones of the network in terms of the relative probability of offenses against persons. Indeed, whereas the use of $\sigma = 125$ allows the user to obtain quite extreme relative probability values at some segments of the network (the values vary from 0 to 0.99), choosing $\sigma = 1000$ causes that all the relative probabilities lie in the interval $[0.07, 0.32]$.

In view of Figure 8.11, $\sigma = 250$ seems a reasonable choice (although some procedures for bandwidth selection could be explored for taking this decision). Therefore, this selection of the bandwidth parameter is maintained to display now how the DRHotspots_k_n function of the package performs.

Detecting differential risk hotspots

The function DRHotspots_k_n needs four parameters to be specified: X (a point pattern on a linear network), rel_probs (an object like the one obtained in the previous step), k and n. Parameters k and n control the differential risk hotspot procedure, as it has been explained before. For example, you can try with k = 1 and n = 30 (computation time: 1.12 minutes):

```
> hotspots_persons <- DRHotspots_k_n(X = X.chicago,
                                    rel_probs = rel_probs_persons,
                                    k = 1, n = 30)
```

The output of the function DRHotspots_k_n presents the following structure:

```
> str(hotspots_persons)
List of 8
 $ DRHotspots   :List of 5
  ..$ : num [1:12] 247 313 314 1150 1151 ...
  ..$ : num [1:12] 551 552 553 554 5080 ...
  ..$ : num [1:8] 1093 1818 2466 4575 6127 ...
  ..$ : num 4271
  ..$ : num [1:2] 5533 5535
 $ k            : num 1
 $ n            : num 30
 $ lixel_length : num 50
 $ sigma        : num 250
 $ mark         : chr "offense_against"
 $ category_mark: chr "persons"
 $ PAI_type     : num 7.82
```

The first component of hotspots_persons contains the differential risk hotspots that have been detected for the values of k and n provided to DRHotspots_k_n. In this case, five differential risk hotspots are found, which are formed by 12, 12, 8, 1 and 2 road segments, respectively. The object hotspots_persons also contains the values of the parameters involved in the computation of the hotspots and a final component that includes the global PAI_{type} for the set, which is 7.82 in this example.

The function SummaryDRHotspots can be used to provide a summary of each of the differential risk hotspots determined by DRHotspots_k_n (computation time: 15.50 seconds):

```
> summary_persons <- SummaryDRHotspots(X = X.chicago,
                                       rel_prob = rel_probs_persons,
                                       hotspots = hotspots_persons)
```

The output of SummaryDRHotspots includes a count of the number of events located within each differential risk hotspot and how many of these correspond to the category ‘‘persons’’:

```
> summary_persons[,c("Events type (ctr)", "All events (ctr)",
                     "Prop. (ctr)")]
  Events type (ctr) All events (ctr) Prop. (ctr)
1                 4               11        0.36
2                 5                8        0.62
3                 0                1        0.00
4                 0                0         NaN
5                 0                0         NaN
```

The summary also contains the length (in meters) of each differential risk hotspot and the PAI_{type} that corresponds to each of them:

```
> summary_persons[,c("Length in m (ctr)", "PAI_type (ctr)")]
  Length in m (ctr) PAI_type (ctr)
1            456.86           9.13
2            366.01          14.24
3            265.71           0.00
4             42.70           0.00
5             68.78           0.00
```

Furthermore, the output of SummaryDRHotspots also provides the same statistics for an extension of each of the hotspots. The reason to include this information is because if there are not many events available in the dataset (as it happens in this example), the method can determine differential risk hotspots where very few events, if any, have taken place. Indeed, in the output of SummaryDRHotspots shown above, there are two hotspots including zero events, one including only one, and two more containing a very reduced number of crimes. The fact of employing kernel density estimation to infer a relative probability surface makes more convenient to think of each differential risk hotspot as the union of a center or core (what DRHotspots_k_n returns, the hotspot itself) and an extension of it. Hence, by considering an extension of the differential risk hotspot one can better appreciate the zone of the network that has been accounted for in the estimation of the relative probability values corresponding to the segments of the hotspot.

By default, the extension computed by SummaryDRHotspots coincides with a neighbourhood of the segments forming each differential risk hotspot of order $o = \frac{\sigma}{\text{Lixel length}}$ (rounded to the nearest integer), although a different order can be specified through the parameter order_extension. In this example, we have $o = \frac{250}{50} = 5$, which is used by SummaryDRHotspots if no other order is indicated:

```
> summary_persons[,c("Events type (ext)", "All events (ext)",
                     "Prop. (ext)")]
  Events type (ext) All events (ext) Prop. (ext)
1                14               42        0.33
2                14               50        0.28
3                11               33        0.33
4                10               30        0.33
5                 5               15        0.33
> summary_persons[,c("Length in m (ext)", "PAI_type (ext)")]
  Length in m (ext) PAI_type (ext)
1           3526.69           4.14
2           3285.57           4.44
3           1921.33           5.97
4           1318.60           7.91
5           1072.17           4.86
```

It can be observed that all extensions of the differential risk hotspots include a reasonable number of events and that the corresponding proportions of offenses against persons are clearly above the global proportion for the dataset, which is $268/1326 \approx 0.20$.

Assessing the statistical significance of the hotspots

Following with the choice of k = 1 and n = 30, it only remains to estimate a p-value for each of the differential risk hotspots detected. This can be done calling the function SummaryDRHotspots again and specifying compute_p_value = T. A total of 200 iterations are selected for performing the Monte Carlo simulation process (computation time: 18.89 minutes):

```
> summary_persons <- SummaryDRHotspots(X = X.chicago,
                                       rel_prob = rel_probs_persons,
                                       hotspots = hotspots_persons,
                                       compute_p_value = T,
                                       n_it = 200)
> summary_persons$‘p-value‘
[1] 0.015 0.010 0.035 0.035 0.000
```

Therefore, the five differential risk hotspots detected with $k = 1$ and $n = 30$ are statistically significant ($p < 0.05$). It is worth noting, however, that the usual significance level of 0.05 should be reduced (corrected) if many differential risk hotspots are detected to avoid the presence of multiple comparison problems.

Choosing *k* and *n*

Remember that a higher value of either `k` or `n` represents using a more stringent criterion regarding hotspot detection. This is illustrated through the four maps available in Figure 8.12, which can be generated with the function `PlotHotspots` of DRHotNet. For instance, as in the following example for the object `hotspots` created previously (which corresponds to Figure 8.12d):

```
> PlotHotspots(X = X.chicago, hotspots = hotspots)
```

Indeed, Figure 8.12 shows the differential risk hotspots that `DRHotspots_k_n` detects for the choices of `k = 0.5` and `n = 20` (Figure 8.12a), `k = 1.5` and `n = 20` (Figure 8.12b), `k = 1` and `n = 10` (Figure 8.12c), and `k = 1` and `n = 30` (Figure 8.12d). Two of these combinations of conditions on k and n are implicitly represented by the other two. Consequently, the differential risk hotspots shown in Figure 8.12b are contained in Figure 8.12a, and those in Figure 8.12d are contained in Figure 8.12c. The choice of `k = 1` and `n = 30` leads to the highest global PAI$_{type}$ among the four combinations of parameters indicated, with the value of 15.6 mentioned before. In this regard, it is recommended to perform a sensitivity analysis on the values of k and n to decide which combination is more convenient. The sensitivity analysis carried out in Chapter 3 yielded that a choice around $k = 1.5$ and $n = 45$ was optimal in terms of the PAI$_{type}$ for the traffic accident dataset that was investigated. However, each dataset should require a specific analysis.

Thus, a sensitivity analysis on k and n can be carried out with the function `Sensitivity_k_n`. The user has to provide a point pattern (`X`), an object containing the relative probabilities of a type of event along the network (`rel_probs`) and a set of values for k and n (`ks` and `ns`, respectively):

```
> sensitivity_analysis <- Sensitivity_k_n(X = X.chicago,
                                          rel_prob = rel_probs_persons,
                                          ks = seq(0,3,0.5),
                                          ns = seq(10,30,5))
> sensitivity_analysis
        n = 10 n = 15 n = 20 n = 25 n = 30
k = 0     3.05   3.60   4.36   4.72   5.58
k = 0.5   3.93   4.90   5.96   6.23   6.52
k = 1     4.62   5.82   6.11   6.66   7.82
k = 1.5   5.44   7.63   4.52   0.00   0.00
k = 2     5.66  18.14   0.00   0.00     NA
k = 2.5   7.06     NA     NA     NA     NA
k = 3     0.00     NA     NA     NA     NA
```

The output from Sensitivity_k_n is a matrix that contains the PAI$_{type}$ values that correspond to each combination of k and n indicated by ks and ns. A NA value represents that no differential risk hotspots can be found for such combination of parameters. According to this matrix, the highest PAI$_{type}$ value is achieved for $k = 2$ and $n = 15$.

Therefore, one can choose the values of k and n that maximize the PAI$_{type}$ (considering the parameters provided in ks and ns) to determine the final set of differential risk hotspots. However, this criteria may lead sometimes to detect a very low number of differential risks hotspots and hence to miss zones of the network that may also deserve some attention. Hence, a more conservative approach could be considering several combinations of k and n that yield some of the highest values of PAI$_{type}$ and explore each set of differential risk hotspots associated. Then, one could investigate the output of SummaryDRHotspots for each combination of parameters (including the computation of p-values) to better decide which zones of the network are relevant for the type of event of interest.

Other applications of the methodology

This final section shows the results that are obtained for other type of events for comparative purposes. First, the marks X.chicago are recoded into binary outcomes as follows:

```
> year_after_2012 <- ifelse(X.chicago$data$year>2012, "Yes", "No")
> month_winter <- ifelse(X.chicago$data$month%in%c(12,1,2), "Yes", "No")
> hour_21_3 <- ifelse(X.chicago$data$hour%in%c(21:23,0:3), "Yes", "No")
> marks(X.chicago) <- data.frame(as.data.frame(marks(X.chicago)),
                               year_after_2012 = year_after_2012,
                               month_winter = month_winter,
                               hour_21_3 = hour_21_3)
```

The relative probability surfaces have to be computed. The values for lixel.length and sigma used in the previous examples are chosen again.

```
> rel_probs_after_2012 <- RelativeProbabilityNetwork(X = X.chicago,
                                                   lixel_length = 50,
                                                   sigma = 250,
                                                   mark = "year_after_2012",
                                                   category_mark = "Yes")
> rel_probs_winter <- RelativeProbabilityNetwork(X = X.chicago,
                                               lixel_length = 50,
                                               sigma = 250,
                                               mark = "month_winter",
                                               category_mark = "Yes")
> rel_probs_21_3 <- RelativeProbabilityNetwork(X = X.chicago,
                                                lixel_length = 50,
                                                sigma = 250,
                                                mark = "hour_21_3",
                                                category_mark = "Yes")
```

The corresponding sensitivity analyses are carried out:

```
> sensitivity_after_2012 <- Sensitivity_k_n(X = X.chicago,
                                             rel_prob = rel_probs_after_2012,
                                             ks = seq(0,3,0.5),
                                             ns = seq(10,30,5))
> sensitivity_after_2012
        n = 10 n = 15 n = 20 n = 25 n = 30
k = 0     2.09   2.67   3.45   4.98   8.71
k = 0.5   2.58   2.99   4.28   7.16  16.81
k = 1     5.43   7.61   9.71  17.64  28.91
k = 1.5   3.40     NA     NA     NA     NA
k = 2       NA     NA     NA     NA     NA
k = 2.5     NA     NA     NA     NA     NA
k = 3       NA     NA     NA     NA     NA
> sensitivity_winter <- Sensitivity_k_n(X = X.chicago,
                                         rel_prob = rel_probs_winter,
                                         ks = seq(0,3,0.5),
                                         ns = seq(10,30,5))
> sensitivity_winter
        n = 10 n = 15 n = 20 n = 25 n = 30
k = 0     2.88   4.02   5.06   7.36   9.89
k = 0.5   3.14   4.60   4.30   6.97   6.03
k = 1     5.18   7.35   1.32   0.00     NA
k = 1.5   0.83   0.00     NA     NA     NA
k = 2     0.00     NA     NA     NA     NA
k = 2.5     NA     NA     NA     NA     NA
k = 3       NA     NA     NA     NA     NA
> sensitivity_21_3 <- Sensitivity_k_n(X = X.chicago,
                                       rel_prob = rel_probs_21_3,
                                       ks = seq(0,3,0.5),
                                       ns = seq(10,30,5))
> sensitivity_21_3
        n = 10 n = 15 n = 20 n = 25 n = 30
k = 0     2.42   3.20   4.49   4.69   4.81
k = 0.5   3.00   4.08   5.30   5.55   6.03
k = 1     3.04   3.94   4.74   4.11   0.00
k = 1.5   3.67   4.58   6.37   5.94     NA
k = 2     5.46   5.79   7.30   5.58     NA
k = 2.5  12.03  11.36  14.41   5.67     NA
k = 3     0.00   0.00   0.00     NA     NA
```

The highest PAI$_{type}$ values for `year_after_2012`, `month_winter` and `hour_21_3` are 28.91, 9.89 and 14.41, respectively. The differential risk hotspots that are obtained for the combination of k and n that lead to these PAI$_{type}$ values can be visualized with `PlotHotspots` (Figure 8.13):

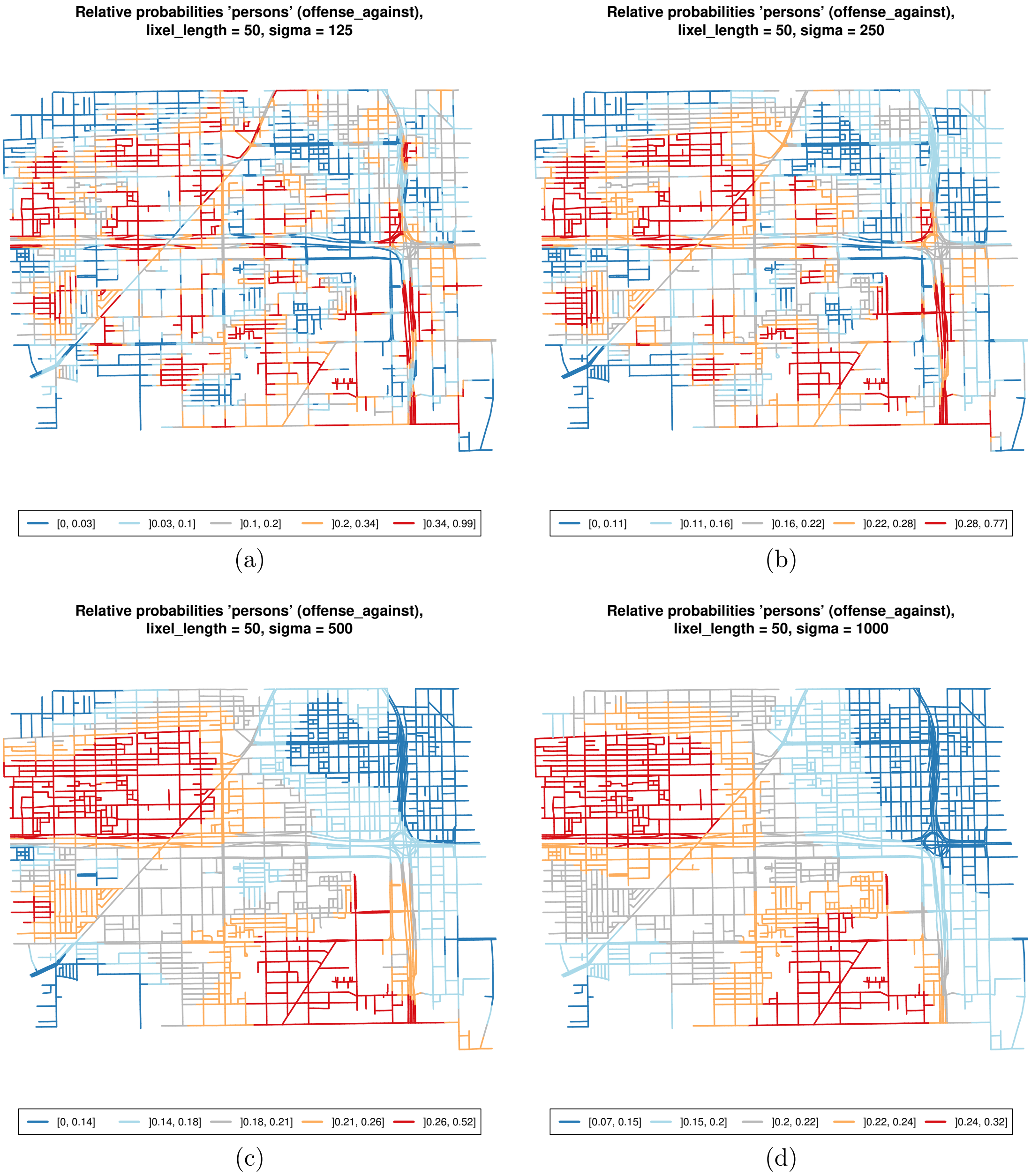
```
> hotspots_after_2012 <- DRHotspots_k_n(X = X.chicago,
                                        rel_prob = rel_probs_after_2012,
                                        k = 1,
                                        n = 30)
> PlotHotspots(X = X.chicago, hotspots_after_2012)

> hotspots_winter <- DRHotspots_k_n(X = X.chicago,
                                    rel_prob = rel_probs_winter,
                                    k = 0,
                                    n = 30)
> PlotHotspots(X = X.chicago, hotspots_winter)

> hotspots_21_3 <- DRHotspots_k_n(X = X.chicago,
                                  rel_prob = rel_probs_21_3,
                                  k = 2.5,
                                  n = 20)
> PlotHotspots(X = X.chicago, hotspots_21_3)
```

8.2.2 Summary

The R package DRHotNet for detecting differential risk hotspots on linear networks has been described. The use of linear networks in the context of hotspot detection is becoming more important over the years, particularly in the fields of criminology and traffic safety. In addition, it is also of great interest sometimes to detect zones of a linear network where certain type of event is especially overrepresented. Hence, DRHotNet consists of an easy-to-use tool implemented in R to accurately locate the microzones of a linear network where the incidence of a type of event is considerably higher than in the rest of it.


Relative probabilities 'persons' (offense_against), lixel_length = 50, sigma = 125
[0, 0.03]
]0.03, 0.1]
]0.1, 0.2]
]0.2, 0.34]
]0.34, 0.99]
Relative probabilities 'persons' (offense_against), lixel_length = 50, sigma = 250
[0, 0.11]
]0.11, 0.16]
]0.16, 0.22]
]0.22, 0.28]
]0.28, 0.77]
Relative probabilities 'persons' (offense_against), lixel_length = 50, sigma = 500
[0, 0.14]
]0.14, 0.18]
]0.18, 0.21]
]0.21, 0.26]
]0.26, 0.52]
Relative probabilities 'persons' (offense_against), lixel_length = 50, sigma = 1000
[0.07, 0.15]
]0.15, 0.2]
]0.2, 0.22]
]0.22, 0.24]
]0.24, 0.32]

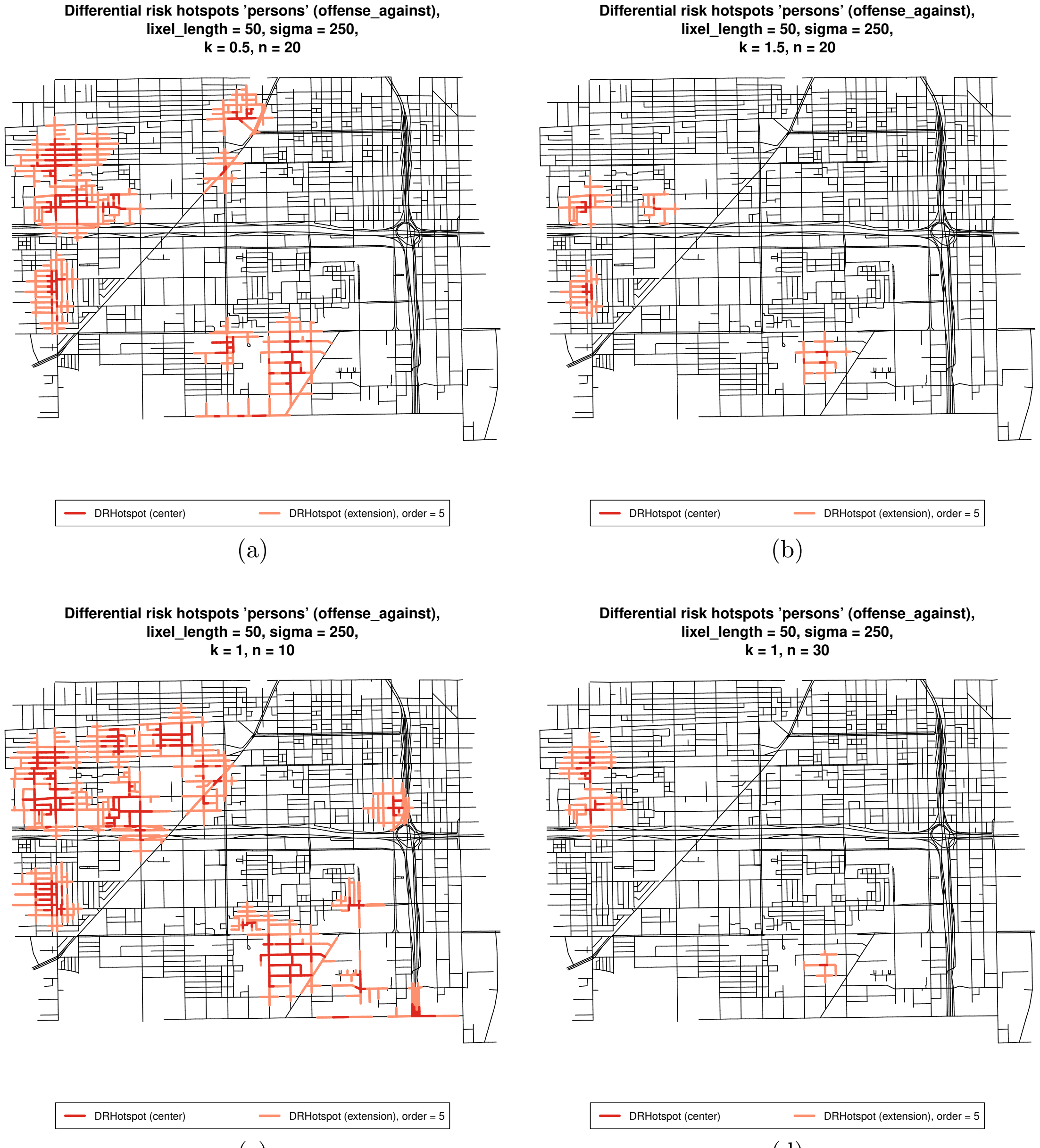
(a)

(b)

(c)

(d)

Figure 8.11: Outputs from the function `PlotRelativeProbabilities` for the following choices of σ: 125 (a), 250 (b), 500 (c) and 1000 (d).



Differential risk hotspots 'persons' (offense_against), lixel_length = 50, sigma = 250, k = 0.5, n = 20

DRHotspot (center)
DRHotspot (extension), order = 5

(a)

Differential risk hotspots 'persons' (offense_against), lixel_length = 50, sigma = 250, k = 1.5, n = 20

DRHotspot (center)
DRHotspot (extension), order = 5

(b)

Differential risk hotspots 'persons' (offense_against), lixel_length = 50, sigma = 250, k = 1, n = 10

DRHotspot (center)
DRHotspot (extension), order = 5

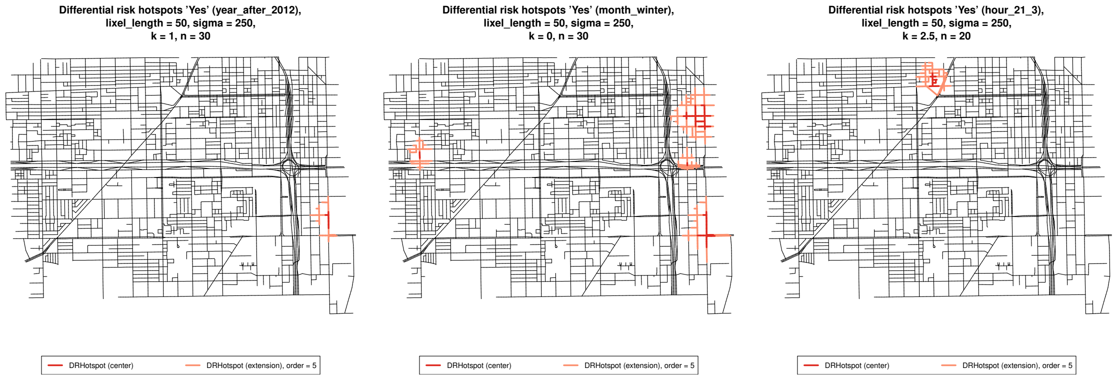
(c)

Differential risk hotspots 'persons' (offense_against), lixel_length = 50, sigma = 250, k = 1, n = 30

DRHotspot (center)
DRHotspot (extension), order = 5

(d)

Figure 8.12: Outputs from the function `PlotHotspots` for the following choices of k and n: $k = 0.5$ and $n = 20$ (a), $k = 1.5$ and $n = 20$ (b), $k = 1$ and $n = 10$ (c) and $k = 1$ and $n = 30$ (d).


Differential risk hotspots 'Yes' (year_after_2012),
lixel_length = 50, sigma = 250,
k = 1, n = 30
DRHotspot (center)
DRHotspot (extension), order = 5
Differential risk hotspots 'Yes' (month_winter),
lixel_length = 50, sigma = 250,
k = 0, n = 30
DRHotspot (center)
DRHotspot (extension), order = 5
Differential risk hotspots 'Yes' (hour_21_3),
lixel_length = 50, sigma = 250,
k = 2.5, n = 20
DRHotspot (center)
DRHotspot (extension), order = 5

(a) (b) (c)

Figure 8.13: Outputs from the function `PlotHotspots` for the marks `year_after_2012`, `month_winter` and `hour_21_3` and the categorical value `Yes` for the three, considering the combinations of k and n that maximize the PAI_{type} (for the values of k and n tested).

Chapter 9

Conclusions and future work

This final chapter summarizes the main outcomes that have been obtained in the context of developing and applying spatio-temporal techniques to analyze crime and traffic safety data, while also indicating some lines of research that could be explored in the future.

A capital goal from the start was to analyze spatial data at the road segment level. Using linear networks as a spatial window is becoming more relevant in several fields, especially in those two in which the focus has been put on: criminology and traffic safety. The technical complexity of managing linear networks led to the creation of the R package SpNetPrep (Chapter 8). With the aid of this package, several studies were carried out, including multivariate analyses (Chapter 2) and the development of a differential risk hotspot detection procedure for linear networks (Chapter 3), which also gave raise to another R package, DRHotNet (Chapter 8), which allows the user to carry out the complete procedure, providing several capabilities such as the visualization of the hotspots or the performance of sensitivity analyses on the two parameters involved. This hotspot detection methodology could be further investigated by integrating statistical models to fit event counts at the road segment level, allowing for the inclusion of covariate information. In this regard, specifying a suitable modelling approach would be essential, as the analysis of event counts at the road segment level usually leads to the presence of overdispersion in the data and to an excessive number of zeros.

Another analysis that made use of the road network structure of Valencia is presented in Chapter 4. In this analysis, the classical version of the Knox test for assessing the presence of space-time interaction is modified by accounting for spatio-temporal risk heterogeneity for the investigation of a dataset of burglaries. Two main mechanisms that can be used to adjust the Knox test are described: the one proposed by Schmertmann (2015), which considers fixed covariate effects that vary in space and time as the covariates, and an alternative approach that contemplates the existence of space-time varying covariate effects (with the covariate values remaining constant). The latter was chosen due to data unavailability for a long enough period, but both approaches are reasonable to achieve the goal of adjusting the usual version of the Knox test in order to get more accurate results. In this regard, carrying out more specific investigations on the choice of both mechanisms, including the consideration of different spatial and temporal domains, could be of interest.

From the analysis of spatial data at the road segment level, the research moved towards area-level analyses. Despite the necessity of considering road networks for gaining accuracy, there are still enough reasons to analyze data at the level of areal spatial units. First, it is hard many times to obtain valuable covariates at the road segment, which can detract from the usefulness of the road-level analysis. Furthermore, for practical reasons such as the implementation of preventive measures, it can be more convenient to analyze event counts per census tract or police district.

Chapter 5 is devoted to one fundamental issue with regard to area-level spatial analyses: the modifiable areal unit problem (MAUP). This chapter includes a detailed case study that focuses on examining the possible effects of the MAUP in macroscopic safety analysis, although the findings could be extended to other fields of research. In particular, the investigation was carried out by accounting for the two main spatial characteristics whose variation is usually implicated in the MAUP: scale and zoning. The results suggested that the variations in scale were less influential in model estimations (for both CAR and GWR models), although there were remarkable differences among the set of covariates considered. In this regard, the analysis also showed that a count-based covariate that is defined on the basis of a highly-clustered point pattern may be more sensitive to the MAUP than a covariate that is defined from a more regular pattern. In addition, one should not overlook how the basic characteristics of the response variable involved in model fitting are affected by any change in scale and/or zoning. Hence, the choice of a certain modelling approach should also account for the possible variations in the statistical properties of the response variable. In the future, further research could be conducted by performing simulation studies accounting explicitly for the main factors that have been already outlined: scale, zoning and characteristics of both response and covariates. These simulation studies could be helpful to determine with greater accuracy the extent of the consequences of the MAUP and, more importantly, the steps that can be carried out in order to attempt minimize to its effects.

Chapter 6 includes a comparative analysis of different methodologies that can be considered for analyzing how certain fixed landscape features (as schools) increase (or decrease) the presence of an event of interest (as traffic accidents). The statistical techniques that were used for the comparison were: observed vs expected ratios (defined for several spatial radii), conditional autoregressive models, multiple source regression models and logistic regression models under the context of a case-control design. Although the different approaches yielded quite coherent results, some differences emerged, especially in view of the results provided by the logistic regression models based on case and control schools. Thus, from a methodological perspective, the main conclusion obtained from this research was the necessity of testing more than one modelling approach to achieve more robust conclusions and the potential of case-control designs to gain more singular insights from the data. All these approaches could be further investigated in the future and compared with other popular modelling techniques in the context of analyzing the effect of fixed locations on crime outcomes such as risk terrain modelling (RTM).

The final study that was conducted during the development of this thesis, which is provided in Chapter 7, was dedicated to investigate how the presence of non-geocoded (missing) data can affect and even reduce the reliability of a spatial analysis. The main goal of the study was to reestimate the first minimum acceptable hit rate (minimum percentage of events that need to be geocoded to rely on the subsequent spatial analysis of the data) that was provided fifteen years ago in the literature, with no more research being done ever since. The new reestimation accounted for some characteristics of both the spatial window (aggregation level) and the point pattern under study (intensity and clustering levels) that were not considered for the first estimation. The analysis revealed that the intensity of the pattern strongly affects the minimum acceptable hit rate. Furthermore, the likely presence of non-uniform mechanisms of non-geocoded data and the possibility that different statistical techniques are more sensitive to the existence of non-geocoded data are also explored. This study could be relevant for spatial statisticians that may have missing data after carrying out a geocoding procedure. In particular, the new insights provided into the minimum acceptable hit rate may be useful for the analysis of both crime and traffic safety data, which are particularly prone to have non-geocoded events for multiple reasons.

To conclude, it is important to remark that, from a practical perspective, the results derived from the investigations carried out during the development of this thesis have unfortunately not yet been applied. The future implementation of preventive measures following some of the statistical models and techniques that are described in this thesis would surely lead to new methodological challenges and hence would provide the right context to make possible the design of more accurate methods for crime and traffic accident prevention. In this regard, it is worth noting that putting these methodologies into practice is important to start to reduce the incidence of crimes and traffic accidents and to gain more insights into the methodologies established.

Resumen

En 1855, un médico inglés llamado John Snow presentó un mapa que mostraba la ubicación de los casos de cólera detectados durante la epidemia londinense de 1854 (Snow, 1855). Snow investigó el patrón formado por estas ubicaciones y sugirió que una fuente de agua pública situada en Broad Street había sido la causante del origen del brote epidémico. La creación de este mapa y el posterior análisis de la epidemia son ampliamente considerados como el primer ejemplo de análisis espacial de datos y uno de los orígenes de la epidemiología (Susser and Bresnahan, 2001).

Desde 1855, han surgido muchos tipos de datos espaciales, especialmente durante las últimas décadas. De hecho, la estadística espacial es el campo de conocimiento dedicado específicamente al análisis de todo tipo de datos espaciales. Existen tres tipos principales de datos espaciales, a saber, los *patrones puntuales*, los *datos sobre agregados espaciales* y los *datos geoestadísticos* (Cressie, 1993), que han dado lugar al desarrollo de tres líneas principales de investigación entre los expertos en estadística espacial. En los párrafos siguientes se describen estos tres tipos de datos y se hace un breve resumen de las técnicas que tradicionalmente se han empleado para el análisis de cada uno de ellos. Antes de eso, cabe destacar que los datos espaciales suelen permitir incorporar una componente temporal, lo que conduce a la definición de datos espacio-temporales y a la existencia de multitud de extensiones de las herramientas fundamentales de la estadística espacial que pueden ser utilizadas en un contexto de datos espacio-temporales.

Un patrón puntual es un conjunto de puntos que se encuentran en un determinado espacio o región. Cada punto de un patrón puntual suele denominarse *evento*. El mecanismo estocástico que genera un patrón puntual se llama proceso puntual (Diggle, 2013). Dado un patrón puntual, el objetivo más común es investigar su estructura. En particular, suele ser de interés determinar si los puntos del patrón presentan alguna dinámica de atracción o inhibición o, por el contrario, si satisfacen la conocida como hipótesis de aleatoriedad espacial completa (Diggle, 2013). El uso combinado de técnicas no paramétricas como la K-función (Ripley, 1977) con múltiples enfoques basados en modelos (Diggle, 2013) facilita la investigación de los patrones puntuales.

Los datos sobre agregados espaciales son observaciones de un proceso aleatorio sobre un número contable de regiones o unidades espaciales. El número de veces que un tipo de evento ha ocurrido en cada unidad administrativa de una ciudad durante un período de tiempo constituye un ejemplo muy típico de datos agregados. El empleo

de modelos con retardo espacial (Anselin, 1988) o de modelos lineales generalizados que tienen en cuenta la autocorrelación espacial entre las regiones (Banerjee et al., 2004) es una práctica común para explicar la variable respuesta en base a un conjunto de covariables. Además, la detección de regiones en las que una variable de red presenta un valor particularmente alto (hotspots) es otra aplicación muy importante (Anselin, 1995; Kulldorff, 1997).

Los datos geoestadísticos corresponden a una colección de medidas disponibles en una región continua. La geoestadística se concibió inicialmente para la geología minera, pero las aplicaciones posibles de esta rama de la estadística espacial son innumerables. El *variograma* y el *kriging* son posiblemente los dos conceptos capitales en el campo de la geoestadística (Cressie, 1993; Krige, 1960; Matheron, 1963).

Pese a la persistencia de esta clasificación de la estadística espacial en tres subáreas, se ha argumentado que la distinción más importante debe hacerse entre los procesos estocásticos espacialmente continuos y los espacialmente discretos (Diggle et al., 2013). Además, la tendencia actual es la de tratar de unificar todo tipo de datos y técnicas espaciales y, en particular, la de utilizar modelos geoestadísticos para el tratamiento de datos agregados y patrones puntuales (Diggle and Giorgi, 2019). Para trabajar en esta dirección, los procesos log-Gaussianos de tipo Cox son una herramienta fundamental (Møller et al., 1998).

La estadística espacial y espacio-temporal puede aplicarse a muchos campos de investigación. Además de la epidemiología y la geología minera, que ya se han mencionado, los datos espaciales suelen surgir en el contexto de otras disciplinas como la agricultura, la astronomía, la biología, la criminología, la hidrología, la meteorología, la seguridad vial y la teledetección, entre otras. De hecho, la colaboración tanto de la Policía Local de València como de la Policía Nacional con el Departamento de Estadística e Investigación Operativa de la Universidad de València ha sido un factor decisivo para orientar esta tesis a la investigación de datos espaciales y espacio-temporales de especial interés en los campos de la criminología y la seguridad vial, y al uso y desarrollo de técnicas estadísticas específicas que faciliten la consecución de este objetivo capital.

De forma más concreta, el objetivo principal de esta tesis era el de identificar "research gaps" entre la amplia literatura existente sobre el uso de métodos estadísticos espacio-temporales para el análisis de datos sobre seguridad vial o criminología. Como parte del desarrollo de esta tesis se ha investigado sobre diferentes estructuras de soporte para el análisis de los datos, sobre modelización estadística de datos de tipo espacial, sobre el estudio espacio-temporal de eventos de tipo contagioso, sobre la calidad de los datos espaciales, o sobre el desarrollo de software específico, entre otros. En todos los casos, la pretensión era mejorar la metodología existente, tanto para abordar algunas cuestiones de tipo más bien teórico, como para realizar estudios prácticos de manera más precisa.

En los siguientes párrafos se describen varios temas relevantes que suelen estar implicados en el análisis espacial y espacio-temporal de los datos procedentes de estos dos campos. En primer lugar, se introduce una estructura espacial singular llamada red lineal. Aunque varios de los análisis que se presentan en esta tesis se realizan sobre unidades con área, el empleo de redes lineales se extiende a través de una parte sustancial de la tesis. En segundo lugar, se describen y contextualizan los diferentes estudios que se han realizado para el desarrollo de esta tesis, perfilando toda la estructura del documento.

Las estructuras de tipo red lineal se introdujeron en el contexto de la estadística espacial hace algunos años, proporcionando la base para el análisis de eventos que normalmente se sitúan sobre estructuras de tipo grafo como puede ser una red de calles o carreteras. En efecto, una red lineal plana, L, se define como una colección finita de segmentos lineales, $L = \cup_{i=1}^{n} l_i$, en la que cada segmento contiene los puntos $l_i = [u_i, v_i] = \{tu_i + (1-t)v_i : t \in [0,1]\}$ (Ang et al., 2012; Baddeley et al., 2015, 2017). Siguiendo la nomenclatura de la teoría de grafos, estos segmentos son a veces denominados *aristas* de la red lineal, mientras que los puntos que determinan los extremos de tales segmentos se conocen como los *vértices* de la red.

Por lo tanto, un proceso puntual X sobre L es un proceso puntual finito en el plano, de tal manera que todos los puntos de X se encuentran en la red L (Ang et al., 2012; Baddeley et al., 2015, 2017). De manera similar, una colección de eventos que se observa en L se conoce como un patrón puntual, x, en L.

La investigación de los patrones espaciales que yacen en redes lineales está ganando atención en los últimos años. El diseño de nuevos y más precisos/eficientes estimadores de densidad de tipo *kernel* (McSwiggan et al., 2017; Moradi et al., 2018, 2019; Rakshit et al., 2019b), la introducción de medidas de intensidad basadas en la estructura de tipo grafo (Eckardt and Mateu, 2018), la construcción de indicadores locales de asociación espacial (Eckardt and Mateu, 2017), o la estimación de riesgos relativos (McSwiggan et al., 2019) son algunos de los temas que se están empezando a desarrollar para las redes lineales.

Las redes lineales son un soporte adecuado para afrontar una gran variedad de análisis espaciales o espacio-temporales sobre datos criminológicos o de seguridad vial. Además, las técnicas espacio-temporales se aplican masivamente tanto en el contexto de la criminología como en el de la seguridad del tráfico, con independencia del tipo de soporte de los datos. En los siguientes párrafos se destacan varios tipos de análisis que son de particular interés en estos dos campos, tanto desde una perspectiva metodológica como práctica.

El análisis estadístico de la seguridad vial se remonta, al menos, a los años 40 del siglo pasado. Gordon (1949) sugirió la conveniencia de tratar las lesiones y muertes que se originan en los accidentes de tráfico como un tipo más de datos epidemiológicos. Así pues, en décadas posteriores los dos objetivos principales para los analistas cuantitativos de la seguridad vial han sido la determinación de los factores que se asocian a un mayor número de accidentes de tráfico (o a una determinada tipología de accidente) y la localización exacta de las zonas ("hotspots") en las que los accidentes de tráfico son especialmente probables.

Ambos objetivos suelen llevar a los investigadores a adoptar un enfoque espacial. La elección de un modelo espacial para medir la relación entre una variable respuesta que representa los conteos o tasas de accidentes a un cierto nivel de agregación espacial y un conjunto de covariables definidas sobre las mismas unidades espaciales es vital para evitar pasar por alto la probable presencia de autocorrelación en los resultados y, por lo tanto, posiblemente sesgar y distorsionar las estimaciones de los parámetros del modelo. Por otra parte, la detección de "hotspots" es una práctica común en múltiples campos que producen datos espaciales, incluyendo el de la seguridad vial. Existen numerosos métodos para afrontar este propósito, que también requieren considerar las características espaciales de los datos.

Ya en los años 70, el uso de técnicas espaciales para analizar datos de accidentes de tráfico se convirtió en una práctica habitual. Moellering (1976) propuso el uso de películas animadas por ordenador para observar la distribución espacio-temporal de los patrones de accidentes de tráfico, lo que puede considerarse como una contribución fundamental que fue seguida por muchos otros estudios publicados en los años y décadas siguientes. Hace unos años, Lord and Mannering (2010) resumió la mayoría de aspectos que merecen ser considerados antes de realizar el análisis estadístico de un conjunto de datos de accidentes de tráfico. Además de ciertas cuestiones como el control de la sobredispersión, la presencia de datos faltantes o la escasez de muestra, también se destacó la necesidad de considerar la presencia de autocorrelación espacial y temporal en los datos. De hecho, más recientemente, Mannering et al. (2016) revisó los métodos y enfoques de modelización que pueden elegirse para incorporar los efectos espaciales en el análisis de la seguridad vial. En el contexto típico de la modelización de los recuentos de accidentes observados en un conjunto de regiones, se recomienda el uso de modelos binomiales negativos e inflados en cero para tener en cuenta la sobredispersión y la posible presencia excesiva de ceros en la respuesta (donde los ceros se refieren a las unidades espaciales donde ningún accidente ha sido registrado).

Las redes lineales son también un soporte adecuado para afrontar los dos objetivos principales que se han descrito en los párrafos anteriores, aunque conllevan ciertas dificultades técnicas. En efecto, la modelización, por ejemplo, de los conteos de accidentes registrados a nivel de tramo de calle o carretera está más sujeta al error que el análisis de conteos sobre la base de una subdivisión de la región de estudio dotada de unidades espaciales con área. En efecto, la representación de una estructura vial a través de una red lineal implica la simplificación de ciertas características urbanas, lo que puede complicar la localización de algún accidente en el segmento de la red que corresponde. Dada esta dificultad, que surgió en la primera etapa del desarrollo de esta tesis, se desarrolló el paquete R SpNetPrep (Spatial Network Preprocessing). Este paquete permite a los usuarios realizar la depuración manual de una red lineal que representa una estructura de calles con la ayuda de una aplicación interactiva basada en R Shiny. Esta aplicación ofrece a los usuarios la posibilidad de agregar y eliminar vértices y aristas de la red, la capacidad de agregar direccionalidad a la red de acuerdo con el flujo de tráfico, y la oportunidad de inspeccionar y editar un patrón puntual que ya se ha situado sobre la red. Además, el paquete también contiene una función que simplifica automáticamente la estructura lineal de la red fusionando, bajo ciertas condiciones de longitud y ángulo, las aristas de la red que están unidas por un vértice de segundo grado. El Capítulo 8 de esta tesis contiene una descripción más amplia de este paquete.

Con la ayuda de SpNetPrep, se han realizado varios análisis espacio-temporales de datos de accidentes de tráfico en la red de carreteras de València, constituyendo una parte sustancial de esta tesis. En los siguientes párrafos se ofrece un resumen de estos estudios. En primer lugar, se ha seleccionado un distrito de València para realizar una modelización espacial de los recuentos de accidentes de tráfico a nivel de calle. La necesidad de depurar tanto la estructura de la red como el conjunto de datos sobre accidentes proporcionado por la Policía Local de València llevó a elegir sólo un distrito de València para el análisis. El objetivo principal de este estudio fue encontrar covariables o factores que incrementen la incidencia de los accidentes de tráfico a nivel de calle. En particular, la modelización de los datos tuvo en cuenta la localización de las zonas de intersección entre calles. Se suele observar que muchos de los accidentes que ocurren en una zona urbana o rural se localizan en el entorno de una intersección, por lo que considerar la proximidad a una intersección en la definición de los modelos estadísticos parece más que conveniente. Más específicamente, la red vial original se dividió para distinguir entre los segmentos de carretera de tipo intersección y los de tipo no intersección (según su situación frente al conjunto de intersecciones detectadas). Además, con el fin de confirmar los factores posiblemente implicados en los accidentes de tráfico, la modelización de los datos se combinó con la detección de “coldspots” y “hotspots”, la cual se basó principalmente en la estimación de la densidad de accidente a lo largo de la red (en su versión “network-contstrained”) y en los indicadores locales de asociación espacial (Anselin, 1995). El Capítulo 2 está totalmente dedicado a este análisis.

El segundo estudio realizado se centró en la detección de “hotspots” sobre la red de carreteras de València donde un tipo de accidente de tráfico está sobrerrepresentado. La metodología propuesta por Kelsall et al. (1995) para producir superficies de riesgo relativo se adaptó para permitir la estimación de la probabilidad relativa de ocurrencia que corresponde a un tipo de evento que tiene lugar a lo largo de una red lineal. Luego se propone una estrategia para detectar “hotspots” de riesgo diferencial en la red, de acuerdo a los valores de probabilidad relativa previamente inferidos. Tanto la magnitud de la estimación de la probabilidad como el tamaño de la muestra involucrada en su cálculo son considerados para la determinación de los “hotspots”. El procedimiento completo se describe en el Capítulo 3, el cual incluye una aplicación de la metodología a través de un conjunto de datos de accidentes de tráfico que proporcionaba información sobre el tipo de colisión y sobre los tipos de vehículos implicados en cada accidente. El paquete de R DRHotNet (Differential Risk Hotspots in a Linear Network), disponible en CRAN, implementa el método al completo. Este paquete de R se describe en detalle en el Capítulo 8.

Aunque el análisis de los accidentes de tráfico a nivel de calle proporciona el mayor nivel de precisión tanto para los investigadores como para los profesionales de la seguridad vial, el estudio de los accidentes a nivel de área sigue siendo de interés. Concretamente, el análisis de los conteos de accidentes de tráfico que se registran en un conjunto de unidades administrativas o áreas policiales durante un período de tiempo facilita la consideración de determinadas variables ambientales, estructurales o sociodemográficas para evaluar su relación con los accidentes. En este mismo sentido, la elección de una determinada unidad de análisis territorial puede afectar a los resultados estadísticos que se realicen. Esta cuestión se conoce como el problema de la unidad de área modificable (MAUP) en el ámbito de la estadística espacial (Openshaw, 1984). Esta tesis incluye un estudio de caso sobre las consecuencias de variar la escala o la zonificación de un conjunto de regiones en el contexto del análisis de la seguridad vial (Capítulo 5). Se investigan en detalle los efectos de la modificación de cualquiera de estos dos factores tanto en la respuesta como en las covariables, y el impacto en la estimación y el rendimiento del modelo.

El estudio de la delincuencia desde una perspectiva espacial comenzó a cobrar cierta importancia en los años 50 del siglo pasado. Por ejemplo, Shannon (1954) analizó la distribución espacial de los índices de criminalidad de todos los estados americanos. En la década de los 70 la disciplina comenzó a consolidarse a través de múltiples estudios enfocados a explicar la incidencia espacialmente variable de los tipos de crímenes más relevantes (Block, 1979; Brantingham and Brantingham, 1975; Georges, 1978; Harries, 1973; Stephenson, 1974). A principios de los años 80, los fundamentos de la criminología espacial fueron establecidos definitivamente por Patricia y Paul Brantingham (Brantingham and Brantingham, 1981, 1984).

El uso de redes lineales para realizar un análisis espacial se ha convertido en una práctica esencial en criminología para capturar adecuadamente la distribución de los eventos criminales en el espacio. De hecho, los últimos años están trayendo al campo muchos estudios enfocados en la evaluación de la *ley de concentración del crimen* en una ciudad o región. La ley de concentración del crimen establece que un cierto nivel de concentración del crimen tiende a suceder en una parte de la región de estudio proporcionalmente menor (Weisburd, 2015). Por lo tanto, es necesario analizar los datos espaciales del crimen a una escala apropiada para apreciar verdaderamente el nivel de concentración del mismo. El empleo de redes lineales y por lo tanto el hecho de ubicar los delitos a nivel de calle usualmente produce niveles más realistas de concentración del evento de interés que los proporcionados por las unidades espaciales con área.

Existen varios mecanismos que explican la incidencia espacialmente variable del crimen en el corto, mediano y largo plazo. La presencia de características criminógenas en una ciudad puede estimular, facilitar o complicar las actividades de los delincuentes y, como consecuencia de ello, explicar las tasas de criminalidad a medio y largo plazo. En la literatura se distinguen tres tipos principales de lugares: los que atraen el crimen, los que favorecen su generación y los que lo perjudican (Brantingham and Brantingham, 1995; Kinney et al., 2008). Los atractores son lugares que constituyen un contexto singular que naturalmente ofrece más oportunidades con respecto a la comisión de un delito. Básicamente, es probable que los lugares de una ciudad en los que cierto tipo de delito es común (por ejemplo, algún lugar en el que los traficantes y consumidores de drogas suelen reunirse) también atraigan otras acciones delictivas. Por otra parte, los generadores de delitos son lugares que atraen la presencia de muchas personas y, por lo tanto, proporcionan más objetivos a los delincuentes. Un centro comercial o un estadio de fútbol son dos ejemplos de generadores de delito. Por último, un detractor de la delincuencia es un lugar que complica el desarrollo de las actividades delictivas. Una comisaría de policía es un ejemplo de detractor de la delincuencia porque implica una mayor presencia de "guardianes" (agentes de policía) en un entorno de la comisaría. Existen múltiples alternativas de modelización estadística para evaluar si cierto lugar (o un conjunto de lugares) de una ciudad influye en las tasas de delincuencia. Los diferentes métodos que se comparan en el Capítulo 6, aunque la aplicación que se muestra en él no corresponde a un conjunto de datos sobre crímenes (de nuevo está relacionado con la seguridad del tráfico), son de uso potencial en este contexto.

Por otro lado, la ocurrencia de un delito a corto plazo está íntimamente relacionada con el fenómeno de la repetición. Prácticamente todos los tipos de delito han sido investigados en el contexto de la repetición o casi-repetición, lo que se refiere a la mayor probabilidad de observar un delito en la proximidad (en el espacio y el tiempo) de un evento delictivo previo del mismo tipo. Por lo tanto, algunos autores se han referido al fenómeno de la repetición del delito como un proceso epidemiológico contagioso (Loftin, 1986). De hecho, la prueba de Knox, concebida originalmente para estudios epidemiológicos (Knox and Bartlett, 1964), sigue siendo la principal herramienta para evaluar la magnitud y la extensión espacio-temporal del fenómeno de casi-repetición para un tipo de delito, aunque cada vez son más importantes algunos nuevos enfoques ajenos al test de Knox (Mohler et al., 2011; Reinhart and Greenhouse, 2018).

El principal inconveniente del test de Knox es que no tiene en cuenta la heterogeneidad del riesgo espacio-temporal, lo que en el contexto de la repetición de delitos complica la distinción entre las casi-repeticiones que están conectadas (explicadas por la llamada teoría "boost") y las que no lo están (explicadas por la llamada teoría "flag"). En el Capítulo 4 se ajusta la versión clásica del test de Knox siguiendo el trabajo de Schmertmann (2015). Una modificación de la propuesta hecha por Schmertmann también se proporciona en este Capítulo. La versión ajustada de la prueba se implementa para analizar la magnitud y extensión del fenómeno de casi-repetición considerando un conjunto de datos de robos registrados en València.

La investigación final que se llevó a cabo en el contexto de esta tesis trata sobre la calidad de la geocodificación y la presencia de datos faltantes como consecuencia de eventos no geocodificados. Este tema es de interés para todas las disciplinas que requieren del uso de datos espaciales, pero ha sido particularmente enfatizado en el contexto de la criminología cuantitativa. En 2004, J. Ratcliffe declaró que una tasa de geocodificación (porcentaje de eventos geocodificados) del 85% era una tasa mínimamente aceptable para llevar a cabo un análisis espacial (Ratcliffe, 2004a). El Capítulo 7 contiene una reestimación de esta tasa teniendo en cuenta algunos factores espaciales (niveles de intensidad, *clustering* y agregación) que se pasaron por alto en la primera estimación. Además, el procedimiento propuesto en Ratcliffe (2004a) (basado en el test de Mann-Whitney) se extiende a otros métodos estadísticos de interés. Así, se ha comprobado que el 85% inicialmente propuesto puede ser demasiado bajo determinadas condiciones.

References

Abdel-Aty, M., Chundi, S. S., and Lee, C. (2007). Geo-spatial and log-linear analysis of pedestrian and bicyclist crashes involving school-aged children. *Journal of Safety Research*, 38(5):571–579.

Abdel-Aty, M., Lee, J., Siddiqui, C., and Choi, K. (2013). Geographical unit based analysis in the context of transportation safety planning. *Transportation Research Part A: Policy and Practice*, 49:62–75.

Aguero-Valverde, J. and Jovanis, P. (2008). Analysis of road crash frequency with spatial models. *Transportation Research Record: Journal of the Transportation Research Board*, (2061):55–63.

Alarifi, S. A., Abdel-Aty, M., and Lee, J. (2018). A Bayesian multivariate hierarchical spatial joint model for predicting crash counts by crash type at intersections and segments along corridors. *Accident Analysis & Prevention*, 119:263–273.

Alba-Fernández, M., Ariza-López, F., Jiménez-Gamero, M. D., and Rodríguez-Avi, J. (2016). On the similarity analysis of spatial patterns. *Spatial Statistics*, 18:352–362.

Allévius, B. (2018). scanstatistics: Space-time anomaly detection using scan statistics. *The Journal of Open Source Software*, 3:515.

Amoh-Gyimah, R., Saberi, M., and Sarvi, M. (2017). The effect of variations in spatial units on unobserved heterogeneity in macroscopic crash models. *Analytic Methods in Accident Research*, 13:28–51.

Anastasopoulos, P. C. (2016). Random parameters multivariate tobit and zero-inflated count data models: Addressing unobserved and zero-state heterogeneity in accident injury-severity rate and frequency analysis. *Analytic Methods in Accident Research*, 11:17–32.

Anderson, J. and Hernandez, S. (2017). Roadway classifications and the accident injury severities of heavy-vehicle drivers. *Analytic Methods in Accident Research*, 15:17–28.

Andresen, M. A. (2009). Testing for similarity in area-based spatial patterns: a nonparametric monte carlo approach. *Applied Geography*, 29(3):333–345.

Andresen, M. A. (2016). An area-based nonparametric spatial point pattern test: The test, its applications, and the future. *Methodological Innovations*, 9:2059799116630659.

Andresen, M. A., Curman, A. S., and Linning, S. J. (2017). The trajectories of crime at places: understanding the patterns of disaggregated crime types. *Journal of Quantitative Criminology*, 33(3):427–449.

Andresen, M. A. and Hodgkinson, T. (2018). Predicting property crime risk: An application of risk terrain modeling in Vancouver, Canada. *European Journal on Criminal Policy and Research*, 24(4):373–392.

Ang, Q. W., Baddeley, A., and Nair, G. (2012). Geometrically corrected second order analysis of events on a linear network, with applications to ecology and criminology. *Scandinavian Journal of Statistics*, 39(4):591–617.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*, volume 4. Springer Science & Business Media.

Anselin, L. (1995). Local indicators of spatial association—LISA. *Geographical Analysis*, 27(2):93–115.

Ashby, M. P. (2019). Studying Crime and Place with the Crime Open Database: Social and Behavioural Scienes. *Research Data Journal for the Humanities and Social Sciences*, 1(aop):1–16.

Ashby, M. P. J. and Bowers, K. J. (2013). A comparison of methods for temporal analysis of aoristic crime. *Crime Science*, 2(1):1.

Assunção, R. M., Neves, M. C., Câmara, G., and da Costa Freitas, C. (2006). Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science*, 20(7):797–811.

Baddeley, A., Nair, G., Rakshit, S., and McSwiggan, G. (2017). "Stationary" point processes are uncommon on linear networks. *Stat*, 6(1):68–78.

Baddeley, A., Rubak, E., and Turner, R. (2015). *Spatial point patterns: methodology and applications with R.* CRC Press.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical modeling and analysis for spatial data.* Chapman and Hall/CRC.

Barua, S., El-Basyouny, K., and Islam, M. T. (2016). Multivariate random parameters collision count data models with spatial heterogeneity. *Analytic Methods in Accident Research*, 9:1–15.

Beavon, D. J. K., Brantingham, P. L., and Brantingham, P. J. (1994). The influence of street networks on the patterning of property offenses. *Crime Prevention Studies*, 2:115–148.

Bernasco, W. (2008). Them again? Same-offender involvement in repeat and near repeat burglaries. *European Journal of Criminology*, 5(4):411–431.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1):1–20.

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.

Bíl, M., Andrášik, R., Svoboda, T., and Sedoník, J. (2016). The KDE+ software: a tool for effective identification and ranking of animal-vehicle collision hotspots along networks. *Landscape Ecology*, 31(2):231–237.

Birch, C. P. D., Oom, S. P., and Beecham, J. A. (2007). Rectangular and hexagonal grids used for observation, experiment and simulation in ecology. *Ecological Modelling*, 206(3-4):347–359.

Bivand, R. and Lewin-Koh, N. (2017). *maptools: Tools for Reading and Handling Spatial Objects*. R package version 0.9-2.

Bivand, R. and Piras, G. (2015). Comparing Implementations of Estimation Methods for Spatial Econometrics. *Journal of Statistical Software*, 63(18):1–36.

Bivand, R. and Rundel, C. (2018a). *rgeos: Interface to Geometry Engine - Open Source ('GEOS')*.

Bivand, R. and Rundel, C. (2018b). *rgeos: Interface to Geometry Engine - Open Source ('GEOS')*. R package version 0.3-28.

Bivand, R. and Yu, D. (2017). *spgwr: Geographically Weighted Regression*. R package version 0.6-32.

Bivand, R. S., Pebesma, E., and Gómez-Rubio, V. (2013). *Applied spatial data analysis with R, Second edition*. Springer, NY.

Black, C., Collins, A., and Snell, M. (2001). Encouraging walking: the case of journey-to-school trips in compact urban areas. *Urban Studies*, 38(7):1121–1141.

Block, R. (1979). Community, environment, and violent crime. *Criminology*, 17(1):46–57.

Borrajo, M., González-Manteiga, W., and Martínez-Miranda, M. (2019). Testing for significant differences between two spatial patterns using covariates. *Spatial Statistics*, page 100379.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31(2):144–152.

Brantingham, P. and Brantingham, P. (1995). Criminality of place. *European Journal on Criminal Policy and Research*, 3(3):5–26.

Brantingham, P. J. and Brantingham, P. L. (1975). The spatial patterning of burglary. *The Howard Journal of Criminal Justice*, 14(2):11–23.

Brantingham, P. J. and Brantingham, P. L. (1981). *Environmental criminology*. Sage Publications Beverly Hills, CA.

Brantingham, P. J. and Brantingham, P. L. (1984). *Patterns in crime*. Macmillan New York.

Briz-Redón, Á. (2019). SpNetPrep: An R package using Shiny to facilitate spatial statistics on road networks. *Research Ideas and Outcomes*, 5:e33521.

Briz-Redón, Á., Martínez-Ruiz, F., and Montes, F. (2019a). Estimating the occurrence of traffic accidents near school locations: a case study from valencia (spain) including several approaches. *Accident Analysis & Prevention*, 132:105237.

Briz-Redón, Á., Martínez-Ruiz, F., and Montes, F. (2019b). Identification of differential risk hotspots for collision and vehicle type in a directed linear network. *Accident Analysis & Prevention*, 132:105278.

Briz-Redón, Á., Martínez-Ruiz, F., and Montes, F. (2019c). Investigation of the consequences of the modifiable areal unit problem in macroscopic traffic safety analysis: a case study accounting for scale and zoning. *Accident Analysis & Prevention*, 132:105276.

Briz-Redón, Á., Martinez-Ruiz, F., and Montes, F. (2019d). Reestimating a minimum acceptable geocoding hit rate for conducting a spatial analysis. *International Journal of Geographical Information Science*.

Briz-Redón, Á., Martínez-Ruiz, F., and Montes, F. (2019e). Spatial analysis of traffic accidents near and between road intersections in a directed linear network. *Accident Analysis & Prevention*, 132:105252.

Briz-Redón, Á., Martinez-Ruiz, F., and Montes, F. (2020). Adjusting the Knox test by accounting for spatio-temporal crime risk heterogeneity to analyze near-repeats. *European Journal of Criminology*.

Brunsdon, C., Fotheringham, A. S., and Charlton, M. E. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis*, 28(4):281–298.

Bürkner, P.-C. et al. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28.

Cai, Q., Abdel-Aty, M., Lee, J., and Eluru, N. (2017). Comparative analysis of zonal systems for macro-level crash modeling. *Journal of Safety Research*, 61:157–166.

Cai, Q., Abdel-Aty, M., Lee, J., Wang, L., and Wang, X. (2018). Developing a grouped random parameters multivariate spatial model to explore zonal effects for segment and intersection crash modeling. *Analytic Methods in Accident Research*, 19:1–15.

Caplan, J. M., Kennedy, L. W., Barnum, J. D., and Piza, E. L. (2015). Risk terrain modeling for spatial risk assessment. *Cityscape*, 17(1):7–16.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1).

Castro, M., Paleti, R., and Bhat, C. R. (2012). A latent variable representation of count data models to accommodate spatial and temporal dependence: Application to predicting crash frequency at intersections. *Transportation Research Part B: Methodological*, 46(1):253–272.

Chainey, S., Tompson, L., and Uhlig, S. (2008). The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal*, 21(1-2):4–28.

Chainey, S. P., Curtis-Ham, S. J., Evans, R. M., and Burns, G. J. (2018). Examining the extent to which repeat and near repeat patterns can prevent crime. *Policing: An International Journal*, 41(5):608–622.

Chang, D. (2011). Social crime or spatial crime? Exploring the effects of social, economical, and spatial factors on burglary rates. *Environment and Behavior*, 43(1):26–52.

Chang, L.-Y. and Mannering, F. (1999). Analysis of injury severity and vehicle occupancy in truck-and non-truck-involved accidents. *Accident Analysis & Prevention*, 31(5):579–592.

Chang, L.-Y. and Wang, H.-W. (2006). Analysis of traffic injury severity: An application of non-parametric classification tree techniques. *Accident Analysis & Prevention*, 38(5):1019–1027.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2018). *shiny: Web Application Framework for R*. R package version 1.1.0.

Chavent, M., Kuentz, V., Labenne, A., and Saracco, J. (2017a). *ClustGeo: Hierarchical Clustering with Spatial Constraints*. R package version 2.0.

Chavent, M., Kuentz-Simonet, V., Labenne, A., and Saracco, J. (2017b). ClustGeo: an R package for hierarchical clustering with spatial constraints. *Computational Statistics*, pages 1–24.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335.

Clark, P. J. and Evans, F. C. (1954). Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology*, 35(4):445–453.

Clifton, K. J. and Kreamer-Fults, K. (2007). An examination of the environmental attributes associated with pedestrian–vehicular crashes near public schools. *Accident Analysis & Prevention*, 39(4):708–715.

Cohen, L. E. and Cantor, D. (1981). Residential burglary in the United States: Lifestyle and demographic factors associated with the probability of victimization. *Journal of Research in Crime and Delinquency*, 18(1):113–127.

Coll, B., Moutari, S., and Marshall, A. H. (2013). Hotspots identification and ranking for road safety improvement: An alternative approach. *Accident Analysis & Prevention*, 59:604–617.

Copas, J. (1983). Plotting p against x. *Applied Statistics*, pages 25–31.

Cressie, N. (1993). *Statistics for spatial data.* John Wiley.

Cronie, O. and Van Lieshout, M. N. M. (2018). A non-model-based approach to bandwidth selection for kernel estimators of spatial intensity functions. *Biometrika*, 105(2):455–462.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, Complex Sy:1695.

Das, A., Pande, A., Abdel-Aty, M., and Santos, J. (2008). Characteristics of urban arterial crashes relative to proximity to intersections and injury severity. *Transportation Research Record: Journal of the Transportation Research Board*, (2083):137–144.

Davies, T. and Johnson, S. D. (2015). Examining the relationship between road structure and burglary risk via quantitative network analysis. *Journal of Quantitative Criminology*, 31(3):481–507.

de Melo, S. N., Andresen, M. A., and Matias, L. F. (2018). Repeat and near-repeat victimization in Campinas, Brazil: New explanations from the Global South. *Security Journal*, 31(1):364–380.

Dell'Acqua, G., Russo, F., and Biancardo, S. A. (2013). Risk-type density diagrams by crash type on two-lane rural roads. *Journal of Risk Research*, 16(10):1297–1314.

DeLuca, P. and Kanaroglou, P. (2008). Effects of alternative point pattern geocoding procedures on first and second order statistical measures. *Journal of Spatial Science*, 53(1):131–141.

Diggle, P., Morris, S., Elliott, P., and Shaddick, G. (1997). Regression modelling of disease risk in relation to point sources. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 160(3):491–505.

Diggle, P., Zheng, P., and Durr, P. (2005). Nonparametric estimation of spatial segregation in a multivariate point process: bovine tuberculosis in Cornwall, UK. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):645–658.

Diggle, P. J. (2013). *Statistical analysis of spatial and spatio-temporal point patterns*. Chapman and Hall/CRC.

Diggle, P. J. and Giorgi, E. (2019). *Model-based geostatistics for global public health: methods and applications*. CRC Press.

Diggle, P. J., Moraga, P., Rowlingson, B., Taylor, B. M., et al. (2013). Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm. *Statistical Science*, 28(4):542–563.

Diggle, P. J. and Rowlingson, B. S. (1994). A conditional approach to point process modelling of elevated risk. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 157(3):433–440.

Dong, H., Wu, M., Ding, X., Chu, L., Jia, L., Qin, Y., and Zhou, X. (2015). Traffic zone division based on big data from mobile phone base stations. *Transportation Research Part C: Emerging Technologies*, 58:278–291.

Duque, J. C., Laniado, H., and Polo, A. (2018). S-maup: Statistical test to measure the sensitivity to the modifiable areal unit problem. *PloS one*, 13(11):e0207377.

Eckardt, M. and Mateu, J. (2017). Second-order and local characteristics of network intensity functions. *arXiv preprint arXiv:1712.01555*.

Eckardt, M. and Mateu, J. (2018). Point patterns occurring on complex structures in space and space-time: An alternative network approach. *Journal of Computational and Graphical Statistics*, 27(2):312–322.

Eugster, M. J. and Schlesinger, T. (2013). osmar: OpenStreetMap and R. *The R Journal*, 5(1):53–63.

Fan, W., Kane, M. R., and Haile, E. (2015). Analyzing severity of vehicle crashes at highway-rail grade crossings: multinomial logit modeling. In *Journal of the Transportation Research Forum*, volume 54, pages 39–56.

Faraway, J. J. (2016). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. Chapman and Hall/CRC.

Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. John Wiley & Sons.

Fotheringham, A. S. and Wong, D. W. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A*, 23(7):1025–1044.

Fox, J. (1991). *Regression diagnostics: An introduction*, volume 79. Sage.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41.

Frith, M. J., Johnson, S. D., and Fry, H. M. (2017). Role of the street network in burglars'spatial decision-making. *Criminology*, 55(2):344–376.

Fuentes-Santos, I., González-Manteiga, W., and Mateu, J. (2017). A nonparametric test for the comparison of first-order structures of spatial point processes. *Spatial Statistics*, 22:240–260.

Fyhri, A. and Hjorthol, R. (2009). Children's independent mobility to school, friends and leisure activities. *Journal of Transport Geography*, 17(5):377–384.

Gabriel, E. and Diggle, P. J. (2009). Second-order analysis of inhomogeneous spatio-temporal point process data. *Statistica Neerlandica*, 63(1):43–51.

Gabry, J. and Mahr, T. (2018). *bayesplot: Plotting for Bayesian Models*. R package version 1.6.0.

Gattis, J. and Low, S. T. (1998). Intersection angle geometry and the driver's field of view. *Transportation Research Record*, 1612(1):10–16.

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Georges, D. E. (1978). The geography of crime and violence: A spatial and ecological perspective. Association of American Geographers Washington, DC.

Getis, A. and Ord, J. (1992). The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis*, 24(3).

Gini, C. (1912). Variabilità e mutabilità. *Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi.*

Glasner, P., Johnson, S. D., and Leitner, M. (2018). A comparative analysis to forecast apartment burglaries in Vienna, Austria, based on repeat and near repeat victimization. *Crime Science*, 7(1):9.

Glasner, P. and Leitner, M. (2016). Evaluating the impact the weekday has on near-repeat victimization: A spatio-temporal analysis of street robberies in the city of Vienna, Austria. *ISPRS International Journal of Geo-Information*, 6(1):3.

Golob, T. F., Recker, W. W., and Leonard, J. D. (1987). An analysis of the severity and incident duration of truck-involved freeway accidents. *Accident Analysis & Prevention*, 19(5):375–395.

Gomes, M. J. T. L., Cunto, F., and da Silva, A. R. (2017). Geographically weighted negative binomial regression applied to zonal level safety performance models. *Accident Analysis & Prevention*, 106:254–261.

Gordon, J. E. (1949). The epidemiology of accidents. *American Journal of Public Health and the Nations Health*, 39(4):504–515.

Graul, C. (2016). *leafletR: Interactive Web-Maps Based on the Leaflet JavaScript Library*. R package version 0.4-0.

Groff, E. R. and Taniguchi, T. (2018). Micro-Level Policing for Preventing Near Repeat Residential Burglary: Final Monograph. Washington, DC: Police Foundation.

Grolemund, G. and Wickham, H. (2011). Dates and Times Made Easy with lubridate. *Journal of Statistical Software*, 40(3):1–25.

Guo, D. (2008). Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). *International Journal of Geographical Information Science*, 22(7):801–823.

Guo, Q., Xu, P., Pei, X., Wong, S., and Yao, D. (2017). The effect of road network patterns on pedestrian safety: A zone-based Bayesian spatial modeling approach. *Accident Analysis & Prevention*, 99:114–124.

Gómez-Rubio, V., Ferrándiz-Ferragud, J., and López-Quílez, A. (2005). Detecting clusters of disease with R. *Journal of Geographical Systems*, 7(2):189–206.

Gómez-Rubio, V., Moraga, P., Molitor, J., and Rowlingson, B. (2019). Dclusterm: Model-based detection of disease clusters. *Journal of Statistical Software, Articles*, 90(14):1–26.

Haberman, C. P. and Ratcliffe, J. H. (2012). The predictive policing challenges of near repeat armed street robberies. *Policing: A Journal of Policy and Practice*, 6(2):151–166.

Hadayeghi, A., Shalaby, A. S., and Persaud, B. N. (2010). Development of planning level transportation safety tools using Geographically Weighted Poisson Regression. *Accident Analysis & Prevention*, 42(2):676–688.

Hahn, U. (2012). A studentized permutation test for the comparison of spatial point patterns. *Journal of the American Statistical Association*, 107(498):754–764.

Haklay, M. and Weber, P. (2008). OpenStreetMap: User-generated street maps. *IEEE Pervasive Computing*, 7(4):12–18.

Hao, W. and Daniel, J. (2014). Motor vehicle driver injury severity study under various traffic control at highway-rail grade crossings in the united states. *Journal of Safety Research*, 51:41–48.

Harada, Y. and Shimada, T. (2006). Examining the impact of the precision of address geocoding on estimated density of crime locations. *Computers & Geosciences*, 32(8):1096–1107.

Harirforoush, H. and Bellalite, L. (2016). A new integrated GIS-based analysis to detect hotspots: a case study of the city of Sherbrooke. *Accident Analysis & Prevention.*

Harries, K. D. (1973). *The geography of crime and justice.* McGraw-Hill.

Harries, K. D. et al. (1999). Mapping crime: Principle and practice. Technical report, US Department of Justice, Office of Justice Programs, National Institute of Justice, Crime Mapping Research Center.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

Harwood, D. W., Council, F., Hauer, E., Hughes, W., and Vogt, A. (2000). Prediction of the expected safety performance of rural two-lane highways. Technical report, United States. Federal Highway Administration.

Harzing, A. W. (2007). Publish or perish [computer software]. *Available from.*

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications.

Hijmans, R. J. (2019). *raster: Geographic Data Analysis and Modeling*. R package version 2.8-19.

Hillier, B. (2004). Can streets be made safe? *Urban Design International*, 9(1):31–45.

Hino, K. and Amemiya, M. (2019). Spatiotemporal analysis of burglary in multifamily housing in Fukuoka City, Japan. *Cities*, 90:15–23.

Hipp, J. R. (2011). Spreading the wealth: The effect of the distribution of income and race/ethnicity across households and neighborhoods on city crime trajectories. *Criminology*, 49(3):631–665.

Hjorthol, R. and Fyhri, A. (2009). Do organized leisure activities for children encourage car-use? *Transportation Research Part A: Policy and Practice*, 43(2):209–218.

Hoppe, L. and Gerell, M. (2019). Near-repeat burglary patterns in Malmö: Stability and change over time. *European Journal of Criminology*, 16(1):3–17.

Hosseinpour, M., Yahaya, A. S., and Sadullah, A. F. (2014). Exploring the effects of roadway characteristics on the frequency and severity of head-on crashes: Case studies from Malaysian Federal Roads. *Accident Analysis & Prevention*, 62:209–222.

Huang, H., Abdel-Aty, M., and Darwiche, A. (2010). County-level crash risk analysis in Florida: Bayesian spatial modeling. *Transportation Research Record: Journal of the Transportation Research Board*, (2148):27–37.

Huang, H., Song, B., Xu, P., Zeng, Q., Lee, J., and Abdel-Aty, M. (2016). Macro and micro models for zonal crash prediction with application in hot zones identification. *Journal of Transport Geography*, 54:248–256.

Huang, H., Zhou, H., Wang, J., Chang, F., and Ma, M. (2017). A multivariate spatial model of crash frequency by transportation modes for urban intersections. *Analytic Methods in Accident Research*, 14:10–21.

Huttenlocher, D. P., Klanderman, G. A., and Rucklidge, W. J. (1993). Comparing images using the Hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence*, 15(9):850–863.

Hwang, J., Joh, K., and Woo, A. (2017). Social inequalities in child pedestrian traffic injuries: Differences in neighborhood built environments near schools in Austin, TX, USA. *Journal of Transport & Health*, 6:40–49.

Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008). *Statistical analysis and modelling of spatial point patterns*, volume 70. John Wiley & Sons.

Imprialou, M.-I. M., Quddus, M., Pitfield, D. E., and Lord, D. (2016). Re-visiting crash–speed relationships: A new perspective in crash modelling. *Accident Analysis & Prevention*, 86:173–185.

Jacquez, G. M. (1996). A k nearest neighbour test for space–time interaction. *Statistics in Medicine*, 15(18):1935–1949.

Johnson, S. D. (2008). Repeat burglary victimisation: a tale of two theories. *Journal of Experimental Criminology*, 4(3):215–240.

Johnson, S. D. (2010). A brief history of the analysis of crime concentration. *European Journal of Applied Mathematics*, 21(4-5):349–370.

Johnson, S. D., Bernasco, W., Bowers, K. J., Elffers, H., Ratcliffe, J., Rengert, G., and Townsley, M. (2007). Space–time patterns of risk: A cross national assessment of residential burglary victimization. *Journal of Quantitative Criminology*, 23(3):201–219.

Johnson, S. D. and Bowers, K. J. (2010). Permeability and burglary risk: are cul-de-sacs safer? *Journal of Quantitative Criminology*, 26(1):89–111.

Kahle, D. and Wickham, H. (2013a). ggmap: Spatial visualization with ggplot2. *The R Journal*, 5(1):144–161.

Kahle, D. and Wickham, H. (2013b). ggmap: Spatial Visualization with ggplot2. *The R Journal*, 5(1):144–161.

Kansky, K. and Danscoine, P. (1989). Measures of network structure. *FLUX Cahiers scientifiques internationaux Réseaux et Territoires*, 5(1):89–121.

Kelly, J. A. and Fu, M. (2014). Sustainable school commuting–understanding choices and identifying opportunities: A case study in Dublin, Ireland. *Journal of Transport Geography*, 34:221–230.

Kelsall, J. E. and Diggle, P. J. (1995). Non-parametric estimation of spatial variation in relative risk. *Statistics in Medicine*, 14(21-22):2335–2342.

Kelsall, J. E. and Diggle, P. J. (1998). Spatial variation in risk of disease: a non-parametric binary regression approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(4):559–573.

Kelsall, J. E., Diggle, P. J., et al. (1995). Kernel estimation of relative risk. *Bernoulli*, 1(1-2):3–16.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Khazraee, S. H., Johnson, V., and Lord, D. (2018). Bayesian Poisson hierarchical models for crash data analysis: Investigating the impact of model choice on site-specific predictions. *Accident Analysis & Prevention*, 117:181–195.

Killick, R. and Eckley, I. (2014). changepoint: An R package for changepoint analysis. *Journal of Statistical Software*, 58(3):1–19.

Killick, R., Fearnhead, P., and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598.

Kim, J.-K., Kim, S., Ulfarsson, G. F., and Porrello, L. A. (2007). Bicyclist injury severities in bicycle–motor vehicle accidents. *Accident Analysis & Prevention*, 39(2):238–251.

Kingham, S., Sabel, C. E., and Bartie, P. (2011). The impact of the 'school run'on road traffic accidents: A spatio-temporal analysis. *Journal of Transport Geography*, 19(4):705–711.

Kinney, J. B., Brantingham, P. L., Wuschke, K., Kirk, M. G., and Brantingham, P. J. (2008). Crime attractors, generators and detractors: Land use and urban crime opportunities. *Built Environment*, 34(1):62–74.

Knox, E. G. and Bartlett, M. S. (1964). The detection of space-time interactions. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 13(1):25–30.

Krige, D. (1960). On the departure of ore value distributions from the lognormal model in south african gold mines. *Journal of the Southern African Institute of Mining and Metallurgy*, 61(4):231–244.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics-Theory and methods*, 26(6):1481–1496.

Kulldorff, M. and Hjalmars, U. (1999). The Knox method and other tests for space-time interaction. *Biometrics*, 55(2):544–552.

Kumfer, W., Harkey, D., Lan, B., Srinivasan, R., Carter, D., Patel Nujjetty, A., Eigen, A. M., and Tan, C. (2019). Identification of Critical Intersection Angle through Crash Modification Functions. *Transportation Research Record*, page 0361198119828682.

Lee, G., Park, Y., Kim, J., and Cho, G.-H. (2016). Association between intersection characteristics and perceived crash risk among school-aged children. *Accident Analysis & Prevention*, 97:111–121.

Lee, J., Abdel-Aty, M., and Cai, Q. (2017). Intersection crash prediction modeling with macro-level data from various geographic units. *Accident Analysis & Prevention*, 102:213–226.

Lee, J., Abdel-Aty, M., and Jiang, X. (2014). Development of zone system for macro-level traffic safety analysis. *Journal of Transport Geography*, 38:13–21.

Lee, S.-I., Lee, M., Chun, Y., and Griffith, D. A. (2018). Uncertainty in the effects of the modifiable areal unit problem under different levels of spatial autocorrelation: a simulation study. *International Journal of Geographical Information Science*, pages 1–20.

Leung, K. Y., Astroza, S., Loo, B. P., and Bhat, C. R. (2019). An environment-people interactions framework for analysing children's extra-curricular activities and active transport. *Journal of Transport Geography*, 74:341–358.

Lindgren, F. and Rue, H. (2015). Bayesian spatial modelling with R-INLA. *Journal of Statistical Software*, 63(19).

Liu, Z. and Zhao, S. (2015). Characteristics of road network forms in historic districts of Japan. *Frontiers of Architectural Research*, 4(4):296–307.

Llera, R. F. and Pérez, M. M. (2012). Colegios concertados y selección de escuela en España: un círculo vicioso. *Presupuesto y gasto público*, 67:97–118.

Loftin, C. (1986). Assaultive violence as a contagious social process. *Bulletin of the New York Academy of Medicine*, 62(5):550.

Loidl, M., Wallentin, G., Wendel, R., and Zagel, B. (2016). Mapping bicycle crash risk patterns on the local scale. *Safety*, 2(3):17.

Lord, D. and Mannering, F. (2010). The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*, 44(5):291–305.

Malczewski, J. and Poetz, A. (2005). Residential burglaries and neighborhood socioeconomic context in London, Ontario: global and local regression analysis. *The Professional Geographer*, 57(4):516–529.

Manley, D. (2014). Scale, aggregation, and the modifiable areal unit problem. *Handbook of Regional Science*, pages 1157–1171.

Mannering, F. L. and Bhat, C. R. (2014). Analytic methods in accident research: Methodological frontier and future directions. *Analytic Methods in Accident Research*, 1:1–22.

Mannering, F. L., Shankar, V., and Bhat, C. R. (2016). Unobserved heterogeneity and the statistical analysis of highway accident data. *Analytic Methods in Accident Research*, 11:1–16.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2 Part 1):209–220.

Marshall, E. and Spiegelhalter, D. (2003). Approximate cross-validatory predictive checks in disease mapping models. *Statistics in Medicine*, 22(10):1649–1660.

Martínez, L. M., Viegas, J. M., and Silva, E. A. (2009). A traffic analysis zone definition: a new methodology and algorithm. *Transportation*, 36(5):581–599.

Martins, T. G., Simpson, D., Lindgren, F., and Rue, H. (2013). Bayesian computing with INLA: new features. *Computational Statistics & Data Analysis*, 67:68–83.

Matérn, B. (1986). Spatial variation, volume 36 of Lecture Notes in Statistics.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, 58(8):1246–1266.

Matkan, A. A., Mohaymany, A. S., Mirbagheri, B., and Shahri, M. (2011). Explorative spatial analysis of traffic accidents using GWPR model for urban safety planning. In *3rd International Conference on Road Safety and Simulation*, pages 14–16.

McElduff, F., Cortina-Borja, M., Chan, S.-K., and Wade, A. (2010). When t-tests or wilcoxon-mann-whitney tests won't do. *Advances in Physiology Education*, 34(3):128–133.

McSwiggan, G., Baddeley, A., and Nair, G. (2017). Kernel density estimation on a linear network. *Scandinavian Journal of Statistics*, 44(2):324–345.

McSwiggan, G., Baddeley, A., and Nair, G. (2019). Estimation of relative risk for events on a linear network. *Statistics and Computing*, pages 1–16.

Mennis, J., Harris, P. W., Obradovic, Z., Izenman, A. J., Grunwald, H. E., and Lockwood, B. (2011). The effect of neighborhood characteristics and spatial spillover on urban juvenile delinquency and recidivism. *The Professional Geographer*, 63(2):174–192.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Meyer, S., Held, L., and Höhle, M. (2017). Spatio-temporal analysis of epidemic phenomena using the r package surveillance. *Journal of Statistical Software*, 77(11).

Miaou, S.-P. and Lord, D. (2003). Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes methods. *Transportation Research Record: Journal of the Transportation Research Board*, (1840):31–40.

Miles, J. (2014). Tolerance and variance inflation factor. *Wiley StatsRef: Statistics Reference Online.*

Moellering, H. (1976). The potential uses of a computer animated film in the analysis of geographical patterns of traffic crashes. *Accident Analysis & Prevention*, 8(4):215–227.

Mohler, G. (2014). Marked point process hotspot maps for homicide and gun crime prediction in Chicago. *International Journal of Forecasting*, 30(3):491–497.

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., and Tita, G. E. (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108.

Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998). Log Gaussian Cox Processes. *Scandinavian journal of statistics*, 25(3):451–482.

Moradi, M. M., Cronie, O., Rubak, E., Lachieze-Rey, R., Mateu, J., and Baddeley, A. (2019). Resample-smoothing of Voronoi intensity estimators. *Statistics and Computing*, 29(5):995–1010.

Moradi, M. M., Rodríguez-Cortés, F. J., and Mateu, J. (2018). On kernel-based intensity estimation of spatial point patterns on linear networks. *Journal of Computational and Graphical Statistics*, 27(2):302–311.

Moran, P. A. (1950a). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23.

Moran, P. A. (1950b). A test for the serial independence of residuals. *Biometrika*, 37(1/2):178–181.

Moreto, W. D., Piza, E. L., and Caplan, J. M. (2014). “A plague on both your houses?": Risks, repeats and reconsiderations of urban residential burglary. *Justice Quarterly*, 31(6):1102–1126.

Mullen, K. M., Ardia, D., Gil, D. L., Windover, D., Cline, J., et al. (2011). DEoptim: An R Package for Global Optimization by Differential Evolution. *Journal of Statistical Software*, 40(i06).

Murakami, M., Higuchi, K., and Shibayama, A. (2004). Relationship between convenience store robberies and road environment. In *Recent Advances in Design and Decision Support Systems in Architecture and Urban Planning*, pages 341–356. Springer.

Murray, Å. (1998). The home and school background of young drivers involved in traffic accidents. *Accident Analysis & Prevention*, 30(2):169–182.

Nakaya, T., Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2005). Geographically weighted Poisson regression for disease association mapping. *Statistics in Medicine*, 24(17):2695–2717.

Neyman, J. and Scott, E. L. (1958). Statistical approach to problems of cosmology. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(1):1–29.

Nie, K., Wang, Z., Du, Q., Ren, F., and Tian, Q. (2015). A network-constrained integrated method for detecting spatial cluster and risk location of traffic crash: A case study from Wuhan, China. *Sustainability*, 7(3):2662–2677.

Nightingale, E., Parvin, N., Seiberlich, C., Savolainen, P. T., and Pawlovich, M. (2017). Investigation of Skew Angle and Other Factors Influencing Crash Frequency at High-Speed Rural Intersections. *Transportation Research Record*, 2636(1):9–14.

Nobles, M. R., Ward, J. T., and Tillyer, R. (2016). The impact of neighborhood context on spatiotemporal patterns of burglary. *Journal of Research in Crime and Delinquency*, 53(5):711–740.

Ogata, Y. (1998). Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, 50(2):379–402.

Okabe, A., Satoh, T., and Sugihara, K. (2009). A kernel density estimation method for networks, its computational method and a GIS-based tool. *International Journal of Geographical Information Science*, 23(1):7–32.

Okabe, A. and Sugihara, K. (2012). *Spatial analysis along networks: statistical and computational methods*. John Wiley & Sons.

Oliver, M. N., Matthews, K. A., Siadaty, M., Hauck, F. R., and Pickle, L. W. (2005). Geographic bias related to geocoding in epidemiologic studies. *International Journal of Health Geographics*, 4(1):29.

O'Neill, W. A. (1991). Developing optimal transportation analysis zones using GIS. In *Proceedings of the 1991 Geographic Information Systems (GIS) for Transportation SymposiumCo-sponsored by the Federal Highway Administration, the Highway Engineering Exchange Program, Transportation Research Board, and Urban & Regional Information Systems Association.*

Openshaw, S. (1977). Optimal zoning systems for spatial interaction models. *Environment and Planning A*, 9(2):169–184.

Openshaw, S. (1978). An empirical study of some zone-design criteria. *Environment and Planning A*, 10(7):781–794.

Openshaw, S. (1979). A million or so correlation coefficients, three experiments on the modifiable areal unit problem. *Statistical Applications in the Spatial Science*, pages 127–144.

Openshaw, S. (1984). The modifiable areal unit problem. *Concepts and Techniques in Modern Geography.*

OpenStreetMap contributors (2017). Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org.

Ornstein, J. T. and Hammond, R. A. (2017). The burglary boost: A note on detecting contagion using the Knox test. *Journal of Quantitative Criminology*, 33(1):65–75.

Padgham, M., Rudis, B., Lovelace, R., and Salmon, M. (2017). osmdata. *The Journal of Open Source Software*, 2(14).

Papadimitriou, E., Filtness, A., Theofilatos, A., Ziakopoulos, A., Quigley, C., and Yannis, G. (2019). Review and ranking of crash risk factors related to the road infrastructure. *Accident Analysis & Prevention*, 125:85–97.

Park, J., Abdel-Aty, M., and Lee, J. (2018). School zone safety modeling in countermeasure evaluation and decision. *Transportmetrica A: Transport Science*, pages 1–16.

Pearson, K. (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 187:253–318.

Pebesma, E. J. (2004). Multivariable geostatistics in S: the gstat package. *Computers & Geosciences*, 30(7):683–691.

Pebesma, E. J. and Bivand, R. S. (2005). Classes and methods for spatial data in R. *R News*, 5(2):9–13.

Piza, E. L. and Carter, J. G. (2018). Predicting initiator and near repeat events in spatiotemporal crime patterns: An analysis of residential burglary and motor vehicle theft. *Justice Quarterly*, 35(5):842–870.

Polvi, N., Looman, T., Humphries, C., and Pease, K. (1991). The time course of repeat burglary victimization. *The British Journal of Criminology*, 31(4):411–414.

Powell, Z. A., Grubb, J. A., and Nobles, M. R. (2018). A Near Repeat Examination of Economic Crimes. *Crime & Delinquency*, page 0011128718811927.

Price, K., Storn, R. M., and Lampinen, J. A. (2006). *Differential evolution: a practical approach to global optimization.* Springer Science & Business Media.

Qin, X., Ivan, J. N., and Ravishanker, N. (2004). Selecting exposure measures in crash rate prediction for two-lane highway segments. *Accident Analysis & Prevention*, 36(2):183–191.

Quddus, M. A. (2008). Modelling area-wide count outcomes with spatial correlation and heterogeneity: an analysis of London crash data. *Accident Analysis & Prevention*, 40(4):1486–1497.

R Core Team (2018). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Rakshit, S., Baddeley, A., and Nair, G. (2019a). Efficient Code for Second Order Analysis of Events on a Linear Network. *Journal of Statistical Software*, 90(1):1–37.

Rakshit, S., Davies, T., Moradi, M. M., McSwiggan, G., Nair, G., Mateu, J., and Baddeley, A. (2019b). Fast Kernel Smoothing of Point Patterns on a Large Network using Two-dimensional Convolution. *International Statistical Review*.

Ramis, R., Diggle, P., Cambra, K., and López-Abente, G. (2011). Prostate cancer and industrial pollution: Risk around putative focus in a multi-source scenario. *Environment International*, 37(3):577–585.

Ratcliffe, J. H. (2002). Aoristic signatures and the spatio-temporal analysis of high volume crime patterns. *Journal of Quantitative Criminology*, 18(1):23–43.

Ratcliffe, J. H. (2004a). Geocoding crime and a first estimate of a minimum acceptable hit rate. *International Journal of Geographical Information Science*, 18(1):61–72.

Ratcliffe, J. H. (2004b). The hotspot matrix: A framework for the spatio-temporal targeting of crime reduction. *Police Practice and Research*, 5(1):5–23.

Ratcliffe, J. H. (2009). Near repeat calculator (version 1.3). *Temple University, Philadelphia, PA and the National Institute of Justice, Washington, DC.*

Ratcliffe, J. H. and McCullagh, M. J. (1998). Identifying repeat victimization with GIS. *The British Journal of Criminology*, 38(4):651–662.

Ratcliffe, J. H. and Rengert, G. F. (2008). Near-repeat patterns in Philadelphia shootings. *Security Journal*, 21(1-2):58–76.

Reeve, N. F., Fanshawe, T. R., Keegan, T. J., Stewart, A. G., and Diggle, P. J. (2013). Spatial analysis of health effects of large industrial incinerators in England, 1998–2008: a study using matched case–control areas. *BMJ open*, 3(1):e001847.

Reinhart, A. and Greenhouse, J. (2018). Self-exciting point processes with spatial covariates: modelling the dynamics of crime. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(5):1305–1329.

Ripley, B. D. (1977). Modelling spatial patterns. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 172–212.

Robey, R. R. and Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology*, 45(2):283–288.

Rothman, L., Buliung, R., Howard, A., Macarthur, C., and Macpherson, A. (2017a). The school environment and student car drop-off at elementary schools. *Travel Behaviour and Society*, 9:50–57.

Rothman, L., Howard, A., Buliung, R., Macarthur, C., Richmond, S. A., and Macpherson, A. (2017b). School environments and social risk factors for child pedestrian-motor vehicle collisions: a case-control study. *Accident Analysis & Prevention*, 98:252–258.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392.

Savolainen, P. and Mannering, F. (2007). Probabilistic models of motorcyclists' injury severities in single-and multi-vehicle crashes. *Accident Analysis & Prevention*, 39(5):955–963.

Schmertmann, C. P. (2015). Adjusting for population shifts and covariates in space–time interaction tests. *Biometrics*, 71(3):714–720.

Schmertmann, C. P., Assunção, R. M., and Potter, J. E. (2010). Knox meets Cox: Adapting epidemiological space-time statistics to demographic studies. *Demography*, 47(3):629–650.

Schnell, C., Braga, A. A., and Piza, E. L. (2017). The influence of community areas, neighborhood clusters, and street segments on the spatial variability of violent crime in Chicago. *Journal of Quantitative Criminology*, 33(3):469–496.

Serra, L., Juan, P., Varga, D., Mateu, J., and Saez, M. (2013). Spatial pattern modelling of wildfires in Catalonia, Spain 2004–2008. *Environmental modelling & software*, 40:235–244.

Shankar, V. and Mannering, F. (1996). An exploratory multinomial logit analysis of single-vehicle motorcycle accident severity. *Journal of Safety Research*, 27(3):183–194.

Shannon, L. W. (1954). The spatial distribution of criminal offenses by states. *The Journal of Criminal Law, Criminology, and Police Science*, 45(3):264–273.

Shirazi, M. and Lord, D. (2018). Characteristics Based Heuristics to Select a Logical Distribution between the Poisson-Gamma and the Poisson-Lognormal for Crash Data Modeling. *Paper Presented at the 97th Annual Meeting of the Transportation Research Board.*

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.

Siddiqui, C. and Abdel-Aty, M. (2012). Nature of modeling boundary pedestrian crashes at zones. *Transportation Research Record*, 2299(1):31–40.

Silverman, B. W. (2018). *Density estimation for statistics and data analysis.* Routledge.

Snow, J. (1855). *On the mode of communication of cholera.* John Churchill.

Sparks, R. F. (1981). Multiple victimization: Evidence, theory, and future research. *J. Crim. L. & Criminology*, 72:762.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Steenbeek, W. and Weisburd, D. (2016). Where the action is in crime? An examination of variability of crime across different spatial units in The Hague, 2001–2009. *Journal of Quantitative Criminology*, 32(3):449–469.

Stephenson, L. K. (1974). Spatial dispersion of intra-urban juvenile delinquency. *Journal of Geography*, 73(3):20–26.

Stern, H. S. and Cressie, N. (2000). Posterior predictive model checks for disease mapping models. *Statistics in Medicine*, 19(17-18):2377–2397.

Stoll, P. and Bergius, E. (2005). Pattern and process: competition causes regular spacing of individuals within plant populations. *Journal of Ecology*, 93(2):395–403.

Sturup, J., Rostami, A., Gerell, M., and Sandholm, A. (2018). Near-repeat shootings in contemporary Sweden 2011 to 2015. *Security Journal*, 31(1):73–92.

Susser, E. and Bresnahan, M. (2001). Origins of epidemiology. *Annals of the New York Academy of Sciences*, 954:6–18.

Thakali, L., Kwon, T. J., and Fu, L. (2015). Identification of crash hotspots using kernel density estimation and kriging methods: a comparison. *Journal of Modern Transportation*, 23(2):93–106.

Thomas, I. (1996). Spatial data aggregation: exploratory analysis of road accidents. *Accident Analysis & Prevention*, 28(2):251–264.

Townsley, M., Homel, R., and Chaseling, J. (2003). Infectious burglaries. A test of the near repeat hypothesis. *British Journal of Criminology*, 43(3):615–633.

Ukkusuri, S., Miranda-Moreno, L. F., Ramadurai, G., and Isa-Tavarez, J. (2012). The role of built environment on pedestrian crash frequency. *Safety Science*, 50(4):1141–1151.

Van Patten, I. T., McKeldin-Coner, J., and Cox, D. (2009). A microspatial analysis of robbery: Prospective hot spotting in a small city. *Crime Mapping: A journal of research and practice*, 1(1):7–32.

Walker, I. (2007). Drivers overtaking bicyclists: Objective data on the effects of riding position, helmet use, vehicle type and apparent gender. *Accident Analysis & Prevention*, 39(2):417–425.

Walker, K. (2016). tigris: An R package to access and work with geographic data from the US Census Bureau. *The R Journal*, 8(2):231–242.

Wang, K., Ivan, J. N., Ravishanker, N., and Jackson, E. (2017). Multivariate poisson lognormal modeling of crashes by type and severity on rural two lane highways. *Accident Analysis & Prevention*, 99:6–19.

Warsh, J., Rothman, L., Slater, M., Steverango, C., and Howard, A. (2009). Are school zones effective? An examination of motor vehicle versus child pedestrian crashes near schools. *Injury Prevention*, 15(4):226–229.

Weisburd, D. (2015). The law of crime concentration and the criminology of place. *Criminology*, 53(2):133–157.

Wells, W., Wu, L., and Ye, X. (2012). Patterns of near-repeat gun assaults in Houston. *Journal of Research in Crime and Delinquency*, 49(2):186–212.

Wheeler, A. P., Steenbeek, W., and Andresen, M. A. (2018). Testing for similarity in area-based spatial patterns: Alternative methods to Andresen's spatial point pattern test. *Transactions in GIS*, 22(3):760–774.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wier, M., Weintraub, J., Humphreys, E. H., Seto, E., and Bhatia, R. (2009). An area-level model of vehicle-pedestrian injury collisions with implications for land use and transportation planning. *Accident Analysis & Prevention*, 41(1):137–145.

Wilson, E. J., Marshall, J., Wilson, R., and Krizek, K. J. (2010). By foot, bus or car: children's school travel and school choice policy. *Environment and Planning A*, 42(9):2168–2185.

Xie, K., Wang, X., Ozbay, K., and Yang, H. (2014). Crash frequency modeling for signalized intersections in a high-density urban road network. *Analytic Methods in Accident Research*, 2:39–51.

Xie, Z. and Yan, J. (2008). Kernel density estimation of traffic accidents in a network space. *Computers, Environment and Urban Systems*, 32(5):396–406.

Xie, Z. and Yan, J. (2013). Detecting traffic accident clusters with network kernel density estimation and local spatial statistics: an integrated approach. *Journal of Transport Geography*, 31:64–71.

Xu, P. and Huang, H. (2015). Modeling crash spatial heterogeneity: random parameter versus geographically weighting. *Accident Analysis & Prevention*, 75:16–25.

Xu, P., Huang, H., and Dong, N. (2018). The modifiable areal unit problem in traffic safety: basic issue, potential solutions and future research. *Journal of Traffic and Transportation Engineering (English Edition)*, 5(1):73–82.

Xu, P., Huang, H., Dong, N., and Abdel-Aty, M. (2014). Sensitivity analysis in the context of regional safety modeling: identifying and assessing the modifiable areal unit problem. *Accident Analysis & Prevention*, 70:110–120.

Xu, P., Huang, H., Dong, N., and Wong, S. (2017). Revisiting crash spatial heterogeneity: a Bayesian spatially varying coefficients approach. *Accident Analysis & Prevention*, 98:330–337.

Yang, H., Ozbay, K., Ozturk, O., and Yildirimoglu, M. (2013). Modeling work zone crash frequency by quantifying measurement errors in work zone length. *Accident Analysis & Prevention*, 55:192–201.

Yasmin, S., Eluru, N., Lee, J., and Abdel-Aty, M. (2016). Ordered fractional split approach for aggregate injury severity modeling. *Transportation Research Record*, 2583(1):119–126.

Ye, C., Chen, Y., and Li, J. (2018). Investigating the Influences of Tree Coverage and Road Density on Property Crime. *ISPRS International Journal of Geo-Information*, 7(3):101.

Ye, X., Pendyala, R. M., Washington, S. P., Konduri, K., and Oh, J. (2009). A simultaneous equations model of crash frequency by collision type for rural intersections. *Safety Science*, 47(3):443–452.

Youstin, T. J., Nobles, M. R., Ward, J. T., and Cook, C. L. (2011). Assessing the generalizability of the near repeat phenomenon. *Criminal Justice and Behavior*, 38(10):1042–1063.

Yu, C.-Y. (2015). How differences in roadways affect school travel safety. *Journal of the American Planning Association*, 81(3):203–220.

Yu, C.-Y. and Zhu, X. (2016). Planning for safe schools: impacts of school siting and surrounding environments on traffic safety. *Journal of Planning Education and Research*, 36(4):476–486.

Zandbergen, P. A. (2009). Geocoding quality and implications for spatial analysis. *Geography Compass*, 3(2):647–680.

Zeng, Q., Wen, H., Huang, H., and Abdel-Aty, M. (2017). A Bayesian spatial random parameters Tobit model for analyzing crash rates on roadway segments. *Accident Analysis & Prevention*, 100:37–43.

Zhai, X., Huang, H., Gao, M., Dong, N., and Sze, N. (2018a). Boundary crash data assignment in zonal safety analysis: an iterative approach based on data augmentation and Bayesian spatial model. *Accident Analysis & Prevention*, 121:231–237.

Zhai, X., Huang, H., Xu, P., and Sze, N. (2018b). The influence of zonal configurations on macro-level crash modeling. *Transportmetrica A: Transport Science*, pages 1–18.

Zhang, Y., Zhao, J., Ren, L., and Hoover, L. (2015). Space–time clustering of crime events and neighborhood characteristics in Houston. *Criminal Justice Review*, 40(3):340–360.

Zhao, M., Liu, C., Li, W., and Sharma, A. (2018). Multivariate Poisson-lognormal model for analysis of crashes on urban signalized intersections approach. *Journal of Transportation Safety & Security*, 10(3):251–265.

Zimmerman, D. L., Fang, X., and Mazumdar, S. (2008). Spatial clustering of the failure to geocode and its implications for the detection of disease clustering. *Statistics in Medicine*, 27(21):4254–4266.

Zimmerman, D. L. and Li, J. (2010). The effects of local street network characteristics on the positional accuracy of automated geocoding for geographic health studies. *International Journal of Health Geographics*, 9(1):10.