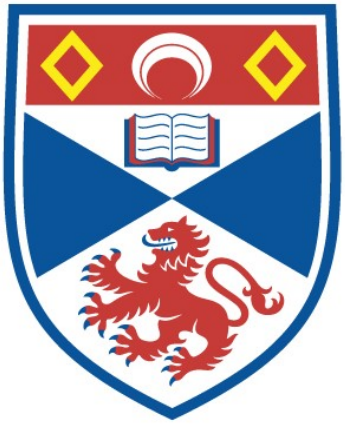
GEOGRAPHICALLY WEIGHTED SPATIAL INTERACTION (GWSI)

Maryam Kordi

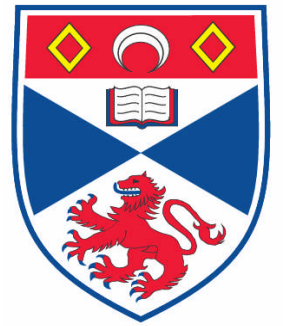
A Thesis Submitted for the Degree of PhD
at the
University of St Andrews



2013

Geographically Weighted Spatial Interaction (GWSI)

Maryam Kordi



A Thesis submitted
for the degree of Doctor of Philosophy
at the
Centre for GeoInformatics (CGI)
School of Geography and Geosciences
University of St Andrews

July 2013

Candidate's declarations:

I, Maryam Kordi, hereby certify that this thesis, which is approximately 42,001 words in length, has been written by me, that it is the record of work carried out by me and that it has not been submitted in any previous application for a higher degree.

I was admitted as a research student in October 2011 and as a candidate for the degree of PhD in October 2011; the higher study for which this is a record was carried out in the University of St Andrews between 2011 and 2013.

Date signature of candidate

Supervisor's declaration:

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date signature of supervisor

Permission for electronic publication:

In submitting this thesis to the University of St Andrews I understand that I am giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. I also understand that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that my thesis will be electronically accessible for personal or research use unless exempt by award of an embargo as requested below, and that the library has the right to migrate my thesis into new electronic forms as required to ensure continued access to the thesis. I have obtained any third-party copyright permissions that may be required in order to allow such access and migration, or have requested the appropriate embargo below.

The following is an agreed request by candidate and supervisor regarding the electronic publication of this thesis: Embargo on both all of printed copy and electronic copy for the same fixed period of two years on the following ground: publication would preclude future publication.

Date signature of candidate signature of supervisor

Abstract

One of the key concerns in spatial analysis and modelling is to study and analyse similarities or dissimilarities between places over geographical space. However, "global" spatial models may fail to identify spatial variations of relationships (spatial heterogeneity) by assuming spatial stationarity of relationships. In many real-life situations spatial variation in relationships possibly exists and the assumption of global stationarity might be highly unrealistic leading to ignorance of a large amount of spatial information. In contrast, local spatial models emphasise differences or dissimilarity over space and focus on identifying spatial variations in relationships. These models allow the parameters of models to vary locally and can provide more useful information on the processes generating the data in different parts of the study area.

In this study, a framework for localising spatial interaction models, based on geographically weighted (GW) techniques, has been developed. This framework can help in detecting, visualising and analysing spatial heterogeneity in spatial interaction systems. In order to apply the GW concept to spatial interaction models, we investigate several approaches differing mainly in the way calibration points (flows) are defined and spatial separation (distance) between flows is calculated. As a result, a series of localised geographically weighted spatial interaction (GWSI) models are developed.

Using custom-built algorithms and computer code, we apply the GWSI models to a journey-to-work dataset in Switzerland for validation and comparison with the related global models. The results of the model calibrations are visualised using a series of conventional and flow maps along with some matrix visualisations. The comparison of the results indicates that in most cases local GWSI models exhibit an improvement over the global models both in providing more useful local information and also in model performance and goodness-of-fit.

Dedication

To my parents, Pedar va Madar e azizam
and to mon cher mari Christian.

I am forever grateful for your unconditional love and support.

Acknowledgements

First, I would like to express my sincere thanks to everyone who contributed in many ways to the success of this study, both at work and in my private life.

I gratefully thank my thesis supervisor, Professor A. Stewart Fotheringham, for his support and guidance throughout my PhD research. I am very much thankful to him for providing me this opportunity to pursue my doctorate under his supervision and also for providing necessary resources to accomplish my research work. I also gratefully thank him for providing me with the freedom to explore and try new ideas which made my PhD study an unforgettable experience. I can never thank Professor Fotheringham enough for his help and support with my transferring to St Andrews and during my residence in Cellardyke.

I would also like to extend my appreciation to my examining committee members: Prof. Graham Clarke from School of Geography, University of Leeds and Dr. Urška Demšar from Centre for GeoInformatics (CGI), University of St Andrews.

I also wish to thank everyone in the National Centre for Geocomputation (NCG) at the National University of Ireland Maynooth (NUIM) where I started my PhD study. My special thanks go to Martin Charlton, Dr. Alexei Pozdnoukhov and Dr. Carson Farmer who were always available for answering my academic questions and to Melina and Ann-Marie for their friendship and help during my study in NCG.

Finally I would like to express my gratitude to the faculty and staff in the School of Geography & Geosciences at University of St Andrews and also gratefully acknowledge my friends and colleagues in CGI. I truly appreciate Sila's and Tommy's offer for staying in their place whenever I traveled to St Andrews in the last few months of my PhD study.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Spatial interaction is broadly defined as the movement, flow, or communication of people, goods or information over space resulting from a decision-making process (Haynes and Fotheringham, 1984; Fotheringham and O'Kelly, 1989; Fotheringham et al., 2000; Fotheringham, 2001; Fischer, 2000). Examples include a wide variety of behaviours such as migration, shopping patterns, commuting, commodity or communication flows, telephone calls, airline passenger traffic, attendance at events such as theatre, conferences and sport events (Haynes and Fotheringham, 1984), all of which form important components of social and urban complex systems. Researchers in a variety of fields have modelled spatial movements through mathematical equations known as spatial interaction models (Fotheringham et al., 2000). These models are particularly useful for better understanding and analysing the patterns of and the underlying structure of the spatial flows in the interaction systems. One of the early spatial interaction models was the gravity model and its related family of models (Wilson, 1967, 1970; Haynes and Fotheringham, 1984; Fotheringham and O'Kelly, 1989; Sen and Smith, 1995; Fotheringham et al., 2000; Roy and Thill, 2004). Later, the underlying formulations of spatial interaction models have been modified further and more sophisticated models have been developed such as competing destinations models (see Fotheringham, 1983, 1984b, 1986).

Spatial interaction is fundamental in regional science (Fischer and Getis, 1999; Clarke and Clarke, 2001) and is also an important aspect of modern society and economy. As a consequence, spatial interaction modelling is one of the most applied geographical analysis and modelling techniques (Fotheringham et al., 2000) (see for instance applications and references in Fotheringham and O'Kelly, 1989; Haynes and Fotheringham, 1984). Traditionally, spatial interaction models have been calibrated globally in which one set of parameter estimates is provided for a study region (Fotheringham and Brunsdon, 1999). The resulting global parameter estimates represent an average type of interaction behaviour and are assumed to be equally valid across the entire study area. The global validity of the results is due to the assumption of spatial stationarity in relationships

being investigated (Lloyd, 2011). However, in many real-life situations relationships may vary across space and then important variations in interaction behaviour could be completely hidden (Linneman, 1966; Greenwood and Sweetland, 1972; Fotheringham et al., 2000, 2002) because the global results may fail to represent the true specification of the reality (Fotheringham and Brunsdon, 1999; Fotheringham et al., 2002; Unwin, 1996a,b; Fotheringham, 1997; Boots and Okabe, 2007).

The global model misspecification came to light through local parameter estimates being obtained for each separate origin or destination region (Fotheringham et al., 2000, 2002) (see sections 5.4.1 and 5.4.2 for further information). The origin- and destination-specific models provide a set of parameter estimates for each origin or destination in the system (see for instance Haynes and Fotheringham, 1984; Fotheringham and O'Kelly, 1989). Although the origin- and destination-specific models provide more disaggregated information compared to global interaction models, these models are localised at the level of discrete origins/destinations. An important drawback of these models is the fact that they ignore a substantial amount of data that can be potentially useful for calibration. For example, an origin-specific model ignores flows from surrounding origins that might have similar flows to the destinations, leading to a lower number of considered data points and potentially to a less reliable parameter estimation. Another possible problem with origin- and destination-specific models is the fact that these models might ignore significant geographical variations of parameters in the interaction system. For example, an origin-specific model provides a single set of parameters for a given origin, ignoring potential differences across destinations. Therefore, identifying spatial variations in relationships, (sometimes referred to as spatial heterogeneity, spatial non-stationarity or spatial drift (Charlton et al., 1997)), is still an ongoing problem in spatial interaction modelling requiring further study. This leads to the following research questions:

- *How can spatial heterogeneity be detected and taken into account in spatial interaction processes?*
- *How can spatial interaction models be localised to consider spatial heterogeneity?*

Interest in local forms of spatial analysis and modelling is not new (Fotheringham, 2000); over the last few decades there has been a powerful movement from global modelling to local modelling (Fotheringham and Brunsdon, 1999; Fotheringham, 1999b; Openshaw et al., 1987; Getis and Ord, 1992; Anselin and Getis, 1992; Fotheringham and Rogerson, 1993; Lloyd, 2011, 2006) which focuses on identifying and understanding differences across space rather than similarities (Fotheringham, 1999b; Fotheringham et al., 2000). Several types of local analytical techniques for spatial data have been developed and used in the literature (see section 5 for a review). One local technique that has become increasingly popular in detecting spatial non-stationarity in spatial analysis is Geographical Weighted Regression (GWR) (Brunsdon et al., 1996; Fotheringham et al., 1998; Brunsdon et al., 1998a, 1999a; Fotheringham et al., 2002) which will be introduced in detail

in section 5.3. Within the GWR framework, relationships under study are allowed to vary spatially and a set of local parameter estimates is produced for each regression location and all observations are spatially weighted with respect to the regression point. The GWR technique has been used in a wide range of applications for spatial data (see section 5.4.3) and found to be efficient in detecting spatial heterogeneity in relationships that may be missed in a global regression analysis (Foody, 2004). The local parameter estimates derived from a GWR analysis can be mapped to show how a relationship varies over space and then to investigate the spatial pattern of the local estimates for better understanding of possible causes of this pattern (Fotheringham et al., 2002).

Considering that GWR could successfully contribute to modelling "*spatially heterogeneous processes*" (Brunsdon et al., 1996; Fotheringham et al., 1996, 1997b, 2002) and has efficiently worked in a range of studies for spatial data, questions we are therefore addressing are:

- *How it is possible to use the experience from GWR in order to apply the geographical weighting concept to spatial interaction models?*
- *Within a geographically weighting framework for spatial interaction, how do we define distances between spatial flows?*"
- *How do we visualise the local parameter estimates for spatial interaction models?*

On a practical side, given the availability of GWR software, the following question is of interest:

- *Will existing GWR software work for spatial interaction flows or is a specific adaptation to spatial interaction required?*

Given the importance of intra-zonal flows in many spatial interaction processes, and the local nature of assessing spatial heterogeneity, we also include the following important question into this research:

- *How can intra-zonal trip distance be estimated in spatial interaction systems in which intra-zonal flows are taken into account in the analysis?*

1.2 Aim and objectives of the thesis

The aim of this research is to develop a framework for localising spatial interaction models using the geographically weighted concept (known from GWR) in which spatial heterogeneity can be detected, visualised and analysed in spatial interaction systems.

The following research objectives represent the required steps for achieving this aim:

- Designate and specify a real-world spatial interaction problem for the case study (data collection and processing).

- Explore and investigate different existing spatial interaction models.
- Investigate possible ways to incorporate intra-zonal flows in the spatial interaction models along with exploring and analysing existing approaches.
- Study and analyse the spatial interaction patterns of the case study using both global and existing local techniques.
- Explore possible approaches for applying the geographically weighted concept to spatial interaction models.
- Create an algorithm and, if necessary, write computer code for calibrating geographically weighted spatial interaction (GWSI) models.
- Perform verification and validation of the GWSI models by applying them to the case study for detecting and analysing spatial heterogeneity.
- Develop and apply appropriate techniques for the visualisation of the conventional and local interaction model results.
- Analyse and compare the spatial patterns of the parameter estimates resulting from global and various local models.
- Demonstrate some application examples of the GWSI models especially for the case of forecasting spatial interaction patterns.

1.3 Short review of techniques and contributions of the thesis

The idea of applying the GWR approach to spatial flows was pointed out by Berglund and Karlström (1999) where the potential applicability of this approach was suggested without performing the method. Later Nakaya (2001, 2003, repeated publications) investigated this by applying GWR to the calibration of an origin-specific model of a migration case study in Japan. Nakaya's study is interesting from two points of view. First, this study has shown that GWR can be a promising approach for local calibration of spatial interaction models or at least for the origin-specific model in the case of migration in Japan. Nissi and Sarra (2011) have also applied this approach similarly for an origin-specific model of migration flows in Poland. Second, in this approach of using GWR for calibrating an origin-specific model for each origin in the study region, a surface of parameter estimates could be estimated where each value in this surface describes the relationship being measured around the destinations close to that location (see Fotheringham et al., 2002, page 244). In Nakaya's study, however, only an origin-specific model is calibrated where the restriction of the model makes the application of GWR on flows easier since only destinations are involved in the geographically weighting process.

Therefore, the question of "how to apply the geographically weighted concept on spatial interaction models" still remains unsolved.

In order to fulfil the main aim of this study which is to develop a framework for localising spatial interaction models using the geographically weighted concept, we have investigated several approaches differing mainly in the way calibration points (flows) are defined and spatial separations (distance) between flows are estimated. As result, a family of geographically weighted spatial interaction (GWSI) models is developed throughout this thesis allowing for the detection of spatial variations in interaction behaviour. The family of GWSI models is composed of the following local models:

- Origin-focused spatial interaction model
- Destination-focused spatial interaction model
- Destination-specific origin-focused spatial interaction model
- Origin-specific destination-focused spatial interaction model
- Flow-focused spatial interaction model

In all these models, following the principle of the geographically weighing concept, around each calibration point a spatial kernel is considered and observations are weighted according to their proximity to the calibration point. However, the main difference between the models is the way the calibration point and the distances between flows are defined. In the first four models, the resulting local parameter estimates are associated with the spatial flows, although the calibration points in the geographically weighting approach are actual locations within the study region; i.e. origin locations in the case of origin-focused models and destination-specific origin-focused models, and destination locations in the case of destination-focused models and origin-specific destination-focused models. The distance between flows is defined by the distance between the calibration point and origins or destinations of the observed flows in the origin-focused and destination-focused models respectfully.

The flow-focused model is different from others because the calibration points in the geographically weighting approach are spatial flows. Therefore, in a flow-focused model a spatial kernel is considered around each calibration flow and the observed flows within this kernel are weighted. Two feasible ways for estimating distances between flows are suggested in this thesis; one based on a four-dimensional spatial kernel and one based on a spatial trajectory distance measure.

All the above models along with a global gravity model and an origin-specific and a destination-specific models are calibrated using a journey-to-work case study in Lausanne, Switzerland. The results of the models are visualised using conventional maps, matrix visualisations and a series of flow maps. The results of the local models are compared with the related global models and in most cases the local GWSI models exhibit an

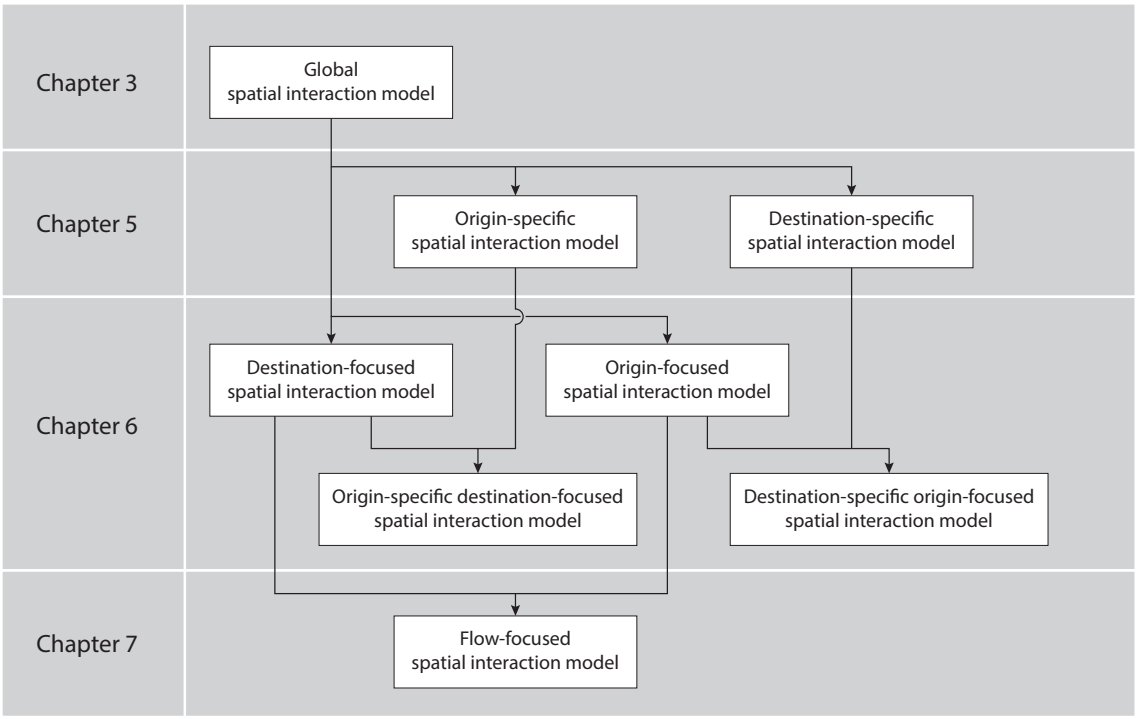


Figure 1.1: Overview of the spatial interaction models discussed in the thesis, with their dependencies.

improvement over the global models both in providing more useful local information and also in terms of model performance and goodness-of-fit.

Figure 1.1 gives a visual overview of the different spatial interaction models discussed in this thesis and relates them to each other. An additional contribution of this work is the discussion on how intra-zonal flows can be considered in spatial interaction models by introducing a method for intra-zonal distance measures. Internal flows are important in local spatial interaction models, because these local flows will be given more weight in a geographically weighting procedure since they are always closer to the calibration location compared to other observed flows. Also, the Lausanne journey-to-work dataset contains an important proportion of internal flows (about 45%), which is an additional reason to include these flows in the models.

1.4 Outline of the thesis

This thesis is structured in 9 chapters, as follows:

- In chapter 2 we introduce the dataset of journey-to-work in the Lausanne agglomeration in Switzerland which is used throughout the work in order to test and validate the interaction models.

- Chapter 3 provides an overview of spatial interaction models and their underlying

theoretical framework along with a discussion on calibration techniques for the interaction models. Also a global Poisson gravity model is calibrated using the journey-to-work dataset in Lausanne described in chapter 2.

- In chapter 4 we discuss the intra-zonal flows problem and introduce a methodology for estimating the average intra-zonal trip length allowing for integration of intra-zonal flows in spatial interaction models.

- Chapter 5 gives a brief overview of the existing local methods for spatial data analysis with attention given to the models which are used in this thesis such as Geographically Weighted Regression (GWR) and origin- and destination-specific spatial interaction models.

- Chapter 6 combines the geographically weighted concept with spatial interaction models and introduces four members of a family of local GWSI models: origin-focused, destination-focused, destination-specific origin-focused, and origin-specific destination-focused models and discusses their application to the Lausanne journey-to-work dataset.

- Chapter 7 introduces the last version of the GWSI, flow-focused model and discusses distance measures between flows along with bandwidth calibration and spatial kernel issue. This model is also applied to the Lausanne journey-to-work dataset and the results are briefly analysed.

- Chapter 8 discusses some further issues around local spatial interaction and shows some example applications where GWSI models are used for prediction.

- Chapter 9 concludes the thesis by giving a short overview of the models and the contributions of the individual chapters and finally with a discussion on the possible future work.

Chapter 2

Data and case study

2.1 Introduction

This thesis focuses on the methodologies of localising spatial interaction models. In order to illustrate and examine local interaction methods we need to use an appropriate spatial dataset. Spatial interaction, by definition, takes place between a pair of locations in space (i.e. origin and destination points) and data should contain information on the volume of flows between these points as well as attributes and locational information about the origins and destinations (see Thompson, 1974; Bailey and Gatrell, 1995; Banerjee et al., 2000; Rae, 2009). Here in this study, the data are used only for model validation and exposition so the ideas herein are not limited to any particular type of spatial interaction data but instead have broad application.

The Swiss Federal Statistical Office provides fine scale data on journey-to-work (commuting) between different communes in the country. Communes, also known as municipalities, are the smallest administrative district in Switzerland. In order to provide more detailed information about the dataset, we first describe some general definitions and information about the subdivisions of Switzerland.

2.2 Subdivisions of Switzerland

There are several administrative divisions in Switzerland that divide the country into smaller units. The highest administrative subdivision in the country are known as cantons. There are 26 cantons in Switzerland which are the member states of the Swiss Confederation. Each canton is divided in a number of districts and each district is divided in communes which is the smallest administrative unit. Switzerland had 2896 communes in 2000. Communes have a local government and are responsible for basic public services. Communes vary in the size from 22 residents for Corippo (Ticino) to 363,273 residents for the city of Zurich (for year 2000, data from Population Census 2000).

Besides the administrative levels of cantons, districts and communes, Switzerland is

divided in a series of other spatial subdivisions based on several statistical variables. One of these subdivisions is the concept of agglomeration, corresponding roughly to what is a Metropolitan Area in the US (Berry et al., 1969; Dahmann and Fitzsimmons, 1995). Other countries have very similar concepts, for instance the Urban Areas in the UK.

2.2.1 Agglomerations in Switzerland

The agglomerations try to define the spatial extent of urban areas. According to Schuler et al. (2005, published by Federal Swiss Statistical Office), the definition of agglomerations is based on different characteristics. More specifically, in Switzerland, a commune belongs to an agglomeration if at least 3 of the 5 following criteria are met:

- Continuity of built zone with the central city
- High human density (sum of residential population and number of jobs)
- Population growth higher than average
- Low agricultural activity
- Strong commuting relationships with the central zone of the agglomeration

Figure 2.1 shows the Swiss agglomerations according to the definition of the year 2000. They contain a central city (in red) and surrounding functionally and economically dependent areas (in orange). In some particular cases, an agglomeration can also consist of a single isolated city (in yellow). The main purpose of the definition of agglomerations is to be able to compare urban areas with very different administrative limits. All agglomerations together define the urban area of Switzerland, as opposed to rural zones. With progressing urbanisation, the definition of agglomeration has changed over time and is periodically updated by the Swiss Federal Statistical Office (usually every 10 years). In some cases, an agglomeration can also contain neighbouring areas abroad where the functional and economic relationships are strong. A total of 979 of 2896 communes were considered as being 'urban' in 2000, with a total of 73% of the population (Schuler et al., 2005).

2.3 The Swiss journey-to-work (commuting) dataset in the literature

As the quality of the Swiss journey-to-work dataset is quite high with fine scale information about Switzerland, it is not surprising that the data have been used in previous studies. For instance, Bavaud (2010a) used the Swiss commuting dataset for illustrating the procedure of his study of distances on weighted graphs enabling thermodynamic graph clustering. The dataset has also been used by Tuia and Bavaud (2007) as a case study for a dimensionality reduction algorithm based on a flow matrix. The commuting

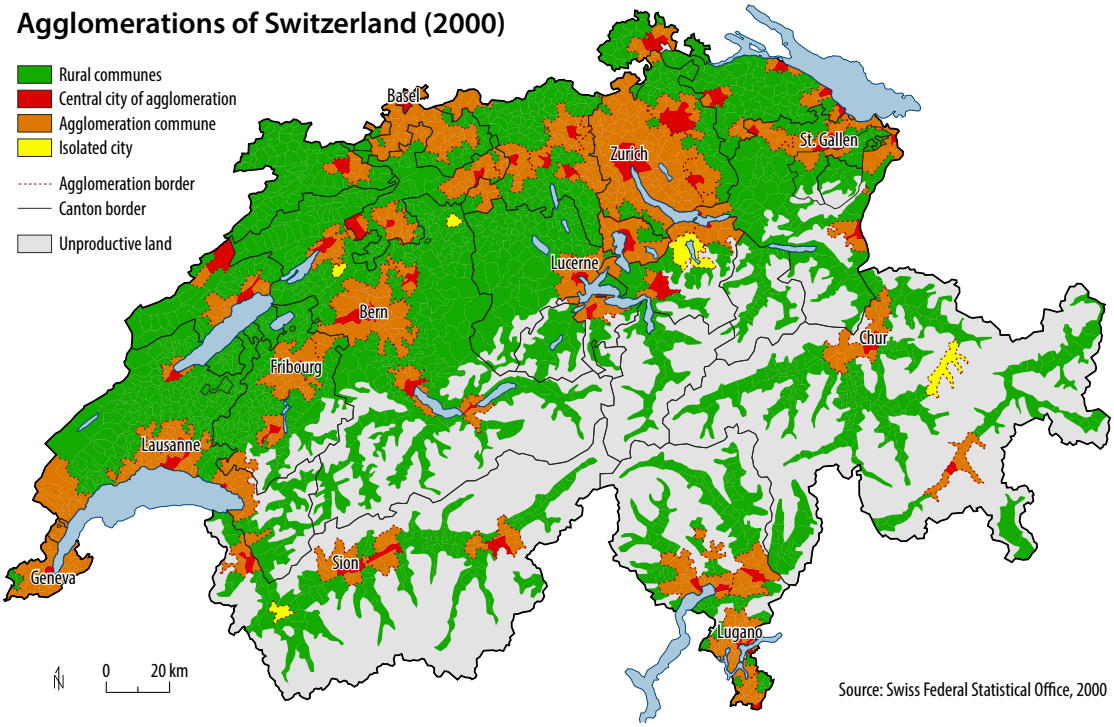


Figure 2.1: Agglomerations of Switzerland for year 2000

dataset also contains information about the means of transportation used for the journey to work. Kanevski et al. (2009) have used these data to illustrate the ability of General Regression Neural Networks (GRNN) to predict the spatial pattern of the usage of different means of transportation in commuting. The data has also been used for visualisation purposes in Killer and Axhausen (2010) or Kaiser (2011). Kaiser et al. (2011) have used the commuting dataset to demonstrate the calibration of a local spatial interaction model using a variant of geographically weighted regression. Dessemontet et al. (2010) have made an extensive study of the commuting network of Switzerland by using the same dataset and Dessemontet (2011) has used these journey-to-work data along with other data to study the evolution of employment and accessibility in Switzerland over 60 years.

2.4 A spatial interaction model for journey-to-work

Due to the importance of commuting and trip distribution for spatial planning, (e.g. traffic and infrastructure development), there has been intensive research covering this subject in geography and regional science (de Vries et al., 2009; O'Kelly and Niedzielski, 2007) (also see papers Wilson, 1967; O'Kelly and Lee, 2005; Farmer and Fotheringham, 2011; Batty, 1976; O'Kelly and Niedzielski, 2008; Wilson, 1974; O'Kelly et al., 2012; Sang et al., 2011, inter alia). According to de Vries et al. (2009), the concepts of labour and

housing markets are connected through commuting flows and so the size of flows between regions and their effect on the housing and labour markets can be analysed with spatial interaction models (see Batten and Boyce, 1986; Fotheringham and O'Kelly, 1989).

A comprehensive introduction to spatial interaction modelling is given in section 3. However, briefly there are three essential elements in spatial interaction models. The first is travel cost which often is measured as distance between interaction (origin and destination) regions; the second and third elements are attributes (or sizes) of interaction regions which measure propulsiveness of origins and attractiveness of destinations respectively. Based on the type of interaction problem and the purpose of the model, different origins and destinations attributes can be considered in the model. For instance in a shopping expenditure model, the origin attribute might be defined as the average household income or unemployment rate whereas in a migration model, living cost or average house price can be considered as destination attributes (see Fotheringham and O'Kelly, 1989).

In this thesis, according to the available elements in our dataset, we consider four components in our journey-to-work model. The first component is that of commuting flows which will act as the independent variable in the calibration of the spatial interaction model. The other three components are:

- Number of economically active population (working people) in each commune, considered as an origin propulsiveness attribute
- Number of jobs in each destination region, considered as a destination attractiveness attribute
- Euclidean distance between centroids of origin and destination communes, considered as a surrogate for travel cost

Although the selection of variables in the interaction case study in this work is mainly guided by available census elements, there are a number of studies in the literature using the same variables for commuting analysis, supporting these choices (see for instance Uboe, 2004; Lloyd and Shuttleworth, 2005; Shuttleworth and Lloyd, 2005; O'Kelly and Niedzielski, 2007, 2008; Lloyd et al., 2007).

In this thesis we use the Swiss commuting dataset for the year 2000 which was the latest available version of the data at the time of analysis. This dataset is issued from the population census, which contains also other data including population. The number of jobs is available through the firms census which has been conducted during 2001. The qualifying date for the population census 2000 is the 5 December, while the firms census contains data for the 1 January 2001, less than one month later. The following sentences give a summary of general information about the data used in this thesis for spatial interaction modelling:

- The journey-to-work data were acquired during the population census 2000. Frick et al. (2004) describe extensively this dataset including methodological issues on

data acquisition and descriptive statistics. The Swiss Federal Statistical Office offers this commuting data at the fine communal level, freely available at `http://www.pendlerstatistik.admin.ch`

- Information about active population has been acquired during the population census 2000, along with other information such as residential population. Most of the data from the population census is freely available at `http://www.stattab.bfs.admin.ch` at the level of the communes, including the active population we are using for our case study.

- The number of jobs has been acquired by the Statistical Office during the firms census in 2001, where information about all companies in Switzerland was collected. Information about the economic sectors is also available, including the split of the number of jobs between secondary and tertiary sectors. Again, most of the data from this census is also freely available at `http://www.stattab.bfs.admin.ch` at the level of the communes.

2.5 Case study: Lausanne

The agglomeration of Lausanne, located in Western Switzerland was selected as the study area for the calibration of various spatial interaction models. This agglomeration is composed of 70 communes, covering a total area of roughly 317 km^2 with a population of slightly more than 310,000. The agglomeration is well separated from neighbouring agglomerations which limits undesired inter-agglomeration interactions. Also the agglomeration of Lausanne does not extend behind the country borders as it is the case in some areas in Switzerland; e.g. in the agglomeration of Geneva where over 9% of the workforce lives in France, or in Basel where over 13% of the workforce lives in France or Germany. In Lausanne only 1% of commuters are cross-border. Consisting of 70 communes with populations ranging from 61 inhabitants (commune of Malapalud) up to 124,914 for the city of Lausanne, (overall population about 310,000 in 2000), makes the agglomeration of Lausanne the 5th biggest agglomeration in Switzerland. Figure 2.2 shows an overview of the agglomeration of Lausanne with the road and train network. Lausanne is bordered in the south by Lake Geneva. The main transportation routes are along the lake and also in a northerly direction.

Figure 2.3 shows the location of the 70 communes in the agglomeration of Lausanne in 2000. Figure 2.4 shows the residential population in the agglomeration. The city of Lausanne is by far the most populated commune with nearly 125,000 people which accounts for roughly 40% of the population of the agglomeration. The remainder of the population is mainly concentrated in the communes around Lausanne, mainly to the West between the communes of Prilly to Morges. To the North, we find mainly small communes, with the exception of the slightly more important town of Echallens. Figure 2.5 shows the active population in the agglomeration. As active population is considered the working

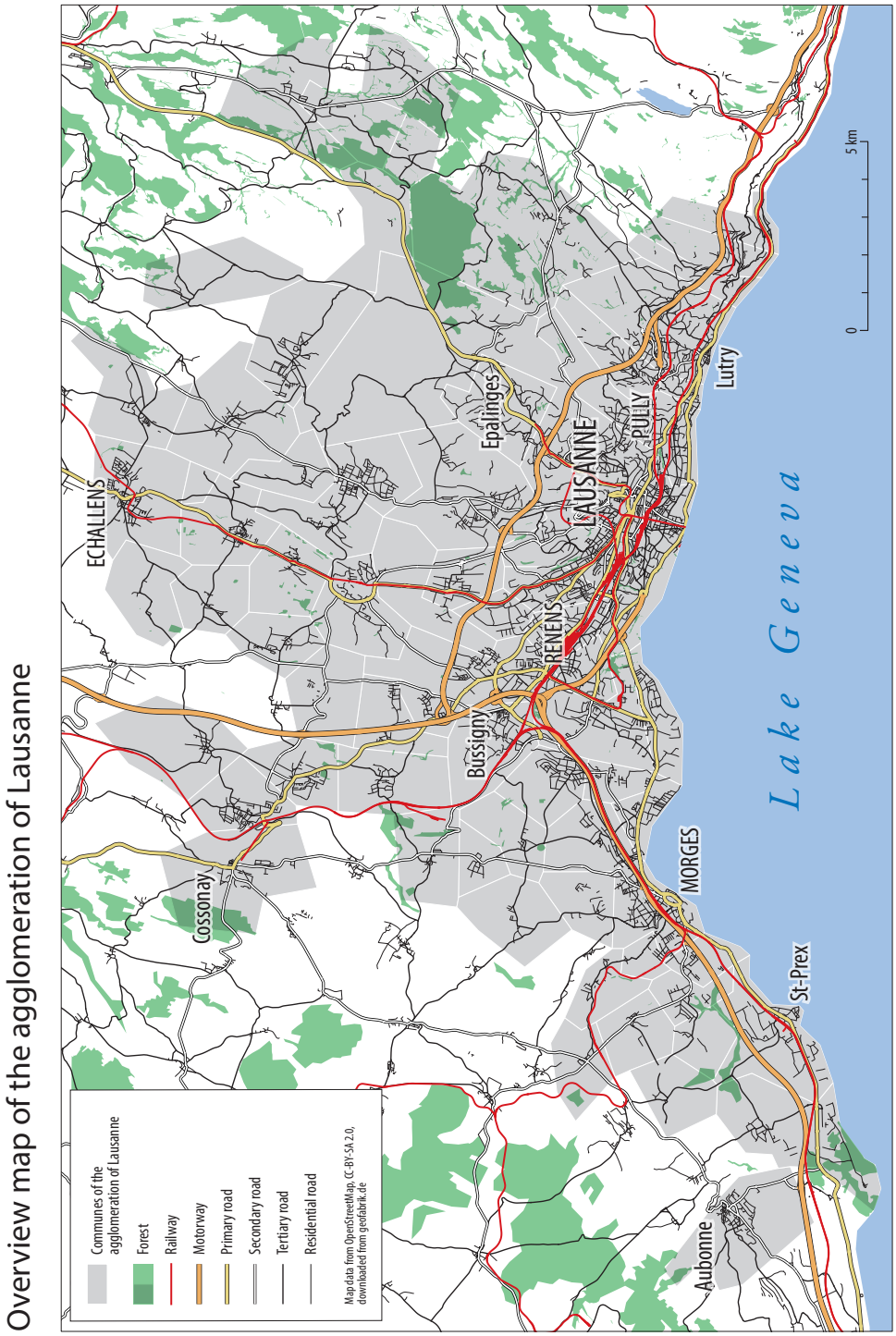


Figure 2.2: Overview map of the agglomeration of Lausanne

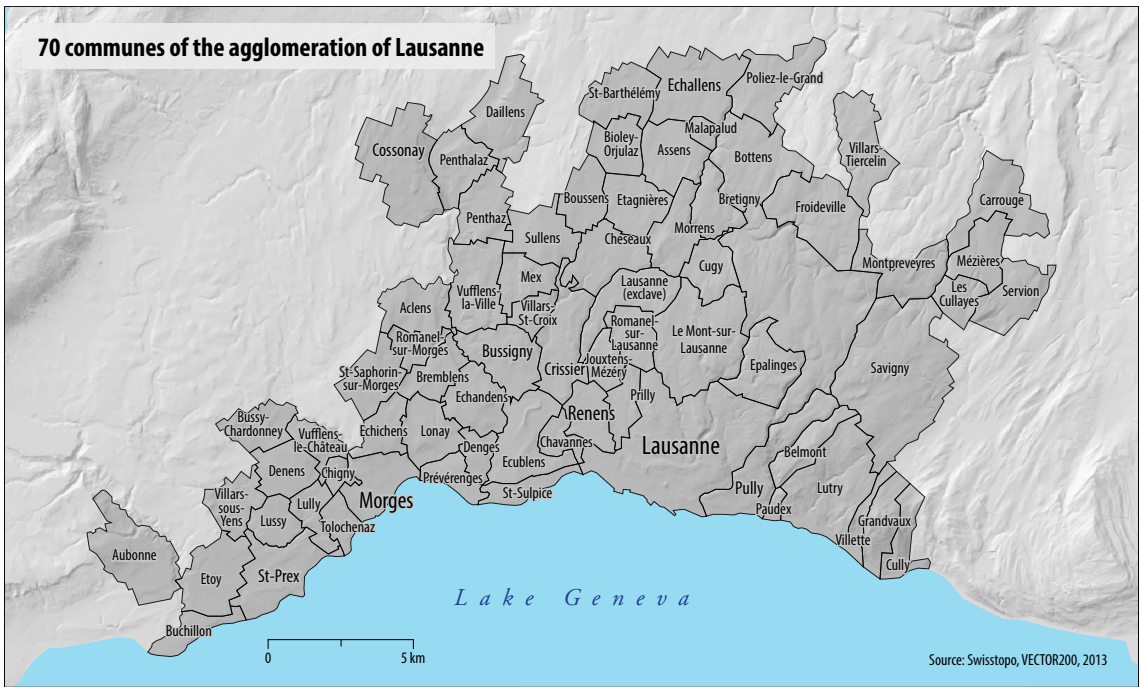


Figure 2.3: Communes of the agglomeration of Lausanne

population together with the people seeking actively employment, students are not considered in the active population unless they are working part time. The spatial structure of the active population is similar to the residential population; the city of Lausanne has an active population of nearly 60,000 which is roughly 38% of the overall active population in the agglomeration. The ratio of active population to residential population is of 48% in the city of Lausanne, and 50% in the whole agglomeration.

Figure 2.6 shows the number of jobs in the agglomeration in 2001. A job is considered as an occupied working place in a company so vacant jobs are not counted in this statistic. There is also no difference between part time and full time jobs; both are counted as a job. Nearly half of the jobs (roughly 86,000 of 175,500, or 48.9%) are located in the city of Lausanne, which shows its important role as a centre of this agglomeration. The remaining jobs are mainly located in the west and north-west of the city, not far from the junction of the motorways going east, west and north. Figure 2.7 compares the number of jobs with the active population. Blue circles depict a surplus of jobs, while red circles represent communes with a greater number of population than jobs. This map shows a clear gap between the economic and the residential communes in the agglomeration. The communes east of Lausanne are typical residential communes with a wealthier population.

Figures 2.8 and 2.9 both show the percentage of internal flows for each commune. In figure 2.8, the number of internal flows is compared to the *total incoming flows* to the commune (including the internal flow), while in figure 2.9, the percentage of internal

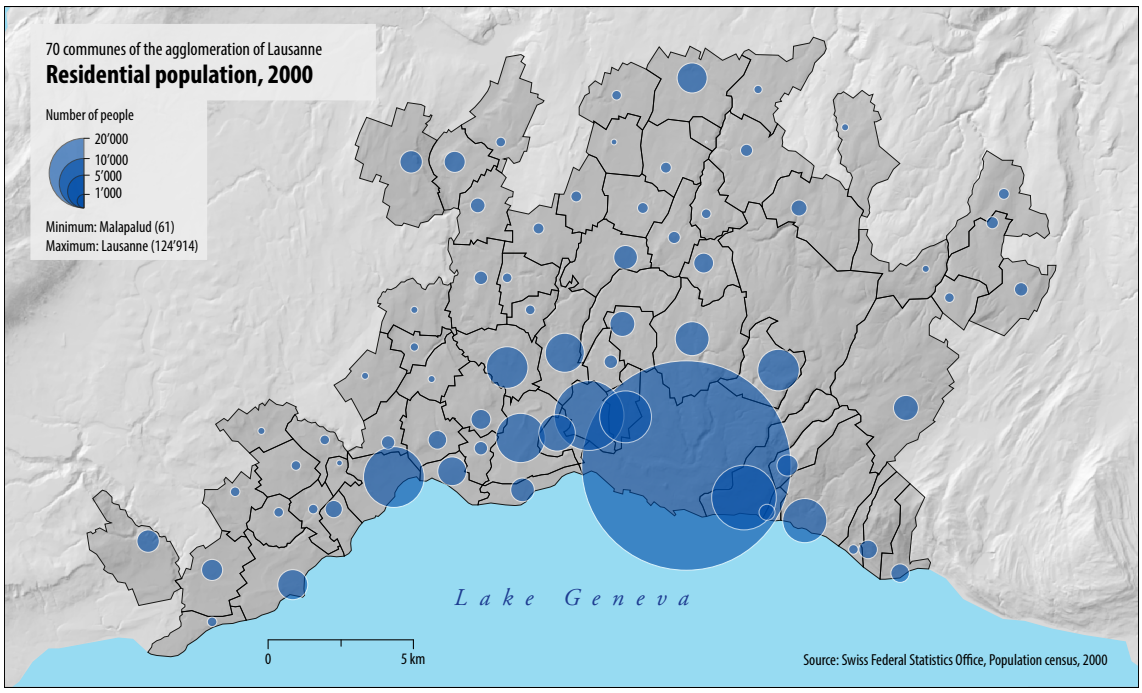


Figure 2.4: Population of the agglomeration of Lausanne in 2000

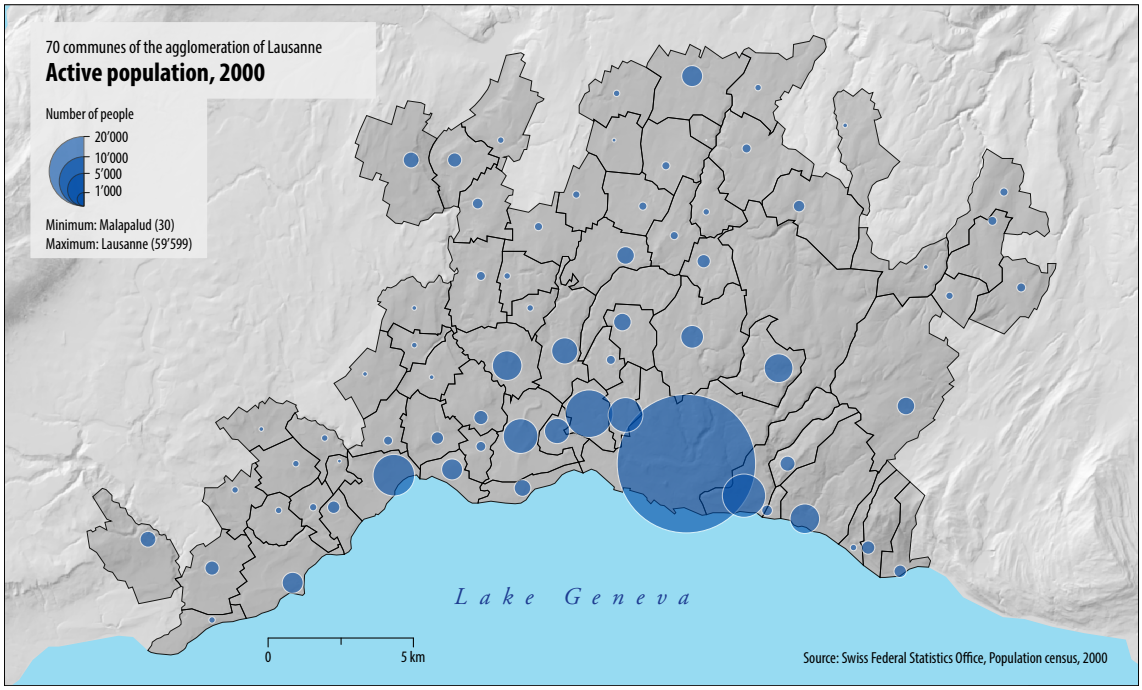


Figure 2.5: Active population of the agglomeration of Lausanne in 2000

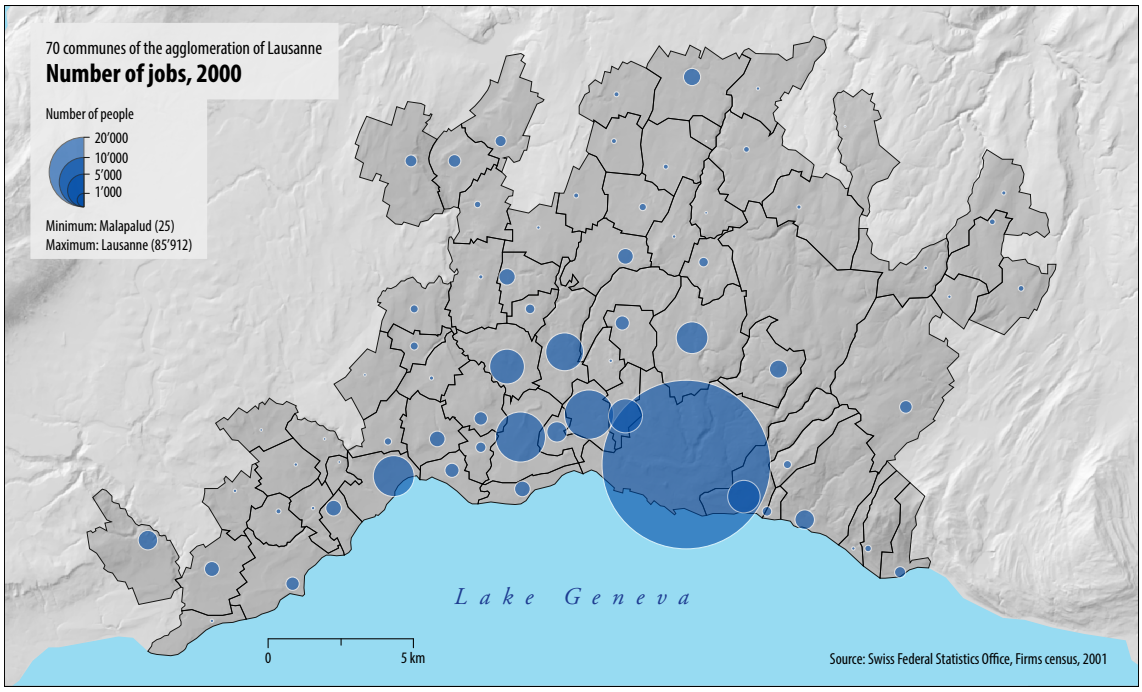


Figure 2.6: Number of jobs in the agglomeration of Lausanne in 2001

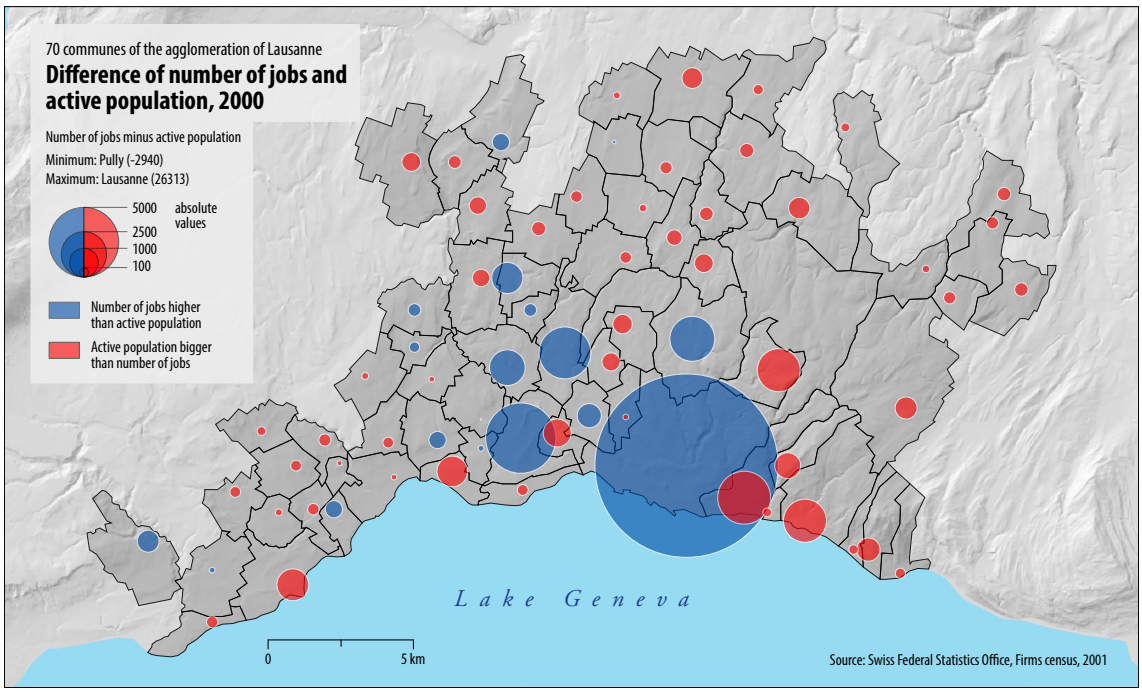


Figure 2.7: Number of jobs (2001) minus active population (2000) in the agglomeration of Lausanne. Proportional symbols depict the absolute value of the difference.

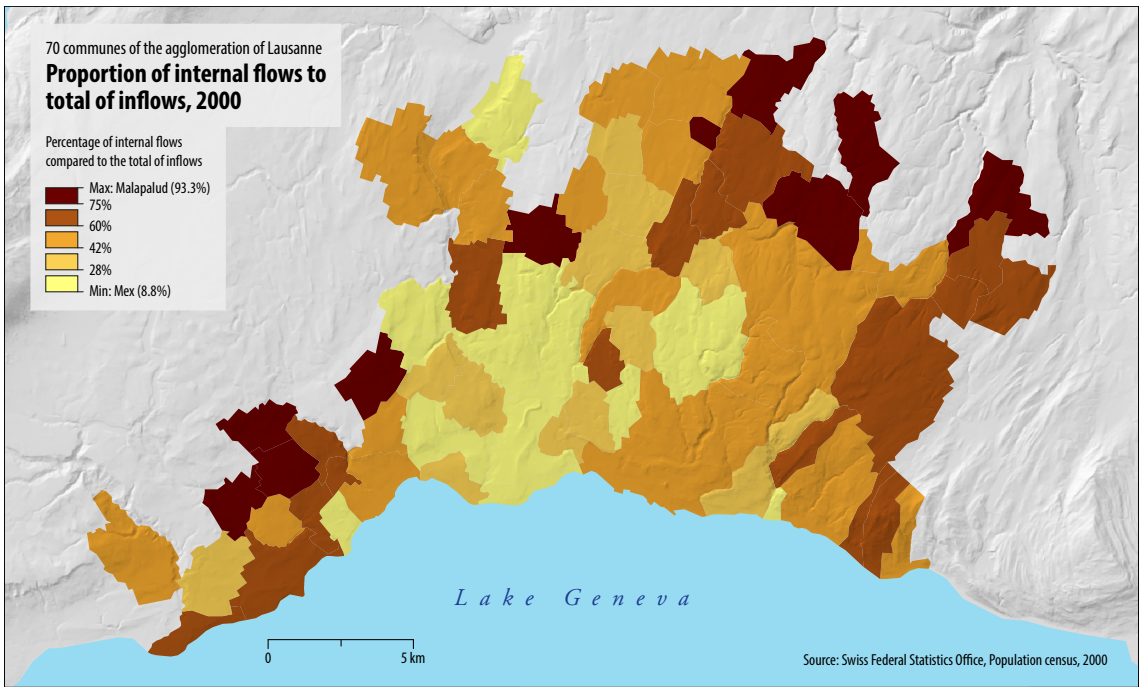


Figure 2.8: Percentage of internal flows compared to the total of inflows for the communes of the agglomeration of Lausanne

flows is computed compared to the *total outgoing flows*. Figure 2.8 gives information on the percentage of the commuters, working in a commune living in this same commune. As an example, from the 64,717 people commuting to Lausanne, 35,585 or 55% live in Lausanne itself. The spatial pattern shows clearly that the smaller communes at the border of the agglomeration have a high percentage of internal flows compared to the total of incoming flows, suggesting few people commute to these communes. On the other hand, the communes west of Lausanne have more jobs than active population (see also figure 2.7) and typically have a high percentage of people commuting from other communes. The information in figure 2.9 represents the percentage of workforce staying inside their commune of residence for their work. As an example, from the 47,071 commuters of the city of Lausanne, 35,585 or 75.6% do not leave the city for employment. This percentage of internal flows compared to the total of outflows is higher than average in the communes having more jobs than active population. But there are a series of communes in the western border of the agglomeration towards Geneva showing even higher proportions of internal flows to total of outflows. These communes seem to offer a high proportion of jobs to the local population while having, with the exception of Aubonne, a surplus of active population compared to the number of jobs, which is rather unusual in the agglomeration of Lausanne.

Figure 2.10 represents the journey-to-work flows in the agglomeration of Lausanne. Flows smaller than 20 commuters are not considered. The internal (intra-zonal) flows are

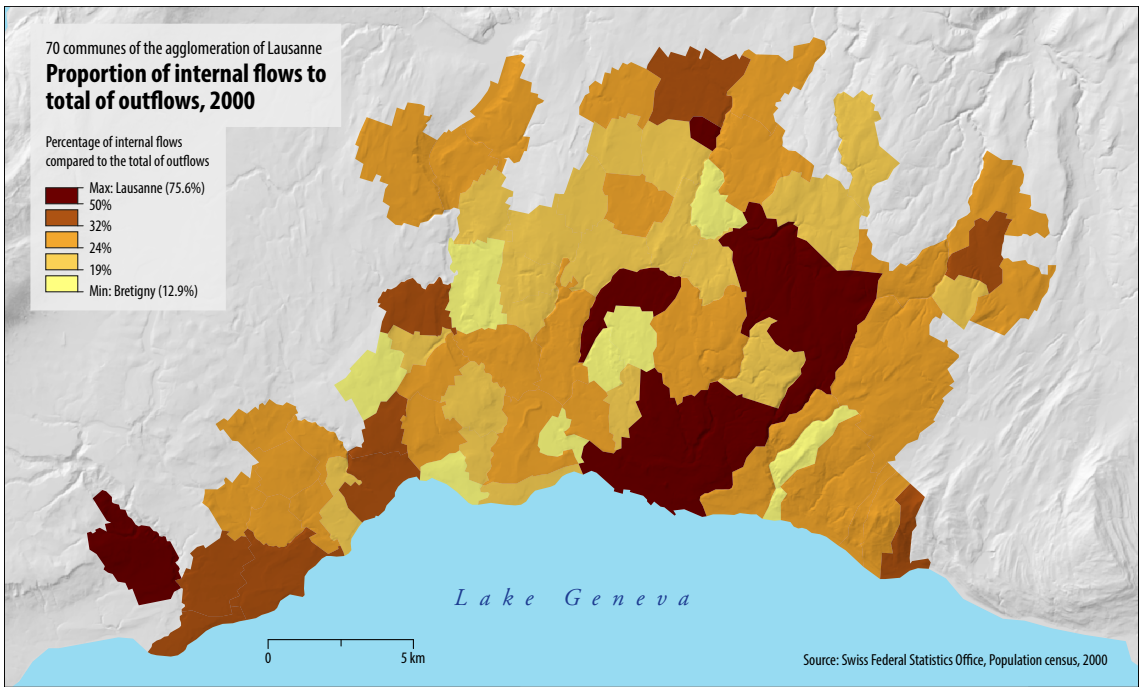


Figure 2.9: Percentage of internal flows compared to the total of outflows for the communes of the agglomeration of Lausanne

represented by proportional circles while the inter-zonal flows are depicted by lines with proportional width. The map shows clearly an inner part of the agglomeration which is very well connected in terms of journey-to-work flows. This inner part runs roughly from Morges in the West to Lutry in the East and contains the communes North-West of Lausanne having more jobs than active population (see figure 2.7). The communes in the inner part of the agglomeration are well connected between themselves, while the flows in the surrounding communes focus mainly towards the inner part of the agglomeration. This pattern shows a concentric organisation of the agglomeration where the central part has the biggest parts of the jobs, and the surrounding communes are mostly residential.

Figure 2.11 shows the relationship between the total incoming versus the total outgoing flows. The diagonal line represents equality of incoming and outcoming flows. Both axes of the chart have logarithmic scales. Only a few communes have higher inflows than outflows. This chart shows that a few central communes have more jobs than active population, and many small, less central communes are typical residential communes with much more outflows than inflows.

The frequency histogram of the distance from home to work is shown in figure 2.12. The average commuting distance is around 3.3 km, while the median is 2.2 km due to the skewed nature of the frequency distribution. It has to be noted that the commuting distances have been computed using the polygon centroids, resulting in internal flows having a distance of 0. Given that about 45% of the flows are internal flows and that

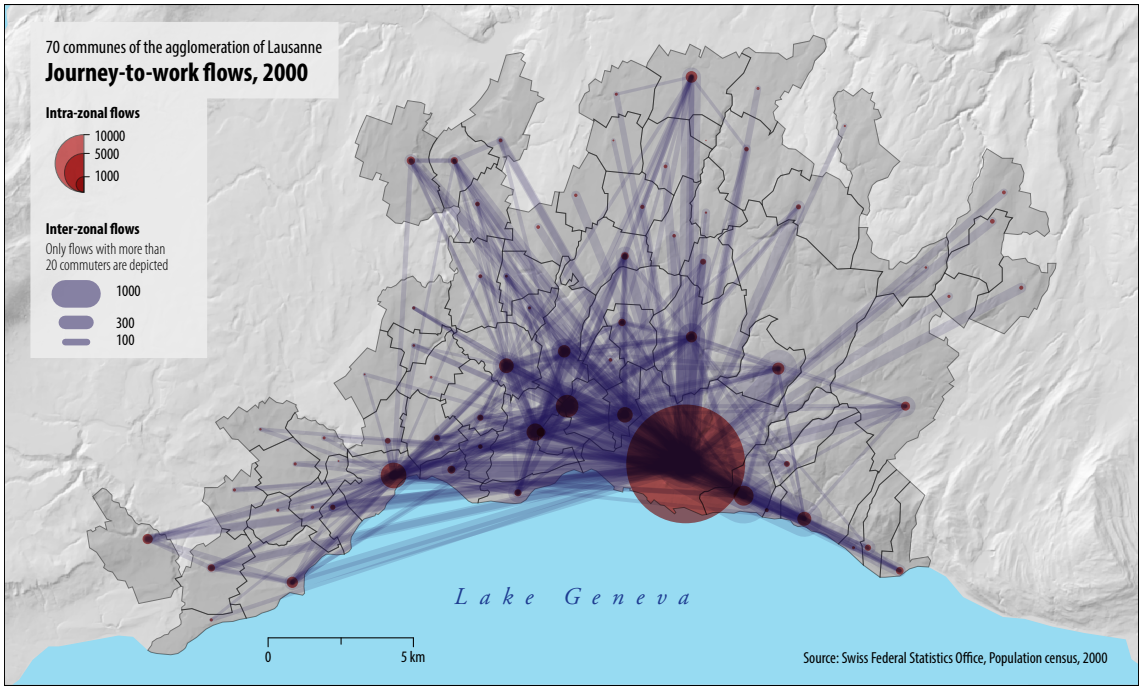


Figure 2.10: Journey-to-work flows in the agglomeration of Lausanne, 2000

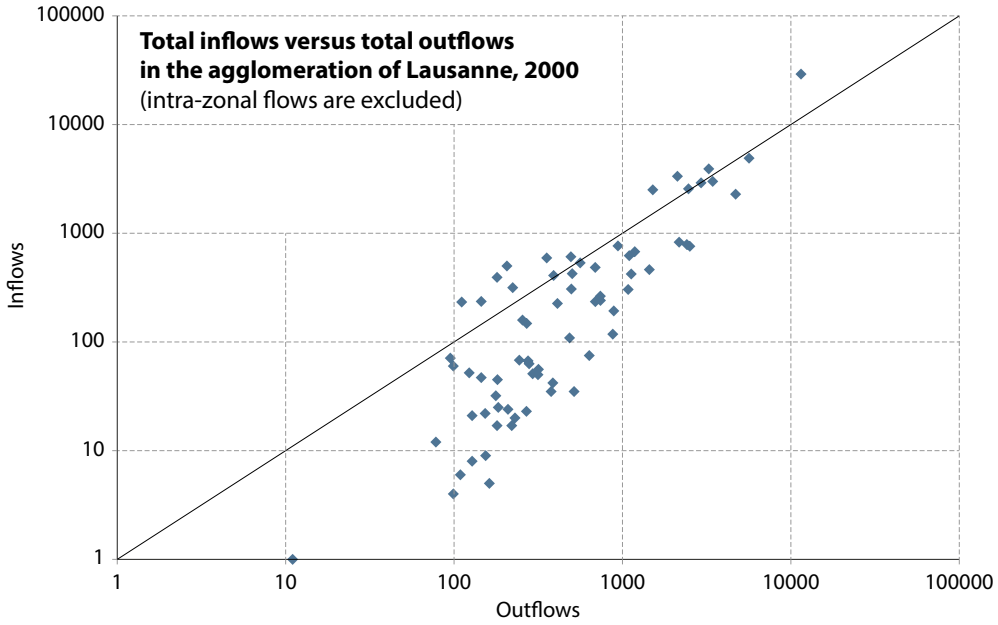


Figure 2.11: Inflows versus outflows in the agglomeration of Lausanne, 2000

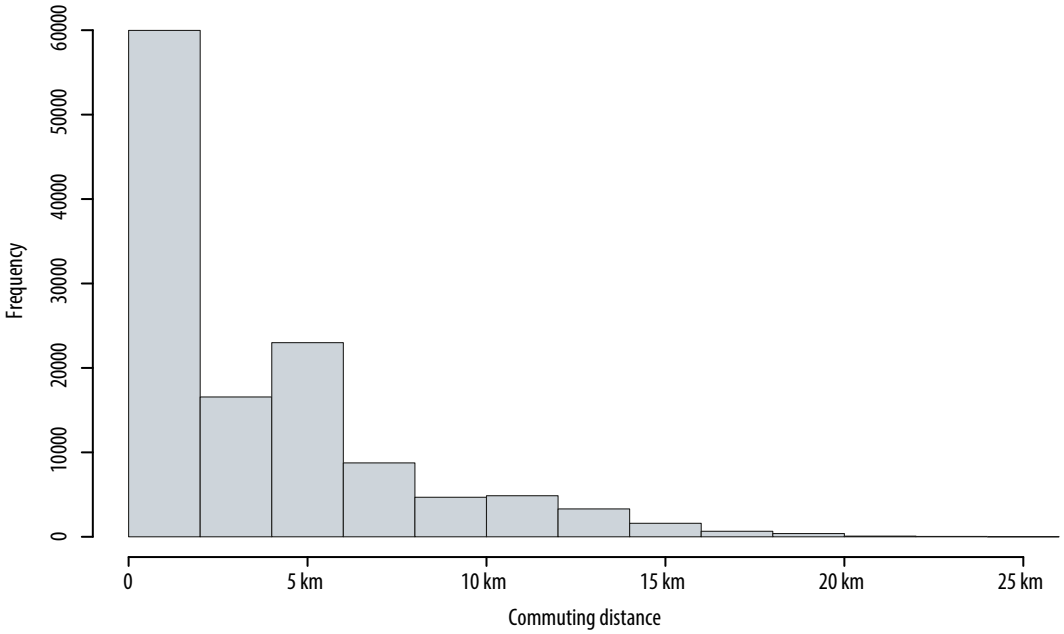


Figure 2.12: Frequency histogram of the commuting distance in the agglomeration of Lausanne, 2000

most of the internal flows are in reality not of distance 0, this results in a under-estimation of the commuting distance. Neverthless, figure 2.12 shows an exponential decrease in the frequency of the flows with increasing distance.

Another interesting map describing the median income of the resident population in the agglomeration of Lausanne in 2005 is shown in figure 2.13. The data are provided by the tax administration of the canton of Vaud. The number of communes in the agglomeration has been reduced to 65 in 2005, as some communes have merged with neighbouring communes. According to the map, residents of city of Lausanne have the lowest income. This is perhaps due to more students and lower-income workers living here. This is also true for the communes west of the city of Lausanne such as Renens, Prilly or Ecublens. In contrast, the commune of Jouxtens-Mézery located just in the north of Lausanne is the commune with the highest median income. This is possibly because of the low tax rate in this commune in comparison with other parts of the agglomeration which make this commune an attractive place to live for wealthier people. The communes of Saint-Sulpice and Buchillon in the middle south and west are famous for being wealthy locations lying along Lake Geneva.

In this chapter the dataset to be used is described in detail. The dataset consists of commuting flows in the agglomeration of Lausanne, located in Western Switzerland. The commuting model in this thesis considers three independent variables of active population, number of jobs and distance between origins and destinations. The data related to

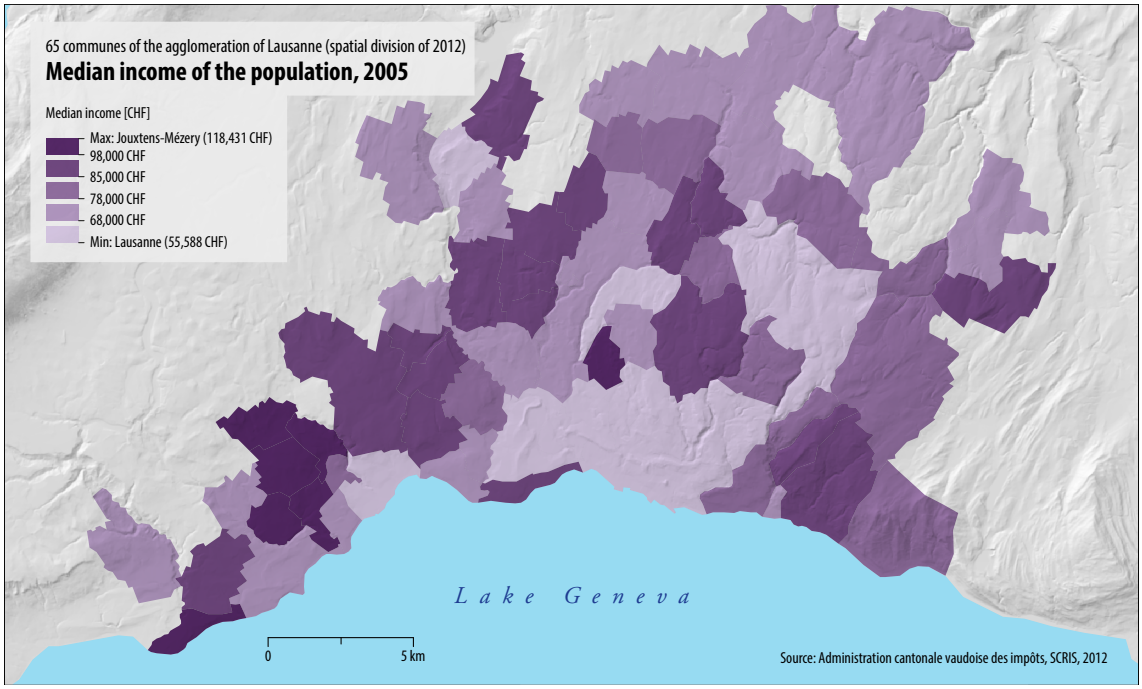


Figure 2.13: Median income in the agglomeration of Lausanne, 2005

these variables are used to produce the maps presented in this chapter.

In the following chapter, spatial interaction models and their underlying theoretical frameworks will be reviewed and a global Poisson gravity model will be calibrated using MLE for the Swiss commuting data described above.

Chapter 3

Spatial flow modelling: an overview

3.1 Introduction

According to Fotheringham and O'Kelly (1989); Fischer (2000) and Fotheringham (2001), spatial interaction can be broadly defined as *the movement or communication of objects such as people, goods and information over geographic space that results from a decision-making process* (also Batten and Boyce, 1986). By this definition, spatial interaction covers a wide variety of behaviours and movements such as migration, shopping trips, commuting, commodity or communication flows, trips for educational purposes, airline passenger traffic, the choice of health care services, the spatial pattern of telephone calls, emails and the World Wide Web connections and even attendance at events like conferences, cultural and sport events (Haynes and Fotheringham, 1984). All of these behaviours form important components of social and urban complex systems. In each case, an individual or group of individuals trade off the benefit of the interaction with the cost of overcoming the separation between them and their possible destinations; hence, these decision-making processes are particularly related to spatial choice. The decision where to relocate in case of migration, where to shop in case of shopping and the decision where to live or where to work in case of journey-to-work are examples of spatial choice in spatial interaction.

Usually spatial interaction systems are complex and multi-dimensional; according to Van-Lierop (1986), this could be due to the fact that *"in reality, multiple dimensions are involved in the geographical dispersion of human activities and the spatial relationship between them"*. This sort of complex system is difficult to model and analyse. It has been shown that even simple spatial interaction processes can show complex, chaotic behaviour (see e.g. Dendrinos and Sonis, 1990; Chen, 2009). However, to facilitate understanding and analysis of the patterns and underlying structures of spatial flows in interaction systems, during the years, researchers in various fields have tried to model spatial

flows through mathematical equations, known broadly as *"spatial interaction models"*. Spatial interaction models can be used for explanatory purposes when each determinant of flows is examined through an associated parameter estimate (Fotheringham and O'Kelly, 1989). These models also can provide the opportunity to predict flows patterns when changes in the interaction system occur; e.g. in a shopping behaviour model, to forecast how patterns of spatial flows will change when a shop in the study area either opens or closes.
In this chapter an overview of spatial interaction models is provided with the following structure: the general form and basic elements of the interaction models are described in section 3.2; the gravity model is introduced in section 3.3 as an early spatial interaction model; entropy and utility maximisation frameworks of spatial interaction models are presented in sections 3.4 and 3.5 respectively; section 3.7 covers the calibration techniques for interaction models and includes the Poisson form of spatial interaction model; and in a final section of 3.9, a Poisson gravity model is applied to journey-to-work flows in Lausanne, as an empirical example of a global interaction model.

3.2 General form and elementary components of spatial interaction modelling

The most general form of a spatial interaction model can be formulated (see e.g. Wilson, 1967; Alonso, 1978; Sen and Sööt, 1981) as:

$$T_{ij} = f(V_i\ W_j\ C_{ij}) \tag{3.1}$$

where the interaction between any pair of origins i and destinations j is specified as T_{ij}, V_i represents a vector of origin factors measuring the propulsiveness of origin i, W_j is a vector of destination attractiveness factors, and C_{ij} represents a vector of separation factors, measuring the separation between zones i and j usually in term of distance, cost or travel time between i and j (Fischer, 2000; Haynes and Fotheringham, 1984).

In spatial interaction analysis, a so-called *"origin-destination matrix"* is often used to display the interactions between different origins and destinations. The size of this matrix is defined by the number of origins and destinations in the interaction system. Table 3.1 represents an origin-destination matrix for an interaction system with m origins and n destinations. The elements, T_{ij}, of this $(m \times n)$ matrix indicate the number of flows between origin i and destination j. Each row of the matrix is allocated to an origin i and the columns are aligned with each destination j. The total number of interactions emanating from each origin i and the total interactions terminating in each destination j are summed in corresponding O_i rows and D_j columns respectively; the sum of all flows in the matrix which represents the total number of interactions in the system is shown by T in table 3.1 (for a reference see e.g. Van-Lierop, 1986). Besides T_{ij}, the variables V_i, W_j and C_{ij} of a spatial interaction model (see equation 3.1) can also be represented in matrix notation. When there are p origin attributes and q destination attributes in a

Table 3.1: Origin-destination matrix

Origin \ Destination	1	2	3	·	·	·	·	·	·	n	Total
1	T_{11}	T_{12}	·	·	·	·	·	·	·	T_{1n}	O_1
2	T_{21}	T_{22}	·	·	·	·	·	·	·	T_{2n}	O_2
3	·	·	·	·	·	·	·	·	·	·	·
·	·	·	·	·	·	·	·	·	·	·	·
·	·	·	·	·	·	·	·	·	·	·	·
·	·	·	·	·	·	·	·	·	·	·	·
·	·	·	·	·	·	·	·	·	·	·	·
·	·	·	·	·	·	·	·	·	·	·	·
m	T_{m1}	T_{m2}	·	·	·	·	·	·	·	T_{mn}	O_m
Total	D_1	D_2	·	·	·	·	·	·	·	D_n	T

system with m origins and n destinations, an $(m \times p)$ matrix V and an $(q \times n)$ matrix W can be used for representing origin and destination attributes respectively, and a $(m \times n)$ matrix C where its elements c_{ij} represent separation between origin i and destination j (generally in terms of distance) can be considered as components in spatial interaction models (Fotheringham and O'Kelly, 1989):

$$V = \begin{pmatrix} v_1^1 & v_1^2 & \cdots & v_1^p \\ v_2^1 & v_2^2 & \cdots & v_2^p \\ \vdots & \vdots & \vdots & \vdots \\ v_m^1 & v_m^2 & \cdots & v_m^p \end{pmatrix} \quad W = \begin{pmatrix} w_1^1 & w_2^1 & \cdots & w_n^1 \\ w_1^2 & w_2^2 & \cdots & w_n^2 \\ \vdots & \vdots & \vdots & \vdots \\ w_1^q & w_2^q & \cdots & w_n^q \end{pmatrix} \quad C = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mn} \end{pmatrix}$$

3.3 Gravity model: an early spatial interaction model

One of the most widely used modelling frameworks for spatial interaction is the Gravity model (Haynes and Fotheringham, 1984) which has a long history in the social sciences (Sen and Smith, 1995), and for which many review texts exist (e.g. Roy and Thill, 2004; Sen and Smith, 1995; Batten and Boyce, 1986; Roy, 2004). The early attempts of understanding regularities in patterns of spatial flows, which can be seen as the starting point of gravity models, date back at least to the works of Carey (1858) and Ravenstein (1885) who observed a greater number of migrants to move between larger and closer cities, *ceteris paribus* (O'Kelly, 2009; Fotheringham et al., 2000).

The essence of the gravity model framework is based on Newton's law of universal gravitation: the attraction between every entity is proportional to their masses and inversely proportional to their distance. During the mid-1850s, Newton's theory began to be used for modelling certain types of human activity between entities physically separated in geographical space (Roy and Thill, 2004). In determining spatial interaction based on Newton's theory, initially the gravitational force was replaced with the number of interactions between origin i and destination j as T_{ij}; the masses were specified by the measured sizes of the interaction regions, for example by their populations: P_i for origins and P_j for destinations (see Stewart, 1941), and the distance factor was expressed as the

centroid-to-centroid distance between the interacting regions d_{ij}, (Roy, 2004), as:

$$T_{ij} = k \, \frac{P_i \, P_j}{d_{ij}} \tag{3.2}$$

where k is a scaling parameter relating the magnitude of T_{ij} to the ratio $P_i P_j / d_{ij}$ (Fotheringham et al., 2000).

The basic formulation of the gravity model in equation 3.2 has evolved over the years to better correspond to complex spatial interaction systems. In order to consider the variation of relationships in real-world situations, the formula 3.2 has been modified to include a freely varying exponent for each of the model variables:

$$T_{ij} = \kappa \, \frac{P_i^{\alpha} \, P_j^{\gamma}}{d_{ij}^{\beta}} \tag{3.3}$$

or alternatively:

$$T_{ij} = \kappa \, P_i^{\alpha} \, P_j^{\gamma} d_{ij}^{\beta} \tag{3.4}$$

where κ (balancing factor), α, γ and β are parameters of the model to be estimated empirically and reflect the nature of the relationship between spatial flows and each of the explanatory variables (Fotheringham et al., 2000). The only difference between equation 3.3 and 3.4 is the value sign of the parameter β. The parameter β, known as *"distance-decay"* or *"friction of distance"* parameter (Haynes and Fotheringham, 1984), has a general negative influence on the total number of interactions. Therefore, the value sign of β in equation 3.3 will be positive while in equation 3.4 β will be negative to indicate a negative effect on the total interaction. Differently from β, both parameters γ and α have positive signs with a general positive influence on the total number of flows in the interaction system (for more information see Haynes and Fotheringham, 1984).

A further modification of the basic gravity model formula has the form of an expanded model which considers a number of attributes of the origins and destinations, rather than only the considered size variables in equations 3.3 and 3.4. The expanded version of the gravity model has the following formula:

$$T_{ij} = \kappa \, \frac{V_{i1}^{\alpha 1} \, V_{i2}^{\alpha 2} \, \cdots \, V_{ip}^{\alpha p} \, W_{j1}^{\gamma 1} \, W_{j2}^{\gamma 2} \, \cdots \, W_{jq}^{\gamma q}}{d_{ij}^{\beta}} \tag{3.5}$$

with p origin attributes, V_i, affecting the magnitude of the flows leaving i and q destination attributes, W_j, affecting the magnitude of flows entering j (Fotheringham et al., 2000); (see Haynes and Fotheringham, 1984, for more details).

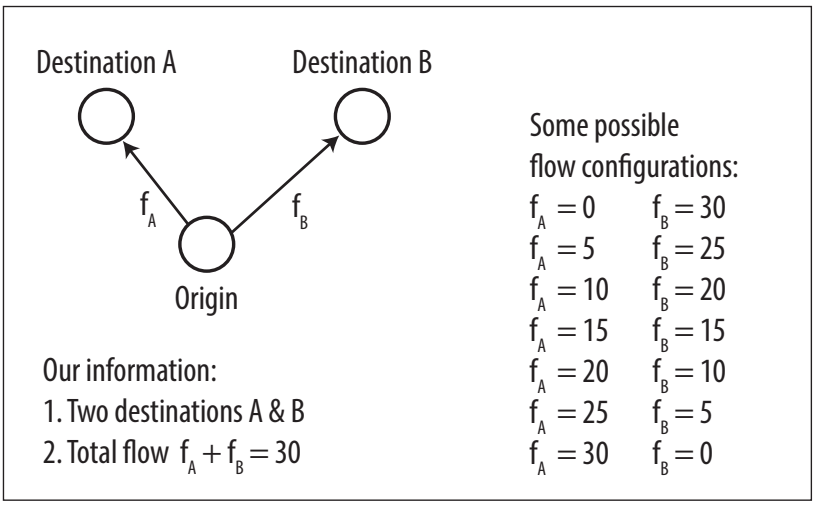
The underlying *"social physical"* framework of the gravity model has been criticised for its lack of theoretical grounding in the way individuals behave (Fotheringham et al., 2000). However, the ability of the model to produce reasonably accurate estimates of

flows and its easy to understand framework made the gravity model one of the most widely used interaction models which is continued to be modified, expanded and used today (Fotheringham et al., 2000; Fischer, 2000). There has been a great deal of effort in the literature to develop a satisfactory theoretical framework for the gravity model, (see e.g. Dodd, 1950; Zipf, 1949; Huff, 1959; Niedercorn and Bechdolt Jr, 1969), where two analytical thought classes of "entropy-maximising" formulation and "utility-theory" approach brought most advances in this regard (Fotheringham et al., 2000).

3.4 Entropy maximisation and the family of spatial interaction models

The entropy maximisation approach estimates the most likely (the most probable) distribution pattern in an interaction system given only limited information, and potentially respecting a set of constraints. The base of this approach was introduced by Wilson (1967, 1970, 1971, 1974). The essence of this approach can be explained with the following example: suppose we have one origin with an outflow of 30 individuals, and two destinations A and B. There are many different possible configurations of the flows between the origin and the destinations (see figure 3.1 for some possible configurations). For a given flow configuration, e.g. $f_A = 5$ and $f_B = 25$, there are R number of ways of

Figure 3.1: Simple spatial interaction system with some possible flow configurations.



assigning each of the 30 individuals to one of the flows f_A or f_B and can be computed as $R = 30!/(f_A! \cdot f_B!)$ or more generally as:

$$R = T! \; / \prod_{ij} T_{ij}! \tag{3.6}$$

If we assign the individuals randomly to one of the flows, the probability of getting one particular configuration is $R/\sum R$; where $\sum R$ is the total number of possible configurations. Therefore, the most probable flow configuration for f_A and f_B is the one where this probability and hence R is maximised. In a spatial interaction model, the flow configuration T_{ij} that maximises R has to be found. Without affecting the maximisation result, we can take the natural logarithm of equation 3.6 and divide it by the total number of flows T, giving a new quantity H:

$$H \equiv \frac{1}{T}\ lnR = \frac{1}{T}\big(ln\ T! - \sum_{ij} ln\ T_{ij}!\big) \tag{3.7}$$

where $ln\ T!$ can be estimated by Stirling's formula of large factorials as: $(T\ lnT - T)$, if all T_{ij} are large; after some rearranging, this would result in:

$$H = -\sum_{ij}(T_{ij}/T)\ \ln(T_{ij}/T) = -\sum_{ij} p_{ij}\ ln(p_{ij}) \tag{3.8}$$

where p_{ij} is the proportion of trips going from i to j (see Fotheringham et al., 2000, for more details). Equation 3.8 is the formula for the entropy of a distribution (Shannon, 1948; Jaynes, 1957; Georgescu-Roegen, 1971, 1986). Therefore, finding the most likely flow configuration in a spatial interaction model can be seen as an entropy maximisation problem. The maximum entropy solution is the one where the different flows T_{ij} have equal values or are all as near to equal as possible; for instance in the above example, the optimal flow configuration is when $f_A = f_B = 15$.

It is possible to add constraints to this entropy maximisation procedure. According to (Fotheringham et al., 2000) some possible constraints that have been imposed on the system are:

$$\sum_{ij} T^*_{ij}\ \ln P_i = P_1 \tag{3.9}$$

where T^*_{ij} is the prediction of the spatial interaction model and P_i is the population of origin i;

$$\sum_{ij} T^*_{ij}\ \ln P_j = P_2 \quad \textit{with } P_j \textit{ the population at destination } j \tag{3.10}$$

$$\sum_{ij} T^*_{ij}\ \ln d_{ij} = D \tag{3.11}$$

where D is the total distance travelled by all individuals together;

$$\sum_{ij} T^*_{ij}\ = K \quad \textit{where } K \textit{ is the known total interaction}; \tag{3.12}$$

$$\sum_{ij} T^*_{ij} = O_i \quad for\ all\ i \tag{3.13}$$

where O_i is the known total flow from each origin (see table3.1); and

$$\sum_{ij} T^*_{ij} = D_j \quad for\ all\ j \tag{3.14}$$

where D_j is the known total inflow into each destination (see table 3.1).
Optimising equation 3.8 using the constraints in equations 3.9 to 3.12 leads to the well known formulation of the gravity model (equation 3.4) (Wilson, 1974). By adding additional constraints on the total of outgoing flows from origins (equation 3.13) and/or on the total of incoming flows to destinations (equation 3.14) it is possible to produce a series of interaction models, called the *family of spatial interaction models* (Wilson, 1974; Fotheringham and O'Kelly, 1989). The application of constraints 3.9 and 3.11 together with constraint 3.14 on the total of incoming flows leads to a spatial interaction model known as *attraction-constrained model*:

$$T_{ij} = \frac{D_j\ P_i^\alpha\ d_{ij}^\beta}{\sum_i\ P_i^\alpha\ d_{ij}^\beta} \tag{3.15}$$

The so-called *production-constrained model* results from maximising equation 3.8 subject to constraints 3.10, 3.11 and the constraint 3.13 on the total of outgoing flows:

$$T_{ij} = \frac{O_i\ P_j^\gamma\ d_{ij}^\beta}{\sum_j\ P_j^\gamma\ d_{ij}^\beta} \tag{3.16}$$

Maximising equation 3.8 subject to constraints 3.11, 3.13 and 3.14 produces a *production-attraction-* or *doubly-constrained* model (Fotheringham et al., 2000):

$$T_{ij} = A_i\ O_i\ B_j\ D_j\ d_{ij}^\beta \tag{3.17}$$

where $A_i = \sum_j (B_j\ D_j\ d_{ij})^{-1}$ and $B_j = \sum_j (A_i\ O_i\ d_{ij})^{-1}$ are balancing factor, to be iteratively adjusted during the model calibration.

Wilson's derivation of spatial interaction based on the entropy maximisation provided an acceptable theoretical justification for the gravity interaction model and the family of spatial interaction models are widely applied in different interaction examples (for instance see Wilson, 1968; Fotheringham and O'Kelly, 1989; O'Kelly, 2009, 2010, 2012; Clarke et al., 1998; Wilson, 2000; Clarke and Clarke, 2001; Nakaya et al., 2007; Singleton et al., 2010). Furthermore, numerous extensions to Wilson's entropy models have been developed in the literature both from a theoretical and practical perspective (see Wilson, 1975; Sen and Smith, 1995; Fotheringham and O'Kelly, 1989; Roy and Thill, 2004; Roy, 2004; Nakaya et al., 2007). For instance, Alonso (1978) proposed a generalised formulation in which each member of the family of interaction models could be obtained from

his framework. Later Fotheringham and Dignan (1984) showed that an infinite number of spatial interaction models can be derived from the Alonso formula (Fotheringham and O'Kelly, 1989). Other examples of extensions to Wilson's entropy models are incorporating neural network into calibration of the models (see for instance Openshaw, 1993; Nakaya, 1995; Fischer, 2002) and integrating the interaction modelling framework with spatial microsimulation (see for instance Birkin and Clarke, 1985, 1988; Clarke et al., 1998; Clarke and Clarke, 2001; Ballas and Clarke, 2001; Birkin et al., 2010; Nakaya et al., 2007). Despite several criticisms of the Wilson's framework in the literature, mainly due to the lack of human behavioural properties in the related models, (see for instance Haynes and Fotheringham, 1984; Fotheringham et al., 2000; Fotheringham, 2001; O'Kelly, 2004), the entropy maximisation remains a popular framework of formulating spatial interaction models. For instance, different entropy maximisation based models have been successfully used in many real-world applications around retailing, where retailing can be seen very broadly as a *"system of interest where there is a flow from a population area to some kind of facility"* (Wilson, 2010). Birkin et al. (2010) show the main issues of using this type of spatial interaction models in an operational environment, and explain some of the modifications and extensions required for addressing issues specific to practical model implementations.

One important modification is a segmentation of the market by customer types, as expressed in the following production-constrained model (for more details, see e.g. Wilson, 2010):

$$S_{ij}^{ngh} = A_i^{ng} \cdot e_i^{ng} \cdot P_i^n \cdot W_j^{gh} \cdot exp(-\beta c_{ij}) \tag{3.18}$$

where S_{ij}^{ngh} is the expenditure by person of type n located in residence zone i for a type of good g in a type of retail store h with location j. e_i^{ng} is the per-capita expenditure by a person of type n located in residential zone i for a good of type g. P_i^n is the population of type n in zone i. W_j^{gh} represents the attraction properties of the retail store h located in destination j, given good type g. c_{ij} is the distance or cost between origin i and destination j, and β is the distance (or cost) deterrance term to be calibrated for each n, g and h. Finally, A_i^{ng} is a balancing factor to ensure that the total expenditure of all persons of type n in zone i for a type of good g in retail store h is met:

$$\sum_j S_{ij}^{ngh} = O_i^{ngh} \tag{3.19}$$

Such a disaggregated entropy-based spatial interaction models takes into account demand, supply and interaction elements. Birkin et al. (2010) note that disaggregation needs to be used carefully, as it might lead to over-fitting of the model and consequently to poor prediction capabilities. This is especially true if the model is calibrated for individual origin or destinations.

This type of spatial interaction model has been used extensively in many studies. As

described by Birkin et al. (2002, 2010), these models have been extensively used in the private sector for retail site location. Additional studies in related fields have also been undertaken. For example, Clarke et al. (2002) have used a entropy-based model in the context of food retailing with the aim of finding zones without sufficient supply in grocery stores, both in quantity and variety. The spatial interaction model is used for estimating the flows for grocery shopping, and indicators are derived based on the model measuring the level of provision for an area or a household. Finding these potential *"food deserts"* is useful for planning purposes. A similar application is retail impact assessment, which tries to assess the impact of new or potential stores (see e.g. England, 2000). If a planned store is likely to have a considerable impact on existing stores, or generates an important amount of traffic, careful evaluation by urban planners is required. While retail impact assessment is traditionally done without spatial interaction model, Khawaldah et al. (2012) have shown that the use of a disaggregated spatial interaction model as described above gives better results.

3.5 Utility maximisation

The family of spatial interaction models derived with the entropy maximisation approach can also be derived in a more behavioural framework using the concept of utility maximisation. The underlying idea of the concept is that the constrained versions of these models allocate the flows emanating from an origin i to a limited number of destinations j which can also be seen as in the framework of discrete choice models (McFadden, 1973, 1978, 1980). For example a person at origin i makes the choice of destination j for a given purpose, such as working or shopping, based on a limited number of attributes of j. The choice of the person will be the one that maximises his/her utility or benefit. The utility is estimated based on two components, V_{ij} and μ_{ij}, where V_{ij} is a measurable component based on the attributes of origin i and destination j, and μ_{ij} is a random component varying between $-\infty$ and $+\infty$. The utility of a person in i derived from selecting destination j can then be computed as:

$$U_{ij} = V_{ij} + \mu_{ij} \tag{3.20}$$

where V_{ij} is a measurable component, μ_{ij} is an unknown and small component relative to V_{ij}.

In theory, it is possible to compute for each destination j the utility a person at i would derive from choosing j and then to find the destination yielding maximum utility. Due to the random term μ, the estimation of the utility cannot be done with certainty. The choice of destination j can only be based on the measurable part of the utility V. If the assumption is made that the random term μ is distributed according to a Gumbel distribution (also named type I extreme value distribution) (Fisher and Tippett, 1928), the probability of choosing destination k out of a set of N destinations j can be computed

as:

$$p_{ik} = \frac{exp(V_{ik})}{\sum_j exp(V_{ij})}. \tag{3.21}$$

Equation 3.21 has the same basic form as a production- or attraction-constrained spatial interaction model. Fotheringham and O'Kelly (1989), Fotheringham et al. (2000) and Pagliara and Timmermans (2009) give a more in-depth discussion of the derivation of the constrained spatial interaction models using the utility maximisation approach and their applications.

3.6 Competing destination model

The derivation of spatial interaction models based on the discrete choice framework (utility maximisation), by providing a human behavior and information processing theoretical foundation for the models, was an improvement over the previous physical analogies to gravitational attraction and entropy Fotheringham et al. (2000). However, the framework of discrete choice is mainly based on aspatial contexts where the number of alternatives are small such as choice of transportation mode. In the derivation of spatial interaction models based on the utility maximisation, it is assumed that an individual can evaluate all alternatives; however, this assumption may not be applicable to spatial choice situations where number of alternatives or choices is usually big. This is because of the fact that human brain capacity for processing information is limited Norman and Bobrow (see e.g. 1975); Bettman (see e.g. 1979); Fotheringham et al. (see e.g. 2000).

In order to improve performance of the spatial interaction models in spatial choice situations, one suggested solution is to assume that individuals consider a hierarchical spatial information processing when evaluating different alternatives and to incorporate relevant measures explicitly, mainly measured with an accessibility variable, in the model (Fotheringham, 1981, 1982b,a, 1983, 1984a,b). Fotheringham in a series of publications introduced an improved form of spatial interaction models, called competing destinations models. In this type of model, an accessibility variable which measures competition between different destinations in the system, is added to the model formula (see Fotheringham, 1983, 1984b, 1986).

The competing destinations models were improved and developed further later and used in different applications (see for instance Fotheringham, 1987, 1988a; Gitlesen and Thorsen, 2000; Thill and Kim, 2005; Chun et al., 2012; Uboe, 2004; Brown and Andreson, 2002; Pellegrini and Fotheringham, 1999; Kwan, 1998; Fotheringham and Trew, 1993; Pellegrini et al., 1997; Thorsen and Gitlesen, 1998; Guldmann, 1999; Fotheringham et al., 2001b; Lo, 1991).

3.7 Calibration of spatial interaction models

Different calibration techniques exist for estimating the parameters of spatial interaction models. Traditionally in the calibration of spatial interaction models it was assumed that these models follow a normal (Gaussian) distribution which allowed a linear regression calibration technique, such as ordinary least squares, to be used for estimating the model parameters. Ordinary Least Squares (OLS) is a calibration technique which tries to find the parameter values of the model which best fit the data set, based on minimising the sum of the squared error term. To show the derivation of OLS estimators, consider a mathematical model with a linear combination of dependent variable y_i and an independent variable x_i as:

$$y_i = \alpha + \beta x_i + e_i \tag{3.22}$$

where e_i is the error (residual) term, α and β are coefficients to be estimated. In order to minimise the sum of squared residuals (SSR) in equation 3.23, OLS solves the equations of the first order derivatives equal to zero (see equations 3.24 and 3.25). The maximisation problem can be written as:

$$\begin{aligned} minimising \quad SSR(\hat{\alpha},\hat{\beta}) &= \sum_i \hat{e}_i^2 = \sum_i (y_i - \hat{y}_i)^2 \\ &= \sum_i (y_i - \hat{\alpha} - \hat{\beta}x_i)^2 \qquad (a) \\ when \quad \frac{\partial(SSR)}{\partial\hat{\alpha}} &= 0 \qquad (b) \\ and \quad \frac{\partial(SSR)}{\partial\hat{\beta}} &= 0 \qquad (c) \end{aligned} \tag{3.23}$$

To derive parameter α based on equation 3.23 (b):

$$\begin{aligned} -2\sum_i (y_i - \hat{\alpha} - \hat{\beta}x_i) = 0 \qquad (divide\ by\ -2n) \qquad (a) \\ \Rightarrow \quad \sum_i y_i - n\hat{\alpha} - \sum_i \hat{\beta}x_i = 0 \qquad (rearrange) \qquad (b) \\ \Rightarrow \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \qquad (c) \end{aligned} \tag{3.24}$$

For estimating the variable β, based on equation 3.23 (c):

$$\begin{aligned} -2x_i \sum_i (y_i - \hat{\alpha} - \hat{\beta}x_i) = 0 \qquad (divide\ by\ -2) \qquad (a) \\ \sum_i x_i y_i = \hat{\alpha}\sum x_i + \hat{\beta}\sum x_i^2 \qquad (b) \end{aligned} \tag{3.25}$$

multiply 3.24 (b) by $\sum_i x_i$ and 3.25 (b) by n:

$$\begin{aligned} \sum_i y_i \sum x_i = n\hat{\alpha} \sum x_i + \hat{\beta}(\sum_i x_i)^2 = 0 \qquad (a) \\ n \sum_i x_i y_i = n\hat{\alpha} \sum x_i + n\ \hat{\beta} \sum_i x_i^2 = 0 \qquad (b) \end{aligned} \tag{3.26}$$

subtracting 3.26 (a) from 3.26 (b):

$$n \sum_i x_i y_i - \sum_i y_i \sum x_i = n\ \hat{\beta} \sum_i x_i^2 - \hat{\beta}(\sum_i x_i)^2 \tag{3.27}$$

and finally

$$\hat{\beta} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2} = \frac{\sum x_i y_i - \bar{y} \sum x_i}{\sum x_i^2 - \bar{x} \sum x_i} \tag{3.28}$$

The OLS technique can also be used for calibrating equations with more than two parameters. In this case, the model formulation can be written in a matrix notation as $y = X\ \beta + e$ where y is a $(m \times 1)$ vector of the dependent variable, X is a $(m \times p)$ matrix of independent variables, including a column of one for the intercept parameter and β is a $(p \times 1)$ vector of parameters to be estimated. Following the same principle of minimising the sum of squared residuals, the estimated parameters of the model can be derived from the following equation:

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad \text{where } X^T \text{ indicates the transpose of } X. \tag{3.29}$$

When using matrix notation of OLS for calibrating the spatial interaction models, the independent variable T_{ij} representing spatial flows from origin i to destination j is normally distributed and a mean of spatial flows can be achieved with a linear combination of the independent variables. For example, consider a general formula of the gravity model between m origins and n destinations: $T_{ij} = \kappa\ P_i^{\alpha}\ P_j^{\gamma}\ d_{ij}^{\beta}$; taking logarithms from both sides of the equation, the formulation of the model is in the form of linear regression:

$$\log T_{ij} = \log \kappa + \alpha \log P_i + \gamma \log P_j + \beta \log d_{ij} + e_{ij} \tag{3.30}$$

where P_i, P_j and d_{ij} are vectors containing the values of explanatory variables of origins and destinations and distance between them; κ, λ, α and β are parameters of the model and the error term e_{ij} is identical and independent and follows a normal distribution with zero mean. Using the matrix notation of the least squares technique for minimising

the sum of square error, a vector of parameter estimates $\hat{\beta}$, can be obtained as:

$$\hat{\beta} = \begin{pmatrix} \log \kappa \\ \alpha \\ \gamma \\ \beta \end{pmatrix} \quad \text{when} \quad y = \begin{pmatrix} \log T_{12} \\ \log T_{13} \\ \vdots \\ \log T_{1n} \\ \log T_{21} \\ \log T_{23} \\ \\ \vdots \\ \log T_{2n} \\ \vdots \\ \log T_{m1} \\ \log T_{m2} \\ \vdots \\ \log T_{mn} \end{pmatrix} \quad \text{and} \quad X = \begin{pmatrix} 1 & \log P_1 & \log P_2 & \log d_{12} \\ 1 & \log P_1 & \log P_3 & \log d_{13} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \log P_2 & \log P_1 & \log d_{21} \\ 1 & \log P_2 & \log P_3 & \log d_{23} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \log P_2 & \log P_n & \log d_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \log P_m & \log P_1 & \log d_{m1} \\ 1 & \log P_m & \log P_2 & \log d_{m2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \log P_m & \log P_n & \log d_{mn} \end{pmatrix}$$

3.8 Poisson spatial interaction model and maximum likelihood

The form of spatial interaction models that are calibrated using regression techniques are often referred to as *log-normal models* (see Flowerdew and Aitkin, 1982). Although regression is one of the most commonly used mechanisms for calibrating the log-normal models, there are several problems associated with this technique. Here we briefly review two of these problems which are most mentioned in the literature (for further information see Flowerdew and Aitkin, 1982; Fotheringham and O'Kelly, 1989, *inter alia*). The zero flow problem is the first one. Regression techniques require linearising the interaction model by applying a logarithmic function to both sides of the model equation. The application of a logarithmic function is problematic if there are any zero flows in the model, because the logarithm of zero is undefined (Fotheringham and O'Kelly, 1989). It is obvious that in many spatial interaction systems, a number of spatial flows might be zero. For instance, in a journey-to-work model covering cities in a country, it is rather unusual if commuters travel more than several hours for their daily travel to work so in this case usually the interaction flows between far locations are zero. Or in a trade flow model, as it is stated in Burger et al. (2009), factors like distance, cost, or absence of historical and cultural links between distant countries can be some reasons for the absence of trade and so having zero spatial flow in the model is fairly common (see also Frankel, 1997; Rauch, 1999).

A second problem associated with the calibration of the spatial interaction models with regression technique is the assumption of normal distribution of the flows. As spatial interactions are discrete phenomena indicating a given (nonnegative integer) number of

flows between a set of origins and destinations, assuming a continuous distribution for the flows might be an unrealistic assumption (see Flowerdew and Aitkin, 1982).

To overcome the above-mentioned problems, some possible solutions have been suggested. The simplest solution would be to remove all zero interactions from the analysis but this would result in a biased dataset and misleading parameter estimates that do not reflect the low interactions (Fotheringham and O'Kelly, 1989; Eichengreen and Irwin, 1998). Another solution for the zero flow problem is to add a small positive constant (usually 0.5 or 1) to elements of the interaction matrix; for instance to add a small constant to every flow in the system, or add this constant only to the zero flows (see e.g. Linders and de Groot, 2006). However in any case, it has been shown that the model calibration is sensitive to this constant in which different constant values added to the flows would result in different parameter estimates (see Flowerdew and Aitkin, 1982; King, 1988). Therefore, adding a constant to elements of the flow matrix, would potentially bring some bias in the model calibration and would not be a proper solution for the problem of zero flows in the log-normal models.

In a more promising approach for solving the above problems, considering the fact that spatial flows are count data, Flowerdew and Aitkin (1982) suggested that in modelling spatial interaction, each T_{ij} variable should be regarded as having a discrete probability distribution such as Poisson. If there is a constant probability of any individual moving between i and j when the population of i is large and movements of individuals are independent, then interactions have a Poisson distribution with mean λ_{ij} (Flowerdew and Aitkin, 1982; Fotheringham and Williams, 1983; Flowerdew and Lovett, 1988; Lovett and Flowerdew, 1989). Therefore, the probability that T_{ij} number of people are moving between i and j is given by:

$$Pr(T_{ij}) = \frac{e^{-\lambda_{ij}}\ \lambda_{ij}^{T_{ij}}}{T_{ij}!} \tag{3.31}$$

where λ_{ij} is the expected mean value of the variable T_{ij} and should be estimated.

In the Poisson regression model, the mean parameter depends on the explanatory variables (Cameron and Trivedi, 1998) through the *generalised linear model (GLM)* (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989). GLM extends the basic regression model, where the dependent variable is assumed to follow a Gaussian distribution, to the case of the exponential family of distributions (Fotheringham et al., 2002), such as Poisson, binomial distribution, *inter alia*. The expected value of the dependent variable in a GLM is defined in terms of a linear function of the explanatory variables:

$$g\big(E(y_i)\big) = a_0 + a_1 x_{i1} + a_2 x_{i2} + \cdots + a_p x_{ip} \tag{3.32}$$

where the function $g()$ is referred to as "link function" and is specified based on the type of model distribution. The Poisson regression corresponds to the GLM with "log" link

function (Kamo et al., 2009); so that:

$$\log\big(E(y_i)\big) = a_0 + \sum_{k=1}^{p} a_k x_{ik} \tag{3.33}$$

where p is the number of explanatory variables, excluding the intercept.

In equation 3.31 where the spatial interaction model is considered to follow a Poisson distribution, the expected mean value of flows λ_{ij} is unknown and can be estimated from a spatial flow model (e.g. a simple gravity model) logarithmically linked to a linear combination of the logged independent variables (see Flowerdew and Aitkin, 1982; Flowerdew, 1982):

$$\ln(\lambda_{ij}) = \kappa + \alpha \ln P_i + \gamma \ln P_j + \beta \ln d_{ij} \tag{3.34}$$

or equally as:

$$\lambda_{ij} = \exp(\kappa + \alpha \ln P_i + \gamma \ln P_j + \beta \ln d_{ij}) \tag{3.35}$$

when an exponential function, $exp()$, is applied to the both sides of equation 3.34.

Initially Flowerdew and Aitkin (1982) calibrated λ_{ij} using an iteratively reweighted least squares techniques described by Nelder and Wedderburn (1972); however, later Fotheringham and Williams (1983) suggested the maximum-likelihood calibration technique for this purpose. Maximum likelihood is a calibration technique which estimates parameters of the model by maximising the probability (likelihood) of the sample data. Suppose that X is a random variable with density function $f(X; \theta_1, ..., \theta_k)$, where $(\theta_1, ..., \theta_k) = \theta$ are k parameters to be estimated. If $x_1, ..., x_n$ is a set of n independent observations, the likelihood function is given by:

$$L(\theta) = f(x_1, x_2, ..., x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta) \tag{3.36}$$

The maximum likelihood estimators (MLE) of θ are obtained by maximising the function $L(\theta)$ by: $\partial L(\theta)/\partial\theta = 0$.

In practice it is easier to work with the logarithm of the likelihood function, called the log-likelihood:

$$\begin{aligned} lnL(\theta) &= \sum_{i=1}^{n} lnf(x_i; \theta) \qquad (a) \\ and\ then: \quad \frac{\partial\ lnL(\theta)}{\partial\theta} &= 0 \qquad (b) \end{aligned} \tag{3.37}$$

In order to calibrate λ_{ij} in the Poisson spatial interaction model represented in equa-

tion 3.31 using the maximum likelihood technique, consider $\lambda_{ij} = k\ P_i^{\alpha}\ P_j^{\gamma}\ d_{ij}^{-\beta}$ which is equivalent to equation 3.34 when $k = \exp(\kappa)$; the log-likelihood function is:

$$lnL(\lambda_{ij}) = \sum_i \sum_j (-\lambda_{ij} + T_{ij}\ ln\lambda_{ij} - ln\ T_{ij}!) \tag{3.38}$$

where T_{ij} is known. The formula is maximised when:

$$\frac{\partial\ lnL(\lambda_{ij})}{\partial \phi} = \sum_i \sum_j T_{ij}\ lnx_i - \sum_i \sum_j \lambda_{ij}\ lnx_i = 0 \tag{3.39}$$

where x_i is the independent variable associated with the parameter ϕ (Fotheringham and O'Kelly, 1989). For example, for estimating the parameter γ:

$$\begin{aligned}
\frac{\partial\ lnL(\lambda_{ij})}{\partial \gamma} &= \frac{\partial}{\partial \gamma} \sum_i \sum_j (T_{ij}\ ln\lambda_{ij} - \lambda_{ij} - lnT_{ij}!) \quad (a)\\
&= \frac{\partial}{\partial \gamma} \sum_i \sum_j \left(T_{ij}\ ln(k\ P_i^{\alpha}\ P_j^{\gamma}\ d_{ij}^{-\beta}) - (k\ P_i^{\alpha}\ P_j^{\gamma}\ d_{ij}^{-\beta})\right) \quad (b)\\
&= \frac{\partial}{\partial \gamma} \sum_i \sum_j \left(T_{ij}\ (lnk + \alpha\ lnP_i + \gamma\ \mathbf{lnP_j} - \beta\ lnd_{ij}) - (k\ P_i^{\alpha}\ \mathbf{P_j^{\gamma}}\ d_{ij}^{-\beta})\right) \quad (c)\\
\frac{\partial\ lnL(\lambda_{ij})}{\partial \gamma} &= \sum_i \sum_j \left(T_{ij}(\mathbf{lnP_j}) - (k\ P_i^{\alpha}\ d_{ij}^{-\beta}\ P_j^{\gamma}\ \mathbf{lnP_j})\right) = 0 \quad (d)
\end{aligned} \tag{3.40}$$

where formula (d) in equation 3.40 is equivalent to equation 3.39. [1]

3.9 Journey-to-work Poisson gravity model in Lausanne

In order to illustrate an application of spatial flow models on a real-world interaction example, we apply a global gravity model on the journey-to-work (commuting) dataset in Lausanne. We consider a Poisson gravity model with three variables as follows:

$$\lambda_{ij} = \exp(\kappa + \alpha \ln P_i + \gamma \ln N_j + \beta \ln d_{ij}) \tag{3.41}$$

where P_i indicates the *number of economically active population* in origin i; N_j is the *number of jobs* in destination j; d_{ij} shows the distance between centroids of the commuting regions; κ, α, γ and β are parameters of the model that will be calibrated using MLE technique in this example.

As already mentioned in chapter 2, the agglomeration of Lausanne covers a total of 70 communes which are the smallest administrative units in Switzerland. In this interaction example, we follow the conventional spatial interaction models that usually consider only inter-zonal flows between origins and destinations but not the interactions within the

[1] Note that $\frac{\partial}{\partial x}(u^{ax}) = u^{ax}\ \ln u\ .\ a, \quad u > 0$

zones themselves. Therefore, in the above interaction model, only inter-zonal commuting flows are modelled and intra-zonal flows are eliminated from the model. This results in a (70×70) origin-destination matrix in equation 3.41, with 4830 inter-zonal flows between the 70 commuting communes. We calibrate this gravity model using the MLE technique and the results are listed in table 3.2, including the estimated t-values, p-values and standard error of each parameter. The McFadden's Pseudo R^2 and Akaike's information criterion (AIC) (Akaike, 1974, 1973) values are also calculated to be used as measures of goodness-of-fit for the model.

Table 3.2: Global Poisson gravity model for journey-to-work in agglomeration of Lausanne

Global Pseudo R^2	0.945
Global AIC	21381.8

Parameter	Estimated value	t-value	p-value	Standard error
k	-5.70	-94.82	0.00	0.0601
α	0.85	316.37	0.00	0.0027
γ	1.01	450.63	0.00	0.0022
β	-0.67	-130.59	0.00	0.0051

The estimated value of the parameter β is -0.67 indicating the negative influence of distance on interaction. The estimated values of the parameters α (equal to 0.85) and γ (equal to 1.01) show positive effects on interaction volume. This is expected as usually increasing the number of workforce and number of existing jobs will increase the number of commuting flows. However, it is interesting to note that according to the estimated values of α and γ and comparing their t-values (316.37 and 450.63 respectively), the influence of the number of jobs at a destination seems to be higher than the number of workers at an origin suggesting that jobs are filled but there might be a surplus of labour. The p-values of all parameters are close to zero which indicate the significance of the model parameters.

The global AIC in this model is equal to 21381.8; however, the AIC is usually a measure for model selection and so it is not interpreted directly but rather compared to AIC(s) from other models fitted to the same data. The McFadden's Pseudo R^2 is based on the maximum log likelihood (see McFadden, 1973; Agresti, 1990; Joost and Kalbermatten, 2010) and reflects the degree of improvement of the model with predictors over the intercept model (Hu et al., 2006). The theoretical range of the $McFadden's\ Pseudo\ R^2$ is between 0 and 1, where 1 indicates perfect predictability. Although according to Hu et al. (2006) and Joost and Kalbermatten (2010), this coefficient has the tendency to underestimate the real R^2 and it never reaches 1, the result of our model shows a reasonable goodness-of-fit with $McFadden's\ Pseudo\ R^2$ equal to 0.945.

3.10 Spatial flow modelling outside geography

In the last few years, considerable research efforts in spatial flow modelling have emerged in fields outside of geography, for example in physics, computer science or complex systems studies. This recent activity can be explained to some extent by the availability of many high-quality spatial interaction datasets. For instance, Simini et al. (2012) try to develop a universal model for spatial interaction by developing a so-called "radiation model". The authors have a background in complex network research and physics, and they claim their model to be superior to the unconstrained gravity model. This radiation model has the interesting property not to rely directly on the geographical distance, and to be completely parameter free. The radiation model is built on the particle emission and absorption processes in physics with the following formulation:

$$\hat{T}_{ij} = T_i \frac{m_i n_j}{(m_i + s_{ij})(m_i + n_j + s_{ij})} \tag{3.42}$$

where $\hat{T}_{ij}$ is the predicted flow, T_i the commuters from population i. m_i is the population in origin i, n_j the population in destination j. s_{ij} is the population in a circle around origin i with radius d_{ij}, (distance between origin i and destination j), without the populations of i and j. In the context of commuting interaction, the idea behind the model can be seen as the choice of the closest job with better benefits (e.g. salary) by each individual. The authors have applied this models to commuting trips between US counties, population migration and commodity flows in the US, and phone calls, with very good analytical results.

Additional work undertaken by Masucci et al. (2013), however, has shown some limitations of the radiation model suggested by Simini et al. (2012). These limitations are mainly related to the model's normalisation to an infinite system that does not accurately represent the situation in many spatial interaction systems. A correction to the radiation model has been suggested by Masucci et al. (2013). In this paper, it is also shown that the calibrated unconstrained gravity model outperforms the radiation model in the case of journey-to-work interaction between wards in England and Wales but the radiation model performs better than the gravity model for interaction systems with longer distances such as migration.

Another new human mobility modelling framework has been presented by Simini et al. (2013), where the space is not considered as being discrete as in previous models such as the gravity model or the radiation model. The framework allows for a unified view of the radiation model presented in Simini et al. (2012), the intervening opportunities model (Stouffer, 1940; Schmitt and Greene, 1978) and a new variant of the radiation model with selection. Further, the authors provide an interesting insight into the theoretical foundations of the unconstrained gravity model in the case of homogeneous space. If the model calibration yields α or β parameters for attraction and propulsiveness variables differing from 1, job offers are not equally distributed in space.

Lenormand et al. (2012) have also presented a model for estimating commuting flows based on the number of outgoing and incoming commuters for each region. According to the authors, their model outperforms the radiation model, especially if the geographic units are small, such as municipalities. The model is close to a doubly-constrained gravity model, but follows a probabilistic simulation approach where individual commuters are allocated one by one to depending on probabilities increasing with the number of commuters towards a destination, and decrease with distance. The model depends on a single parameter β which represents the importance of distance in commuting choice. Compared to the doubly-constrained gravity model, this *"universal model of commuting networks"* can deal with more flexible data such as in- or outflows from other regions. Lenormand et al. (2012) have also adapted the radiation model to a doubly-constrained model, while the authors claim the universal model of commuting networks still outperforms the doubly-constrained radiation model.

Yet another model has been developed by Yan et al. (2013), labeled *"conduction model"*. The authors present their model as an alternative to the radiation model, but better addressing issues with intra-urban mobility. The model is derived from a stochastic decision making process where each individual selects its destinations according to some probabilities. Like the radiation model, the conduction model is parameter free. The application of this model to several cities across the world showed better results compared to the radiation model, but lacks comparison to traditional spatial interaction models. Due to the individual decision making process underlying the model, similarities to models such as the competing destination model presented shortly in section 3.6 could arise, but have not yet been investigated.

Within the field of complex systems, Ruzzenenti et al. (2012) have recently developed a model for extracting spatial and non-spatial effects from spatially embedded networks. The authors analyse the world-wide trade network and recognise that the gravity model can quite accurately estimate the trade volume by using distance and GDP. However, the gravity model seems fails in predicting zero-flows, even in a zero-inflated version (Dueñas and Fagiolo, 2013) resulting in poor estimates for the in- and out-degrees of the network nodes. As a solution, the authors use null models to preserve non-spatial constraints of the model, while still detecting spatial effects. Applications of this model to other datasets are currently missing but are required in order to evaluate how useful the model is in different contexts.

Spatial interaction systems have also been studied using data from mobile phone networks (e.g. Song et al., 2010; Csáji et al., 2013; Lambiotte et al., 2008; Krings et al., 2009). For instance, Expert et al. (e.g. 2011); Blondel et al. (e.g. 2008) focus on the problem of community detection in human networks using large mobile phone datasets. Expert et al. (e.g. 2011) proposed a modularity function adapted to spatial networks to factor out the effect of space and to detect hidden structural similarities between the nodes in the network. Another important study field relating to human networks that

is analysed using spatial interaction models is spreading of epidemics (e.g. Balcan et al., 2009; Colizza et al., 2006; Ferguson et al., 2006; Dalziel et al., 2013).

3.11 Zoning problems in spatial interaction

Data about spatial interaction flows are usually collected at the level of some administrative zones. While zonal aggregation is required to successfully calibrate most spatial interaction models, it also raises several problems. First of all, the known issues related to the Modifiable Area Unit Problem (MAUP; see e.g. Openshaw and Taylor, 1981; Openshaw, 1984) apply also to spatial interaction models, as propulsiveness and attractiveness are assessed at a zonal level. Additionally, problems specific to the flows arise:

- Internal flows are flows where the origin and destination of a trip are within the same zone. Internal flows are difficult to include in a model, as the length of the intra-zonal trips is unknown, and the travelled distance is usually included in the model. As a result, internal flows are often excluded from the calibration process. This issue will be discussed more in depth in chapter 4.

- In some cases, especially if zones are relatively big, it can make a considerable difference where in the zone the origin or a destination of a trip is located. This can be the case for example if a whole city is a single zone, and it is unknown if the commuters have to cross the city or not. In the case of neighbouring zones, origin and destination of a trip can be very close, or very far. Generally, the zone centroids are considered in spatial interaction models in order to estimate travel distance or travel time. The methods that will be discussed in chapter 4 can be used to some extent to deal with this problem.

Possible approaches for dealing with the MAUP exist and are extensively discussed in the literature. One possibility is dasymetric mapping, (e.g. Bentley et al., 2013; Langford et al., 2008; Mennis, 2009), which can also be used for downsampling purposes in order to provide estimations for example at the level of a regular grid. Krider and Putler (2013) have to deal with the opposite problem; due to increasing availability of high quality geographical data at the address level, the authors had to aggregate the individual point locations into a regular grid in order to simplify the spatial interaction modelling process.

Data downsampling is a common problem in microsimulation models, where properties have to be given to individual agents. The whole population of agents should then reflect the statistical properties of the available real-world data. A popular technique for solving this problem is Iterative Proportional Fitting (IPF; see e.g. Ireland and Kullback, 1968; Frick and Axhausen, 2004; Müller and Axhausen, 2010). It is also possible to deal with the zoning problem at the level of the spatial interaction model itself. For example Birkin et al. (2010) describe a "boundary-free model" for retail centers which might have catchment areas overlapping zone boundaries. Another approach is taken bySimini et al.

(2013) who have adapted their radiation model presented as a discrete model in Simini et al. (2012) to account for continuous space and dealing with eventual zoning problems directly at the level of the model.

3.12 Visualisation of spatial interaction

Visualisation of spatial interaction flows is a challenge due to the big quantity of information to show. Flow maps represent interaction as lines where the line width is mostly proportional to the magnitude of the interaction. Flow maps have been extensively used and described in the litterature, e.g. by Tobler (1981, 1987); Phan et al. (2005); Guo (2009); Rae (2009). Dorling (1991) has shown a flow map in a population cartogram. Flow maps quickly become cluttered as the number of flows increase. Different techniques have been developed to make flow maps easier to read, for example flow bundling (Phan et al., 2005; Holten and Van Wijk, 2009), filtering by distance, location or magnitude (Rae, 2009), clustering of origin and destination regions (Guo, 2009), or calculating a field of net exchange vectors (Tobler, 1976, 1981).

An alternative representation method is the non-geographic origin-destination matrix, where the rows and columns of the matrix represent the origins respectively destinations. Each matrix cell is coloured according to the magnitude of the flow. This kind of spatial interaction representation has been used for example in Bertin (1983); Guo (2007); Wong et al. (2006). One of the question in matrix representations is the order of the rows and columns. Many possibilities to this problem have been discussed in the litterature (e.g. Wilkinson, 1979; Friendly, 2009; Mäkinen and Siirtola, 2000), such as using clustering techniques. In an interactive visualisation, the user can be given the possibility to reorder the rows and columns.

Wood et al. (2010) divide the geographical space in cells, and inside each cell, they insert a small matrix covering a miniature version of the map. Inside each cell, the flows towards all locations are shown as a small heatmap matrix. The authors refer to this kind of visualisation as *"OD maps"*. An application of the OD maps can be found in Slingsby et al. (2012). Graph layouts can also be used to visualise spatial interaction flows, as shown for example in Wong et al. (2006) where an interactive tool is shown allowing for representation as a graph and a compact matrix. However, Ghoniem et al. (2004) found graph representation to be less optimal than simple matrix representations except for very small graphs.

In interactive visualisations, it is possible to link spatial and non-spatial views together. This approach has been used for example by Boyandin et al. (2011), where the origins and destinations are represented in two separate maps, and the temporal evolution of migration flows between countries as a heatmap matrix. The matrix is located between the origin- and destination-map, and all three views are connected through lines appearing when the user interacts with the visualisation.

In the above chapter we briefly reviewed the spatial interaction models and their underlying theoretical frameworks. The most commonly used calibration methods of spatial interaction models were discussed and introduced in detail including Poisson specification of the spatial interaction models. A Poisson gravity model was calibrated with the MLE technique when it was applied on the real-world dataset of journey-to-work interactions in Lausanne. The journey-to-work interaction model covered the inter-zonal commuting flows between 70 communes in the agglomeration while the intra-zonal flows within the zones were eliminated from the analysis.

Although ignoring the intra-zonal flows is one of the common solutions for the problem of estimating the average trip length for the intra-zonal flows in the spatial interaction analysis, it may result in some complications such as a reduced data sample that does not accurately represent the processes under study and a biased model calibration. In the next chapter we address this problem and propose a methodology for estimating the average intra-zonal trip length based by scattering the origins and destinations of the flows within their geographical zones.

Chapter 4

Intra-zonal trip length in spatial interaction models

4.1 Introduction

The problem of estimating average trip-length for intra-zonal flows has a long history in spatial modelling. In spatial interaction modelling, the data used in model calibration are often aggregated to represent flows between different zones and the attributes of these zones are assumed to be located at their centroids rather than distributed continuously in space. In such a case, the origin and destination of the flows are assumed to also be located at the centroids of the zones, so that the separation variable, d_{ij} is often approximated as the centroid-to-centroid travel cost (or distance or time) between the zones. However, this assumption might cause poor estimation of the separation variable in the model, especially when populations are not concentrated around the centroid of the zone. Furthermore, using the centroid-to-centroid approximation for the trip-length can be problematic when we have to calculate the intra-zonal distances. For instance in our case study of journey-to-work in Lausanne where a considerable number of workforce staying inside their commune of residence for their work, the intra-zonal distances make up a significant amount of information and should be considered in the analysis. However, in this case the distance calculation between a zone centroid and itself will lead to a centroid-to-centroid intra-zonal distance (d_{ij}) of zero, whereas in reality the actual average distance commuters travelled within the zone would be positive (see Fotheringham, 1988b; Fotheringham and O'Kelly, 1989, for a more detailed discussion).

The intra-zonal flows are often ignored in model calibration, because of the above problems, but eliminating these flows may result in a reduced data sample that does not accurately represent the processes under study, giving rise to biased parameter estimates. There have been some attempts in the literature to estimate the intra-zonal trip length based on algebraic methods. These methods are mostly based on various strong assumptions related to zone shape and population distribution and are mainly

highly approximate and more of analytical interest than viable solutions. In this chapter a methodology is suggested for estimating the average intra-zonal trip length based on a scattering of the origins and destinations of the flows within their geographical zones (scattering-based method). The process of scattering the origins and destinations can be done randomly or based on available spatial density information, such as population density. The average trip length is then calculated for all possible trip configurations within the system.

The remainder of this chapter is structured as follows: A review of some existing methods for estimating intra-zonal distance are discussed in section 4.2. The methodology of scattered-based models both for random and density-based approaches is discussed in sections 4.3.1 and 4.3.2 respectively. The models are then applied to the Lausanne journey-to-work dataset and the results of the proposed and traditional methods are compared in section 4.4. The comparison of the methods reveals that the density-based scattering models have a better model fit and less error; finally a summary is presented in section 4.5.

4.2 Background

According to Batty (1976), the problem of intra-zonal distance measurement is perhaps one of the main concerns in many spatial modelling applications. A common approach to avoid this problem is to simply exclude the internal flows from the analysis. However, in many cases spatial interaction models are applied to short-distance flows, e.g. journey-to-work, where the intra-zonal flows can make up a significant percentage of all flows; therefore omitting them from the model might not be appropriate. For instance, in a typical urban agglomeration, many people live and work within the main city itself. In the Lausanne journey-to-work dataset, more than 45% of the flows are intra-zonal, showing that a significant number of people work within their commune of residence. In such cases, the model calibration is likely to be very sensitive to the inclusion and exclusion of intra-zonal flows. In fact eliminating the intra-zonal flows from the model will certainly ignore a large amount of information and result in a reduced data sample that does not accurately represent the processes under study. Furthermore, the intra-zonal flows are generally shorter than the inter-zonal trips between regions, both in length and duration (see Greenwald, 2006; Venigalla et al., 1999). Ignoring the intra-zonal flows and considering only the longer flows in the model is not a true representation of reality and can potentially result in a biased dataset and parameter estimates (Bharat and Larsen, 2011).

Although the consideration of intra-zonal flows in the model seems to be necessary, often in practice the average trip length for intra-zonal flows cannot be measured or is difficult to measure. There are a number of approaches in the literature for this purpose. For instance, sometimes a different formulation is used to calculate the intra-

zonal trip length or a modification is applied on the model itself to avoid the intra-zonal flows. LeSage and Pace (2008), for instance, created a separate model with some extra explanatory variables for flows from the main diagonal of the flow matrix to explain intra-state migration flows in the United States. LeSage and Fischer (2010) argue that practitioners often see the intra-zonal flows as a nuisance and consider dummy variables for those observations (see for example Behrens et al., 2010). There are also a number of algebraic models in the literature for estimating intra-zonal distances. One of the earliest methods was introduced by the U.S. Department of Commerce (1965) in which the intra-zonal driving time of a particular zone A is estimated as one-half of the average driving time between the centroid of zone A and the centroids of all neighbouring zones (see Ghareib, 1996). Another example is a similar method suggested by Venigalla et al. (1999) where in a first step, the nearest zone centroid to the centroid of zone A was determined and the intra-zonal trip length for zone A was then computed as half the travel length to the nearest centroid.

Besides the above highly approximate approaches, there have been a number of attempts in the literature for measuring the intra-zonal distances for zones of varying shapes and under various assumptions. The assumptions are mainly related to the geometrical shape of the zone and the internal population distribution. Both Batty (1976) and Fotheringham (1988b) suggested zoning based models for deriving the average trip length for zones that are approximately circular. These models are investigated in more detail in following section.

4.2.1 Circular-shape distance estimates

The intra-zonal distance problem can be seen as a problem in finding the mean trip length within any zone (Batty, 1976). Batty (1976) suggested some assumption-based measures for the mean trip length where one of the simplest is based on the assumptions that the zone is circular in shape and the population is distributed evenly at a constant density throughout the zone. An approximation to the intra-zonal distance then can be found by:

$$d = \frac{r}{\sqrt{2}} \tag{4.1}$$

where r is the radius of the zone in terms of trip length (travel cost). He also suggested a slight variation in this measure when it is assumed that the population density varies in a regular way in the zone which can be modelled by a tractable mathematical function, fitted to a particular zone (see also Batty, 1974).

Another model based on the same above assumptions (i.e. circular-based shape and uniform population distribution assumptions) is suggested by Fotheringham (1988b)

where the intra-zonal trip length is estimated with the following formulation:

$$d = 0.846 \cdot (\frac{1.432}{0.846})^{z/r} \cdot r \tag{4.2}$$

where z is the distance between the zone centroid and the destination point and r is the radius of a circle whose area is equal to that of the zone. The coefficients in this formulation are related to the potential minimum ($0.846.r$) and maximum ($1.423.r$) distances in a circular zone (see Eilon et al., 1971; Fotheringham and O'Kelly, 1989, for further details). When the point is located on the circumference of the zone then $z = r$ and $d = 1.432 \cdot r$; when the point is located between the centre and the circumference then $0.846 \cdot r < d < 1.432 \cdot r$; and in the special case of intra-zonal flows when the origin and destination points are both located at the zone centroid, with z being equal to zero, equation 4.2 will simplify to: $d = 0.846 \cdot r$, where the coefficient of the equation is very close to one of the equation 4.1. Multiplication of such similar coefficients in a distance metric should not make a large difference and so both formulations could produce similar results when distances are calculated for real-world practical problems.

4.3 Scattered intra-zonal distance estimates

The circular-shape based estimates in section 4.2.1 are only a rough approximation and they are more of analytical interest than practical (Batty, 1976). Also the assumptions of the models are rarely all met in practice. The polygons representing the origins and destinations are often not circular and population is unevenly distributed within a zone. Additionally, assuming that the population is located at a zone centroid and ignoring intra-zonal flows from the model (as in a centroid-to-centroid calculation) can still potentially result in poor or biased parameter estimates. To overcome these problems, we suggest a method for estimating the average trip lengths by scattering the origins and destinations of the flows within their respective zones.

In spatial interaction models the precise locations of the origins and destinations of the flows usually depends on where people are starting and terminating their trips, respectively. Scattering the origins and destinations would be a better approximation to what is happening in reality compared with assuming that all origins and destinations are located at the centroid of each zone. The distribution of the origins and destinations of the flows can be done randomly or based on an available spatial density distribution. In both random and density-based models, the average trip length is calculated for all possible trip configurations between scattered origins and destinations. In the case of intra-zonal flows, the scattered origins and destinations are obviously within one zone and the average trip length would be calculated between all possible flow configurations between them.

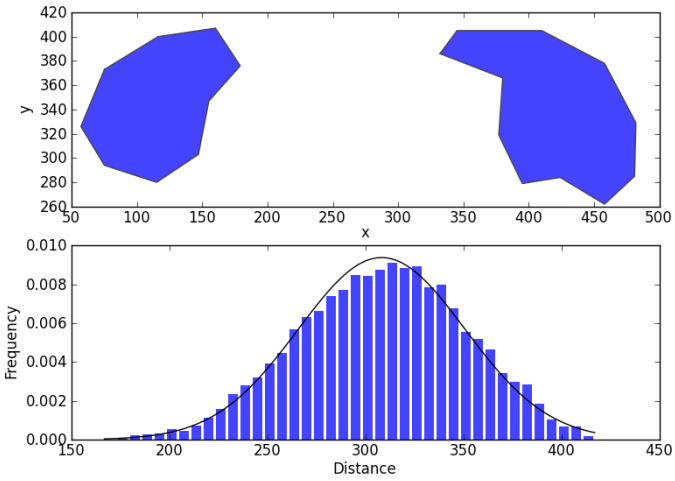
4.3.1 Randomly scattered distance estimates

Ideally the average trip length should be computed using the disaggregated data which give the trip information for all commuters. Generally these data are not available in the model and the average spatial separation (distance, time, cost) needs to be approximated. In spatial interaction models usually the population size in each zone and the number of trips between origin and destination zones are known. However, these data are aggregated and do not provide any information about the location of the population within each zone, nor the origin and destination of the flows. In this case, for estimating the average trip length we can calculate the average geometric distance between two polygons representing the zones. Geometrically speaking, the surface of each polygon covers an infinite number of points. Calculating the average distance between two polygons includes all possible distance configurations between the infinite set of points within both polygons; which is also infinite. The same situation is valid for intra-zonal distance when the average distance should be calculated between the infinite number of points within a zone. By computing the average of all possible distance configurations between two polygons we make the assumption that the trip lengths are a random sample of all possible distances and the sample mean tends towards the population mean with a probability that can be estimated using standard statistical methods.

To overcome the problem of finding the average distance between an infinite number of points, we can randomly select a number of points from both origin and destination polygons to build up a sample of flows. For intra-zonal trips, two random sets of points should be selected within the polygon. The average trip length can then be estimated by calculating the average geometric distance of the sample points. The law of large numbers indicates that as the number of experiments (samples) on a random process increases, the average of all trials will tend towards the expected value. In the case of selecting random flows between polygons (i.e. random distances between points) if the number of sample distances is large enough, the calculated sample mean (i.e. the average of the random distances) is likely to be close to the real value. It should be noted that in this case no information is available about the exact number and location of the individual points, we shall assume these points to be distributed uniformly within the zones. As a consequence, the density of origins and destinations is assumed uniform within each polygon.

Figure 4.1 illustrates an example of estimating the distance between two arbitrary polygons with roughly regular shapes and centrally located centroids. Considering the boundaries of the polygons, a set of $10,000$ flows is selected randomly from the different distance configurations between the points located in each polygon. The average distance of these randomly selected flows is calculated and equals 307.8. The results and information of the experiment are shown in table 4.1. The reliability of the estimated average distance can be investigated using statistical measures such as a confidence interval (CI). The confidence interval for the mean provides an estimated range of values using the

Figure 4.1: Example of average polygon distance.



sample data in which the unknown population parameter, (i.e. in this case the average mean distance between polygons), lies with a given probability. The total population, in our case all the flows between polygons, and its standard deviation are not known. In such a case, the sample mean follows a t-distribution with '*sample size* − 1' degrees of freedom. With increasing sample size, the t-distribution becomes closer to a normal distribution. The confidence intervals of 95% and 99% are also calculated which indicate the probability that the estimated average distance for the polygons lies in the estimated range of values.

Table 4.1: Estimating the distance between two polygons

Centroid for polygon 1 (x, y):	(116.0, 345.3)
Width / height for polygon 1:	122.0 / 127.0
Centroid for polygon 2 (x, y):	(420.3, 340.1)
Width / height for polygon 2:	150.0 / 143.0
Distance between centroids:	304.3
Sample size:	10,000
Average distance between polygons:	307.8
Median distance between polygons:	309.0
Standard deviation for distance between polygons:	42.7
Confidence interval 95%:	[224.03, 391.62]
Confidence interval 99%:	[197.69, 417.96]

Due to the random nature of the sample selection, when the sample size is not large enough, the resulting average distance estimate might vary slightly in different experiments. To check this, the experiment illustrated in figure 4.1 has been repeated with a different set of $10,000$ random samples. The mean distance resulted with new experiment is equal to 308.2, just slightly different from the mean of the previous sample presented

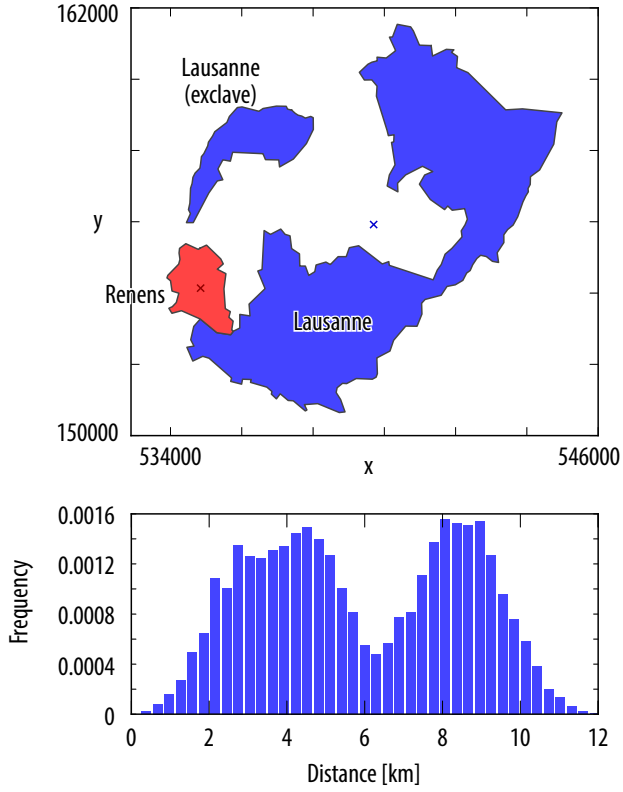
in table 4.1. The close results of these two random trials can be evidence that the sample size of 10,000 in our experiments is big enough considering the law of large numbers. The standard deviation is a measure of the dispersion from the mean of the sampled distances, so the smaller the standard deviation, the more similar are the distances to the mean value. The standard deviation of the first and second trials are also very close to each other, 42.6 compared to 42.7 respectively. The estimated distances and standard deviation for the distribution of distances between random polygon points (see figure 4.1 bottom) is presumably related to the size, shape and orientation of the involved polygons.

The resultant average trip length from the randomly scattered method shown in table 4.1 (and the one of second trial) for the polygons illustrated in the example 4.1 is close to the centroid-to-centroid distance which is equal to 304.3. This is due to the fact that the scattered-based method considers the actual shape of origin/destination zones when estimating average trip lengths. Therefore, if origin/destination zones follow a roughly regular polygon shape especially with centroids located in the middle of polygons (as in the above example), the distance between the centroid of the origin/destination zones (i.e. centroid-to-centroid distance) will be a close representation of the trip length estimated with a scattered-based method. However, in real-world problems the polygons representing the zones are not necessarily very regular. Sometimes the polygon shape is such that the centroid of the zone is located out of the polygon. For instance in figure 4.2, the two blue polygons represent the city of Lausanne. The shape is irregular and divided in two polygons, the zone centroid is out of the polygon (represented with a blue $\times$ in figure 4.2). The roughly regular polygon in red shows the zone polygon of the adjacent city of Renens.

In such cases, as stated earlier, especially when the population is not distributed evenly over space, the centroid-to-centroid distance is not a good approximation of the average trip length between two polygons. Additionally, the centroid-to-centroid distance method fails to address the average distance for intra-zonal flows as the distance between a centroid to itself is zero. To show an example, a similar experience of calculating the average distance between two polygons for 10,000 randomly distributed flows as in figure 4.1 is calculated for Lausanne and Renens (see figure 4.2). The result of the experiment in table 4.2 indicates a considerable difference between the centroid-to-centroid and polygon distances, 5151.9 compared to 5907.0. As stated before, the frequency distribution of the random distances in figure 4.2 (bottom) depends on the shape and the regularity of the polygons, as it is determined by the distances of all possible combinations of points between the polygons. If the polygons are regular, as in figure 4.1, the frequency distribution will be close to Gaussian, with the mean being close to the centroid-to-centroid distance.

The variance (or standard deviation) of the distribution is also influenced by the size of the two polygons in which bigger polygons giving a higher variance. Estimating the average distance between two polygons using a sample of 10,000 flows is computationally

Figure 4.2: Average polygon distance between Lausanne and Renens.



tractable. However, as the number of polygons increases, the number of computations quickly becomes very large. Therefore, we need to find an efficient way to utilise the method described above for approximating the average trip length d_{ij} between polygons and d_{ii} in the case of intra-zonal distance. Instead of generating a random sample of flows, we can use a fine regular grid within the zones. Given that the density of origins and destinations is uniform within each polygon, we can describe the space as a discrete grid with small square units, as long as the grid is fine enough. For both zones i and j we select the set of grid points G_i and G_j lying within zone i and j respectively. The average distance d_{ij} can be then estimated by computing the average distance between all possible pairs of points from G_i to G_j. If the units are small enough, the estimated discrete average distance between centres of the grid units will tend towards the average distance in continuous space between polygons. Figure 4.3 illustrates this approach, the grid in figure 4.3 is very coarse and serves as illustration only.

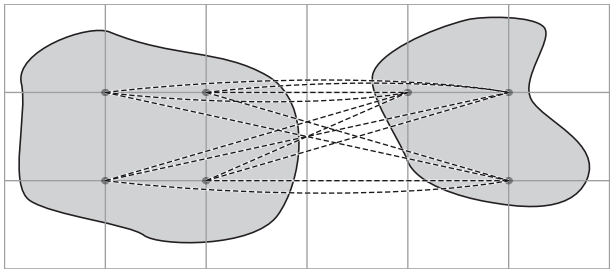
4.3.2 Density-based scattering approach

The randomly scattered approach provides a viable alternative to ignoring zero distance intra-zonal flows, where the populations are assumed to be located at the centre of the zones. However, in the randomly scattered approach the origin and destination points are assumed to be uniformly distributed over space which might not reflect reality in

Table 4.2: Estimating the distance between two polygons (Lausanne and Renens).

Centroid for polygon 1 (x, y):	(539700.2, 155921.6)
Width / height for polygon 1:	10538.0 / 10893.0
Centroid for polygon 2 (x, y):	(534864.6, 154144.2)
Width / height for polygon 2:	1865.0 / 2555.0
Distance between centroids:	5151.9
Sample size:	10,000
Average distance between polygons:	5907.0
Median distance between polygons:	5603.6
Standard deviation for distance between polygons:	2627.1

Figure 4.3: Estimating the average trip length using a regular grid.



most cases. In a uniform distribution, all random locations are equally spaced and equally probable in space. In the case of journey-to-work flows, we would expect the probability of a flow origin at a given location i to be proportional to the density of the active population at zone i and similarly for the destination point with the density of jobs. Sometimes a spatial density surface containing information such as population is available. If such a density surface is available it is possible to choose the locations randomly according to the given probability density surface instead of selecting them from a uniform density function. For instance in the case of journey-to-work, this probability density surface can be approximated via a fine-grained population density surface. If a high resolution population density is not available, it is possible to disaggregate the population data using some kind of areal interpolation to make a fine-grained resolution density surface based on smaller scale data (see e.g. Tobler, 1979; Kyriakidis, 2004; Liu et al., 2008; Pozdnoukhov and Kaiser, 2011).

Similarly to section 4.3.1, the average trip length for the density-based scattering approach can be approximated using a fine regular grid. In order to take into account the probability density surface, a weighted average distance can be computed between all possible pairs of grid points. The considered weight is $w_{ij} = w_i \cdot w_j$, where w_i is the value of the density surface at origin i and w_j is the density surface value at destination j. The density-based average trip length is then computed by:

$$\hat{d}_{ij} = \frac{\sum w_{ij} \cdot d(g_i, g_j)}{\sum w_{ij}} \tag{4.3}$$

where $d(g_i, g_j)$ is the distance between a grid point in zone i and another grid point in zone j (with potentially $i = j$). The sum is taken over all possible pairs of grid points inside zone i and j respectively.

4.4 Application & Results

In following section, we apply different methods presented in sections 4.3.2 and 4.2.1 for calculating the average trip length on the journey-to-work dataset in the Lausanne agglomeration and compare their results. The same global Poisson gravity model presented in section 3.8 is considered here with similar variables and parameters to be estimated with MLE. For the density-based scattering model (section 4.3.2), the population data for Lausanne are available as a regular grid with a spatial resolution of 100 metres (hectare-level population data)[1]. To make the comparison of the methods possible, different variations of treating the intra-zonal flows and distance measures are computed using the following schemes:

1. Excluding the intra-zonal flows and taking the traditional centroid-to-centroid distance between the zones for the inter-zonal flows.

2. Using all flows with the inter-zonal distances again computed using the centroids and the intra-zonal distances are estimated using the circular-shape based model (equation 4.1).

3. Using all flows with both the intra- and inter-zonal distances computed using the randomly scattering model (section 4.3.1).

4. Using all flows with both the intra- and inter-zonal distances computed using the density-based scattering model (section 4.3.2) with the hectare-level population data as additional information to form a density surface.

The parameters of the calibrated model with different average trip lengths are shown in table 4.3 along with some measures of model goodness-of-fit. Parameter α is related to the active population variable and shows a positive influence on the interaction; by increasing the size of the active origin population, the total trip number of interactions from that origin will increase. Parameter γ is associated with the destination attractiveness variable (in our case the number of jobs in each commune) and also has a positive effect on the interaction and β is the distance-decay parameter with a negative influence on the number of interactions.

In our journey-to-work model over 45% of the flows are intra-zonal flows. Hence, ignoring these data, the centroid-to-centroid model considers only half the flows compared with the other models. The resulting model is suitable for inter-zonal flows only, while the

[1] See `http://www.geostat.admin.ch` for more information

Table 4.3: Comparison of different methods for calculating average trip length (distance). The results show the output from calibrating a spatial interaction model with the different distance estimates.

Distance measure model:	Centroid-to-centroid	Circular-shape based	Randomly scattering	Density-based scattering
AIC	21381.82	61360.14	58881.04	41985.05
Pseudo R^2	0.945	0.936	0.938	0.956
SEE	19.35	138.75	144.08	87.69
parameter α	0.846	0.862	0.945	0.791
standard error	0.0027	0.0018	0.0018	0.0020
t-value	316.368	469.904	515.631	414.826
p-value	0.0	0.0	0.0	0.0
parameter γ	1.011	0.969	1.0436	0.949
standard error	0.0022	0.0017	0.00180	0.0018
t-value	450.63	561.289	582.680	534.296
p-value	0.0	0.0	0.0	0.0
parameter β	-0.668	-1.017	-1.3170	-1.297
standard error	0.0051	0.0031	0.00426	0.0039
t-value	-130.589	-325.716	-309.309	-334.594
p-value	0.0	0.0	0.0	0.0
parameter k	-5.70	-2.274	-0.881	0.818
standard error	0.0601	0.0335	0.0383	0.038
t-value	-94.815	-67.876	-23.0196	21.276
p-value	0.0	0.0	0.0	0.0

other models are more general interaction models. A direct comparison of the goodness-of-fit between the centroid-to-centroid model and the other models is not really possible, as the data are not the same. However, a rough comparison of parameter estimates between the four methods shows that the values of α and γ are relatively similar for all the models. The distance-decay parameter β shows a larger variation especially between the centroid-to-centroid model to the other models. The distance-decay parameter in the centroid-to-centroid method is -0.668 while in the other three methods it is smaller than -1. This shows that ignoring the intra-zonal flows yields a distance-decay parameter that is substantially different from when all the flows are considered. The model without intra-zonal flows overestimates people's willingness to accept longer distances for their journey-to-work. The circular-shape based distance calculation and the scattered-based methods, considering all intra-zonal flows, show a stronger distance-decay effect. This is an indication that people view distance as an important criterion for their daily travel-to-work. The intercept parameter k is a balancing factor and not of interest in the behavioural interpretation in the model. For all parameters in different models, the t-values are considerably different from zero and p-values are zero indicating that all parameters are significance in the models.

Other comparisons between the different models can be performed based on statistical measures of goodness-of-fit. The AIC and McFadden's Pseudo R^2 are listed in table 4.3 for all distance measure variants. Although the smallest AIC value is for the model

with the centroid-to-centroid distance, this model has a smaller dataset and so a direct comparison to the models with intra-zonal flows is not appropriate. Between other three models considering the inter- and intra-zonal flows, the density-based scattering model with AIC of 41985.05 shows the best goodness-of-fit. The comparison of the McFadden's Pseudo R^2 is possible only for different models applied on the same data in which a model with larger McFadden's Pseudo R^2 has a better fit compared with models with lower corresponding values. The comparison between models applied on all data (i.e. both inter- and intra-zonal flows) shows that again the model with density-based scattering measure shows a better fit in comparison with the other two models of circular-shape based and randomly scattering.

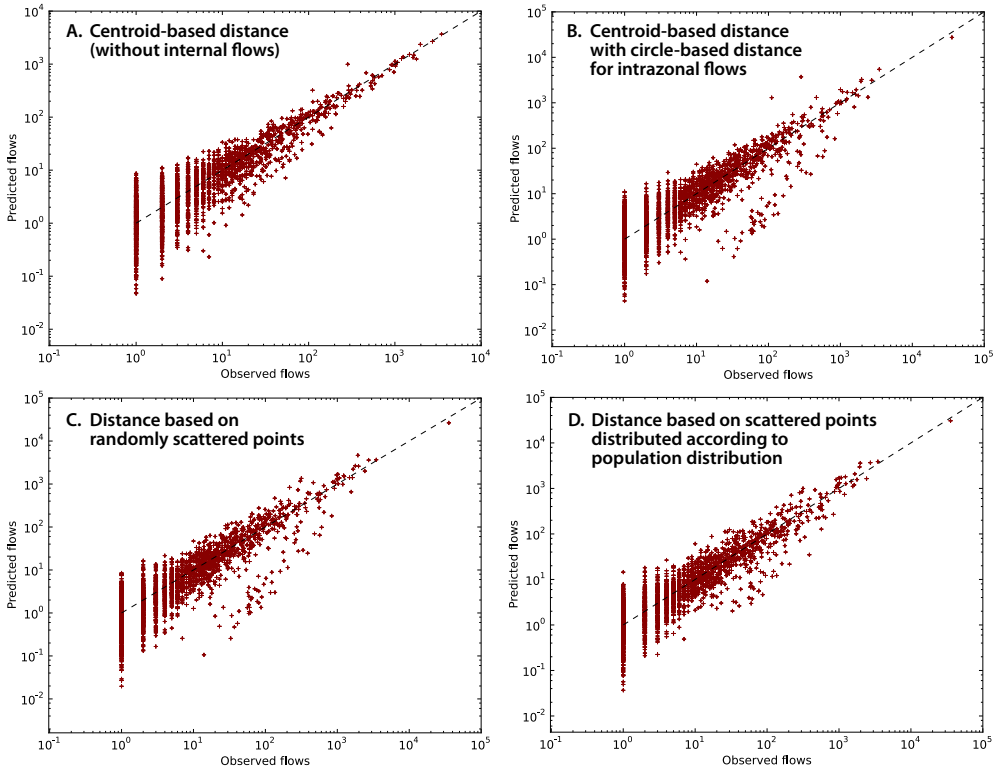
Estimating the average trip distance using the population density seems to enhance the spatial interaction model. Between the methods taking into account intra-zonal flows, the standard error of the estimate (SEE) is higher for the randomly scattering model compared to the circular-shape based model but it is considerably smaller for the density-based scattering model. Figure 4.4 shows the predicted flows versus observed flows for all four model variants. The best fit among the methods considering the intra-zonal flows is given by the density-based scattering model. Considering all the different results and comparison of the models, it can be concluded that the density-based scattering method is a viable alternative to existing methods for calculating the average trip length and for considering the intra-zonal flows in the analysis.

4.5 Summary

In this chapter a method for estimating the average distance for intra-zonal flows is presented. The origins and destinations of the flows in this method are distributed randomly or based on an available density surface within the origin and destination zones, named randomly- and density-based scattering methods respectively. The scattered-based methods provide the average trip-length considering both inter- and intra-zonal flows while the intra-zonal flows were ignored by methods that estimate the average trip length between centroids of the origin and destination zones. The scattered-based methods are not based on any pre-defined assumptions or conditions about the zone shapes. Also this method for estimating the average trip length considers as much information as possible when formulating the model by potentially using available density information such as population density.

The methods presented in section 4.3 are applied on the the Lausanne journey-to-work data where the scattered-based methods are used to estimate average distances between the communes, considering both inter- and intra-zonal flows. The scattered-based method results are compared with existing circular-shape based methods where it is shown that the density-based scattering methods provides a better fit with less error. The scattered-based method can be applied to many different types of spatial interaction

Figure 4.4: Predicted flows vs. observed flows for different distance measures.



models for estimating the average trip length and it could be easily modified to consider travel distance or time instead of Euclidean distance between the set of points of the origin an destination zones.

In following chapter a brief overview of the existing local methods for spatial data analysis will be provided along with existing techniques for local analysis of spatial interaction processes such as origin- and destination-specific models.

Chapter 5

Local spatial analysis

5.1 Introduction

During the last decades there has been a powerful movement within spatial analysis, termed local modelling or local analysis (Fotheringham and Brunsdon, 1999; Fotheringham, 1999b; Openshaw et al., 1987; Getis and Ord, 1992; Anselin and Getis, 1992; Fotheringham and Rogerson, 1993; Lloyd, 2011, 2006) that involves spatial disaggregation of the conventional global approaches (Fotheringham, 1992, 1997; Openshaw, 1993; Fotheringham and Brunsdon, 1999; Openshaw et al., 1987). Local statistics and local models mainly emphasise differences over space rather than similarities and focus on identifying spatial variations in relationships (Fotheringham, 1999b; Fotheringham et al., 2000), while global models assume spatial stationarity of relationships (Lloyd, 2011). In a global model, one set of results representing an average type of relationship is generated providing general information about the entire study region (Fotheringham and Brunsdon, 1999). However, in situations where relationships vary across space, the average set of global results may fail to represent the actual situation in any part of the study area (Fotheringham and Brunsdon, 1999; Fotheringham et al., 2002; Unwin, 1996a,b; Fotheringham, 1997; Boots and Okabe, 2007). In fact in many real-life situations, spatial variation in relationships exists and the assumption of global stationarity or structural stability over space might be highly unrealistic (see Fotheringham et al., 1996; Fotheringham, 1997; Brunsdon et al., 1996; Leung et al., 2000a). Sometimes in the literature, such variations in relationships are referred to as spatial heterogeneity or spatial drift (Charlton et al., 1997).

Despite the importance of spatial heterogeneity in different research areas, a unified definition for it is lacking (Li and Reynolds, 1994). Kolasa and Rollo (1991) have shown that spatial heterogeneity can be defined in different ways and the lack of a unified definition of the concept may be due to the complexity of the phenomena involved. For instance, Li and Reynolds (1994) define spatial heterogeneity as the complexity and variability of a system property in space and LeSage (1999) refers to spatial heterogeneity as variation in relationships over space. So, briefly, spatial heterogeneity can be defined as

the systematic variation of a process by location (spatial non-stationarity). In the case of a model, spatial heterogeneity can be expressed by variation of the model's parameters over space. In a heterogeneous model, the parameters are allowed to vary over space while a stationary model has the same parameters at all locations (Lloyd, 2011).

Fotheringham et al. (2002, 2000); Fotheringham (1999b, 1997) have discussed three different possible reasons why we might expect variation of relationships over space. An obvious reason could be due to sampling variation, i.e. we are likely to generate different parameter estimates resulting from the calibration of the model with different spatial subsets of a dataset. A second possible cause of spatial heterogeneity is that the model is a gross misspecification of reality; for instance, when one or more relevant variables are represented by incorrect functional form or are omitted from the model. The third and in our study probably most important cause of possible spatial heterogeneity is that, for whatever reasons, some relationships vary across space. For instance, in a regression model of housing price, the value of an extra bedroom can vary from place to place; e.g. in areas around a good school, the utility of an extra room is high since in such areas there may be more demand by families with several children (Brunsdon et al., 1999b). This variation can be seen in other fields of study, for instance when modelling the spatial distribution of a certain illness based on social and economical criteria (Fotheringham et al., 1996). Another example is in social processes when the variation of a relationship might depend in part on where the measurement is taken, since the perception of opportunities might vary in space from one individual to another. The individual perception of opportunities and how people evaluate different alternatives to choose between them, is still something of a mystery and relies on an individual's cognition of space. Fotheringham et al. (2002) have discussed that perhaps the reason people have different responses to the same stimuli over space is due to different administrative, political or other contextual issues, or perhaps there are spatial variations in people's attitudes or preferences. For example in the case of a shopping behaviour modelling, people in small cities might accept the need to travel longer distances for clothes shopping than do people in big cities.

The same variation in relationship which represents heterogeneity can be seen in journey-to-work modelling where people working in big cities may accept living farther away and commuting long distances everyday while people working in smaller cities might be less receptive to commuting long distances. This variation in people's behaviour in space represents spatial heterogeneity. There are some studies in contextual effects on variation of people's behaviour over space, for example Cox (1969); Agnew (1996); Pattie and Johnston (2000) in voting behaviour and Yano et al. (2003) in migration. Fotheringham and Brunsdon (1999); Fotheringham et al. (2002) stated that the idea of intrinsic variation of human behaviour over space is consistent with the post-modernist beliefs of the importance of place and locality as frames for understanding such behaviour (see Thrift, 1983), so identification of local variation in relationships can be a useful procedure

to more intensive studies of such differences (Fotheringham, 2000; Fotheringham et al., 1997b).

In order to examine local variations of relationships over space, local statistics and local form of spatial models provide local values for specific spatial subregions that are defined with respect to the complete dataset (Boots and Okabe, 2007). These local values can then be mapped over space to examine and visualise possible significant spatial variations in relationships. Analysing these spatial variations can help us to better understand spatial processes by improving our knowledge of the system under investigation (Fotheringham et al., 2002; Fotheringham, 2000; Fotheringham and Brunsdon, 1999; Fotheringham, 1997). Fotheringham and Brunsdon (1999) stated that "*...an examination of the nature of the spatial variation can suggest to us a more accurate model specification or the nature of some intrinsic variation in spatial behaviour*". Over the last several years, increasing attention has been paid to local models for the analysis of spatial data. However, the earliest attempts for that predate this recent interest (Fotheringham and Brunsdon, 1999; Fotheringham, 1997, 1992, 1984a, 1981; Greenwood and Sweetland, 1972; Casetti, 1972; Monmonier, 1969; Linneman, 1966). There are a number of publications in the literature that attempt to provide a review of local forms of spatial modelling and analysis, for example see Lloyd (2011, 2006); Fotheringham et al. (2002, 2000); Atkinson (2001); Fotheringham and Brunsdon (1999); Unwin and Unwin (1998); Fotheringham (1997); Fotheringham and Charlton (1994).

In the following section we briefly review a sample number of these local models with a greater attention given to a few of the models which will be used in this thesis (e.g. GWR, origin- and destination-specific spatial interaction models).

5.2 Overview of local methods for spatial data analysis

Local models can provide an important link between spatial analysis and the powerful visual display environments of various GIS and statistical graphics packages where local values can be visualised on maps (Fotheringham, 1993, 2000; Fotheringham et al., 2002). The process of information analysis is made easier and quicker, and more spatially disaggregated spatial statistics can be developed (Fotheringham and Charlton, 1994). Many approaches for dealing with local information have been developed. Among the most important arguably are the following:

- **Local point pattern techniques**, such as the Geographic Analysis Machine (GAM) developed by Openshaw et al. (1987) trying to find evidence for local clustering, and further developed e.g. by Openshaw and Craft (1991) and Fotheringham and Zhan (1996). Another set of methods for point pattern analysis are the Spatial Scan Statistics by Kulldorf (1997), testing if a point process is random and aiming for finding local clusters in space and potentially in time.

- **Local spatial autocorrelation** is a descriptive univariate statistic aiming to as-

sess degree of spatial dependency or spatial association among observations. Spatial autocorrelation exists when nearby locations show similar values , i.e. neighbouring observations have dependent values (LeSage and Pace, 2009). Spatial autocorrelation may occur only at some locations and not in others, requiring local indicators to detect its presence. Among these local methods are the Local Indicators of Spatial Association (LISA) described by Anselin (1995). These indicators can be for example a local version of Moran's I (Moran, 1950) or Geary's C (Geary, 1954) statistics. Using these statistics for spatial autocorrelation requires measuring a spatial weight matrix whose elements describe the subjective concept of spatial proximity, such as cardinal distance (e.g. km) between neighbours or ordinal distance (e.g. k nearest neighbours). Some examples of local spatial autocorrelation include (Getis and Ord, 1992; Ding and Fotheringham, 1992; Ord and Getis, 1995, 2001; Brunsdon et al., 1998b; Rogerson, 1999; Rosenberg, 2000; Leung et al., 2000b; Getis and Griffith, 2002; Getis and Aldstadt, 2004; Bavaud, 2008).

- **Local regression techniques** are required when relationships between variables vary over space. Many different approaches can be found in the literature, for example applications of the expansion method to geographical models, allowing the linking of geographical locations to spatially varying phenomena (Casetti, 1972, 1997; Jones and Casetti, 1992; Brown and Kodras, 1987; Fotheringham and Pitts, 1995). The presence of spatial autocorrelation is typically a violation of the underlying assumptions (i.e. independence assumption) of linear regression. The spatial lag model tries to address this issue by incorporating such spatial dependencies as an additional predictor into the model, giving a spatially autoregressive linear regression (Anselin, 1988, 1999, 2001a,b, 2009, 2010). Other variants to this approach exist, such as the spatial lag and spatial error model (Anselin, 1988). Although the spatial lag and spatial error models consider spatial autocorrelation using local information, they still yield a set of global parameter estimates, so in this sense Fotheringham et al. (2002) categorised them as *semi-local* rather than fully local models. Another approach to local regression is Geographically Weighted Regression (GWR) which will be discussed in detail in section 5.3.

- **Local spatial interaction models** try to capture spatial variations in the underlying model parameters in order to investigate differences in interaction behaviour. Spatial disaggregation of spatial interaction models can be achieved by separate calibration of the model for each specific origin or destination in the system. These local model are called origin- and destination-specific spatial interaction models (see e.g. Fotheringham and O'Kelly, 1989), which will be discussed in detail in section 5.4. Further disaggregation has been done by Nakaya (2001, 2003) by combining these location-specific models with the GWR approach. This is also discussed in more detail in section 5.4.3.

- **Local network autocorrelation for spatial flows** is similar to spatial autocorrelation applied on geographic data, but considers dependencies among spatial flows. Network autocorrelation has been first described by Black (1992) who examined a global network autocorrelation using Moran's I statistic. Some attempts in the literature exist for incorporating network autocorrelation in spatial interaction model (see Chun et al., 2012, for a brief review of the recent network autocorrelation research). For instance, Berglund and Karlström (1999) applied Getis-Ord statistics (see e.g. Getis and Ord, 1992; Ord and Getis, 1995) to origin and destination pairs to measure the spatial association in residuals from flow models. Also, network autocorrelation can be integrated into spatial interaction models in a similar way to spatial autocorrelation in spatial lag models, leading to an autoregressive regression variant (see e.g. Griffith, 2007; Chun, 2008; Fischer and Griffith, 2008; Griffith, 2009; LeSage and Pace, 2008, 2009; Chun and Griffith, 2011; Chun et al., 2012). In this approach, spatial lag vectors are formed based on spatial weight matrices that are defined based on proximity of origin and destination regions to their neighbouring zones (see LeSage and Fischer, 2010). Network autocorrelation techniques, similar to spatial lag and spatial error models, incorporate local relationships into their modelling framework but still these models have to be considered as semi-local rather than fully local techniques (see Fotheringham et al., 2002, p. 22), since they yield a set of global parameter estimates. As such, these types of models are outside the remit of this thesis and will not be considered further.

5.3 Geographically Weighted Regression (GWR)

Geographically weighted regression (GWR) (Brunsdon et al., 1996; Fotheringham et al., 1998; Brunsdon et al., 1998a, 1999a; Fotheringham, 1999a; Fotheringham et al., 2002) is a local linear regression technique for analysing spatially varying relationships. In GWR, relationships are allowed to vary spatially and a set of local parameter estimates are produced for each location. These local parameter estimates and their related local statistics can then be mapped over space for further spatial analysis. The *geographically weighted* concept in GWR denotes that data are weighted according to their proximity to a calibration point, in which data in closer proximity carry more weight and have more influence in the parameter estimation. So in GWR, the regression coefficients are location dependent; this extended local version of regression model can be modelled as:

$$\begin{aligned} y_i &= \beta_{0i} + \beta_{1i}\, x_{1i} + \beta_{2i}\, x_{2i} + \ \ldots \ + \beta_{ki}\, x_{ki} + \epsilon_i \\ &= \beta_{0i} + \sum_{k=1}^{p-1} \beta_{ki}\, x_{ki} + \epsilon_i \\ &= \sum_{k=0}^{p} \beta_{ki}\, x_{ki} + \epsilon_i, \qquad when \quad x_{0i} = 1 \end{aligned} \tag{5.1}$$

where y_i is the dependent variable at location i; x_{ik} is the kth independent variable at location i; β_{ki} shows β_k parameters of the model at location i, and ϵ_i is an error term that should be minimised.

The estimation of the parameters of the model then is based on *geographically weighted least squares*, by minimising $\sum_i w_i \ (y_i - y_i')^2$ where w_i is a weighted function applied to each squared difference between the observed y_i and its predicted value y_i' (Brunsdon et al., 1996; Fotheringham et al., 2001a). Rewriting the GWR model expressed in equation 5.1 in a matrix notation gives

$$y_i = \beta_i \ X_i + \epsilon_i \tag{5.2}$$

where β_i is a column vector of the model's parameters matrix; X_i is a row vector of independent variables matrix at location i (Fotheringham, 2009; Wheeler and Páez, 2010). The local parameters of the model at i, β_i', can be estimated by the following matrix notation of weighted least squares:

$$\beta_i' = (X^T \ W_i \ X)^{-1} \ X^T \ W_i \ Y \tag{5.3}$$

where X is the matrix of independent variables with a first column of one values for the intercept variable , X^T denotes the transpose of X, Y is an n by 1 vector of dependent variables: $Y = (y_1, y_2, \ldots, y_n)^T$ and W_i is an n by n matrix of weights whose off-diagonal elements are zero and whose diagonal elements are the geographical weighting of each of the n observed data based on their proximity to the calibration point i (Fotheringham et al., 2002; Fotheringham, 1999a, p. 53). W_i has the form of the following matrix:

$$W_i = \begin{bmatrix} w_{i1} & 0 & \ldots & 0 \\ 0 & w_{i2} & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & w_{in} \end{bmatrix}$$

where w_{in} denotes the weight of the data point n on the calibration of the model around calibration point i (Fotheringham et al., 2000, p.94).

5.3.1 Spatial weighting function

The choice of both weighting function, or kernel, and the bandwidth are two major considerations in the GWR procedure, although the bandwidth selection has a more significant influence on the results (Lloyd and Shuttleworth, 2005; Fotheringham et al., 2002, 1997a). Two major categories of weighting methods exist: one uses a fixed bandwidth and one uses an adaptive bandwidth. With a fixed spatial weighting function, the same bandwidth is applied to each calibration point and so it is assumed that this bandwidth is constant over the study area. Fotheringham et al. (2002, p. 210) refer to this as a fixed kernel where each observation has a weight according to how far it is located from the

centre of the kernel. Observations closer to the centre of the kernel have higher weights while the weights decrease when the observations are located further away. One example of a fixed kernel is the Gaussian kernel which can be written as:

$$w_{ij} = exp\ [\ -\frac{1}{2}\ (\frac{d_{ij}}{b})^2\ ] \tag{5.4}$$

where d_{ij} is the geographical distance, mostly Euclidean distance in GWR, between the locations of calibration point i and observation j. Parameter b, called *bandwidth*, has to be optimally specified using an appropriate technique. Another example of the fixed kernel is the bi-square function:

$$w_{ij} = \begin{cases} [1 - (d_{ij}/b)^2]^2 & \text{if } d_{ij} < b \\ 0 & \text{otherwise} \end{cases} \tag{5.5}$$

which produces a continuous weighting function up to distance b from the regression point and then zero weights to any data point behind b (Fotheringham et al., 2002, p. 57). Examples of using these types of fixed spatial kernels are provided by Brunsdon et al. (1996, 1997); Fotheringham et al. (1998, 2002).

Another category of kernels is when the spatial kernel is not fixed and can adapt its size to the density of data points over space. The size of bandwidth is bigger when the data are sparse and in areas where the data are plentiful, bandwidth size is smaller. This type of spatial weighting function is referred to as an *adaptive spatial kernel* (Fotheringham et al., 2002, p. 46). Adaptive kernels are particularly useful when the density of the observations shows a large spatial variation (Nakaya, 2007). Again, there are different types of spatially adaptive weighting functions for GWR which can be found in the literature (see e.g. Wheeler and Páez, 2010; Páez and Wheeler, 2009; Fotheringham et al., 2002, 2000). For instance, the bi-square function is the type commonly used in GWR for this purpose:

$$w_{ij} = \begin{cases} [1 - (\frac{d_{ij}}{b})^2\ ]^2 & \text{if j is one of the Nth nearest neighbours of i} \\ 0 & \text{otherwise} \end{cases} \tag{5.6}$$

where b is the bandwidth indicating the greatest distance between the regression point and the Nth nearest neighbours, and N is a parameter to be estimated (Fotheringham, 2009) which indicates the number of observations within each kernel.

5.3.2 Calibration of the spatial weighting function

As previously mentioned, the results of GWR are sensitive to the degree of distance-decay so the bandwidth selection is an important consideration in GWR (Fotheringham et al., 2002). A too small bandwidth may lead to a large variance in the results because a small number of data points is used in the local calibration, while a too large bandwidth may lead to biased results since the data are drawn from locations further away from the

regression point (Fotheringham, 2009). Therefore, whichever weighting function is used in GWR, an optimum value for the bandwidth (in the case of the fixed kernel) or the optimal number of nearest neighbours to be considered in the calibration (in the case of the flexible kernel) should be estimated.

There are different methods in the literature that can be used for this purpose (see Fotheringham et al., 2002; Fotheringham, 2009). For instance, the optimum bandwidth parameter can be obtained by minimising a cross-validation (CV) score which is based on minimising the squared error of the dependent variable. The CV approach was initially proposed by Cleveland (1979) for locally weighted regression. Later the CV method has been adapted based on an integrated squared error for kernel density estimation (see e.g. Rudemo, 1982; Bowman, 1984) and eventually was used in GWR (see Brunsdon et al., 1996; Farber and Páez, 2007; Páez and Wheeler, 2009; Wheeler and Páez, 2010). The general form of the CV can be written as:

$$CV = \sum_{i=1}^{n} [\ y_i - y'_{\neq i}(b)\ ]^2 \tag{5.7}$$

where $y'_{\neq i}(b)$ is the fitted value of y_i with the observations for point i omitted from the calibration process and parameter b is the bandwidth. The CV can be estimated using an optimization technique such as golden section search, assuming the cross-validation function is reasonably well behaved (Fotheringham et al., 2000; Greig, 1980). Plotting the CV scores against the bandwidth values provides a guidance for selecting the optimum bandwidth parameter (Fotheringham et al., 2002).

A similar method, which is an approximation to the CV but easier to compute, is the generalised cross-validation criterion (GCV) (see Fotheringham et al., 2000). The formula for the GCV score is:

$$GCV = n\sum_{i=1}^{n} [\ y_i - y'_i(b)\ ]^2/(n - tr(S))^2 \tag{5.8}$$

where $tr(S)$ is the trace of the *hat matrix* S (Hoaglin and Welsch, 1978; Fotheringham et al., 2002) which is equivalent to the effective number of parameters in the model. The hat matrix provides information about the influence of each observed value on each fitted value and gives each fitted value $\hat{y}_i$ as a linear combination of the observed values y_i as $\hat{y}_i = Sy_i$. Each row of the hat matrix can be calculated by:

$$r_i = X_i(X^T\ W(u_i, v_i)\ X)^{-1}X^T\ W(u_i, v_i). \tag{5.9}$$

where (u_i, v_i) is the coordinates of i.

Another possible method for estimating the optimum bandwidth is to minimise the Akaike Information Criterion (AIC), (see Akaike, 1973, 1974), which provides a trade-off

between the complexity of the model and the goodness-of-fit with the following formula:

$$AIC = 2n\ ln(\hat{\sigma}) + n\ ln(2\pi) + n\ + tr(S) \tag{5.10}$$

where n is the sample size and $\hat{\sigma}$ is the estimated standard deviation of the error term (Fotheringham et al., 2002; Hurvich et al., 1998). The model with the smallest AIC, which is called the minimum AIC estimator, is the model with the optimum bandwidth. In some situations, when the number of parameters is relatively big compared to the number of observations, the AIC estimator may perform poorly or may even be biased (see Cheng et al., 2011; Sakamoto et al., 1986; Sugiura, 1978). To avoid this problem, a small sample bias adjustment (second order) (see Hurvich and Tsai, 1989) has been incorporated (notably by Hurvich et al. (1998)) which led to a corrected AIC (AICc) estimator. In local regressions the degrees of freedom is likely to be small, so the use of AICc is more appropriate than AIC. Following Hurvich et al. (1998), Fotheringham et al. (2002) proposed the following AICc formula for use in GWR which provides a trade-off between goodness-of-fit and degree of freedom as:

$$AICc = 2n\ ln(\hat{\sigma}) + n\ ln(2\pi) + n\ \left(\frac{n + tr(S)}{n - 2 - tr(S)}\right). \tag{5.11}$$

When the effective number of parameters in the model is small relative to the number of observations, the difference between AIC and AICc is insignificant (see Nakaya et al., 2005).

5.3.3 Geographically weighted Poisson regression (GWPR)

GWR has initially been developed for linear regression modelling where the dependent variable is assumed to follow a Gaussian (normal) distribution; however, later the geographically weighted method has been extended on the basis of the generalised linear modelling framework (see section 3.8) for Binomial (logistic) distribution as *geographically weighted logistic regression* and for Poisson distribution as *geographically weighted Poisson regression (GWPR)* (Nakaya, 2007; Fotheringham et al., 2002; Nakaya et al., 2007, 2005; Lovett et al., 1986; Lovett and Flowerdew, 1989). In this thesis, where we aim to localise spatial interaction models using the GWR technique, the GWPR provides a more appropriate framework compared to the conventional Gaussian regression since the spatial flows are discrete and nonnegative (see section 3.8 for further details).

The core principle of the GWPR is similar to the basic GWR, in that the parameter estimates are allowed to vary geographically and these variations can be estimated with a spatial weighting kernel (see Nakaya et al., 2005). A comprehensive introduction to the GWPR principle is given by Nakaya et al. (2005) (see also Fotheringham et al., 2002; Nakaya, 2001, 2003). In this section we present a summary of the theory behind the GWPR technique, drawing on insights from Nakaya et al. (2005); Fotheringham et al.

(2002). The traditional form of a Poisson regression model links a dependent variable y_i that follows a Poisson distribution and a number of independent variables x_{ki}, and is generally defined as:

$$y_i = \exp(\beta_0 + \sum_k \beta_k x_{ki} + e_i) \tag{5.12}$$

where β_k are parameters of the model (with β_0 for intercept), and e_i is the error term. A GWR extension of the above model allows the parameters of the model to vary over space by incorporating the coordinates (u_i, v_i) of the regression point in the model formulation:

$$y_i = \exp\big(\beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i) x_{ki} + e_i\big). \tag{5.13}$$

To estimate the GWPR parameters, Nakaya et al. (2005) consider a geographically weighted likelihood principle which is a variant of the local likelihood principle (Loader, 1999) and which is consistent with the geographically weighted least squares in conventional Gaussian GWR. Using this methodology, the model parameters at location i are estimated by maximising the geographically weighted log-likelihood function:

$$maxL(u_i, v_i) = \sum_{j=1}^{n} \big(- \hat{y_j}(\beta_i) + y_j \; ln \; \hat{y_j}(\beta_i)\big) \; . \; w_{ij} \tag{5.14}$$

where $\hat{y_j}(\beta_i)$ is the predicted value of y at location j with parameters at regression point i and w_{ij} is the geographical weight of the jth observation at the ith regression point. As suggested by Fotheringham et al. (2002); Nakaya et al. (2005), the above equation can be maximised using a local Fisher scoring, a form of interactively re-weighted least squares (Hastie and Tibshirani, 1990). In this local scoring method, as the procedure is iterative, an initial guess at the regression coefficients is usually made by an ordinary least squares regression and then the parameter estimates are updated until convergence by repeating a matrix computation of weighted least squares. Computational details for the calibration procedure of GWPR can be found in Nakaya et al. (2005).

Similar to the conventional GWR, GWPR considers a spatial kernel around each calibration point i and the observations are weighted gradually according to their proximity to the centre of the kernel where the observation i has the maximum weight. The choice of the geographical weighting kernel is similar to GWR in which different spatial weighting functions are available such as Gaussian or bi-square function both for fixed and adaptive bandwidths (see section 5.3.1). For calibrating the selected geographically weighting function in GWPR, Nakaya et al. (2005) suggest some possible methods where in a similar way as in GWR, means the optimal bandwidth is selected in term of some criteria. For instance, minimising the AIC of the model is one of the suggested methods:

$$AIC(b) = Deviance(b) + 2k(b) \tag{5.15}$$

where $k(b)$ is the effective number of parameters in the model with bandwidth b. The corrected AIC, AICc, for using in GWPR is formulated as:

$$AICc(b) = Deviance(b) + 2\ k(b)\big(\ \frac{n}{n - k(b) - 1}\big) \\ = AIC(b) + 2\ \frac{k(b)\big(k(b)+1\big)}{n - k(b) - 1} \tag{5.16}$$

As with the AICc in GWR, the model with smallest AICc is the model with the best bandwidth suggestion. As a rule-of-thumb, if the differences between AICc values of two models is less than around 2, then the competition between models is too close and the difference between the models is negligible (Nakaya et al., 2005; Fotheringham et al., 2002).

Nakaya (2001, 2003) argue that as the number of observations increases, the AIC and CV may perform poorly and a more complex model is attained with these estimators as the best model instead of the true one. Therefore, the AIC and CV estimators lead to a smaller bandwidth size for the local modelling. On the other hand, the derivation of the AICc method is based on Gaussian linear models and when the underlying probability distribution is extremely non-normal, using the AICc may not be appropriate (see Burnham and Anderson, 2002; Nakaya et al., 2005). As an alternative, Nakaya (2001, 2003) proposed a Bayesian based bandwidth selection estimator for use in GWPR known as Bayesian Information Criterion (BIC), sometimes referred to as the Schwartz Information Criterion (SIC)(Schwartz, 1978). The BIC is defined as:

$$BIC(b) = Deviance(b) +\ k(b)\ ln(n) \tag{5.17}$$

where $k(b)$ is the effective number of parameters and n is the number of observations. The model with the smallest BIC is the best fit model and guides the optimum bandwidth value. The model complexity penalty in BIC is weighted by the number of observations. Therefore, the same degree of complexity (that is, the same value of k) carries a higher penalty for larger samples in BIC than in similar measures such as AIC. Consequently, using BIC in large samples tends to identify models with fewer parameters as optimal (Fotheringham et al., 2002). Although the estimator selected by BIC may be more biased compared to AIC, the bias is negligible in large samples (Nakaya, 2001, 2003).

5.4 Local calibration of spatial interaction models

One of the earliest attempts of local spatial analysis that perhaps predates all other local models, is that of local spatial interaction (Fotheringham et al., 2000). It was recognised quite early that a global calibration of spatial interaction models may fail to capture

spatial variation of relationships over space and so not represent the true specification of reality (Fotheringham et al., 2002). A large amount of spatial information on interaction behaviour could be completely hidden in the output of a global calibration of spatial flow models, so local parameter estimates potentially provide much more useful disaggregated information (Linneman, 1966; Greenwood and Sweetland, 1972; Fotheringham et al., 2000, 2002). The main evidence for this was shown when the spatial interaction models were calibrated separately for each specific origin and destination instead of a single global estimate (see Fotheringham, 1983, 1984b, 1986). When the resultant local parameter estimates were mapped over space, they showed a clear spatial variation (see Fotheringham, 1981, 1983, 1984a,b, 1991; Meyer and Eagle, 1982; Fotheringham and O'Kelly, 1989; Fotheringham et al., 2000). This spatial variation would be completely ignored with a global calibration (Fotheringham et al., 2002). The spatial interaction models that are calibrated for separate subsets of data to provide local information for each specific origin and destination in the system are referred to as *origin- and destination-specific models* respectively. A large number of examples using origin- and destination-specific models to provide spatial disaggregated information of interaction behaviour over origins and destinations exist in the literature (see for instance Haynes and Fotheringham, 1984; Fotheringham and O'Kelly, 1989).

In the following section, we show the general formulations of the unconstrained gravity origin- and destination-specific models along with an application of the models using the journey-to-work dataset of Lausanne. We map the local results over origin and destination communes in order to visualise possible spatial variations of the model parameters over space and to compare the local results with the global calibrated parameters.

5.4.1 Origin-specific spatial interaction model

Consider a general unconstrained gravity spatial interaction model:

$$T_{ij} = \kappa \, {v_i}^{\alpha} \, w_j^{\gamma} \, d_{ij}^{\beta} \tag{5.18}$$

where T_{ij} represents the flow between regions. Variables are defined as before in chapter 3, v_i represents origin propulsiveness, w_j is destination attractiveness and d_{ij} indicates distance between i and j; κ, α, γ and β are global parameters of the model. These global parameters represent an average type of interaction behaviour and are valid equally for the entire study region. An origin-specific version of this model is applied when only flows from one origin i going to different destinations are considered in the calibration process; see figure 5.1 for an illustration of the origin-specific model. A general origin-specific version of the unconstrained spatial interaction model can be represented by the following formula:

$$T_{ij} = \kappa_i \, {w_j}^{\gamma_i} \, d_{ij}^{\beta_i} \tag{5.19}$$

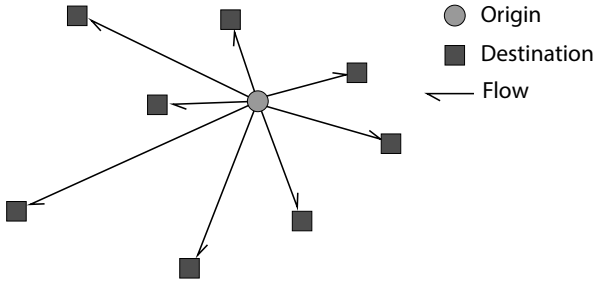


Figure 5.1: A simplified illustration of the origin-specific spatial interaction.

where the value of the origin attribute variable v_i^α becomes a constant (as only one origin is considered in the model calibration) and usually is subsumed into the balancing parameter κ_i. The parameters κ_i, γ_i and β_i are specific to origin i (Haynes and Fotheringham, 1984). The calibration process of the model can be repeated for all origins in the system and the resulting specific parameter estimates can be mapped over space.

In order to provide an example of the origin-specific model, we apply a Poisson version of the model to the journey-to-work dataset in Lausanne agglomeration, using the following formula:

$$\lambda_{ij} = \exp(\kappa_i + \gamma_i \ln N_j + \beta_i \ln d_{ij}) + \varepsilon_{ij} \tag{5.20}$$

where as before λ_{ij} shows the number of interactions between i and j, including intra-zonal flows, N_j the number of jobs in destination j and d_{ij} is the distance between regions calculated with the population density-based scattering method described in section 4.3.2, κ_i, γ_i and β_i are parameters of the model specific to origin i. We calibrated the model for each commune using a subset of data flows from each specific origin in the agglomeration. A set of local parameter estimates is obtained for each origin separately. Table 5.1 shows the results of calibrating the local origin-specific model in the Lausanne agglomeration, where the mean, minimum, maximum, standard deviation and quartiles 25%, 50%, and 75% of the parameter estimates and their t-values along with the Deviance and Pseudo R^2 of the models are presented. The results of a global spatial interaction calibrated for the same area are listed in the table for comparison, although a direct comparison of the model cannot be done because the global model contains more variables (i.e the origin attribute variable, here active population). Also, in the calibration of the global model all the data are considered while in the calibration of an origin-specific model, only a part of data related to that specific origin are involved. The values of parameters in both global and local models show the general expected effect on interaction, i.e. number of jobs at a destination has a positive influence and distance has a negative effect on the total interaction. The global distance-decay parameter of -1.2973 shows a less negative effect in comparison with the mean value of -1.8247 over all communes in the origin-

Table 5.1: The Poisson global and origin-specific models for journey-to-work in the agglomeration of Lausanne.

Global model

Model type	Poisson		
Deviance	32575.3		
Pseudo R^2	0.9561		
Parameters	*Estimates*	*Std error*	*t-values*
Active pop.	0.7908	0.0019	414.8
Jobs	0.9486	0.0018	534.3
Distance	-1.2973	0.0039	-334.6
Intercept	0.8181	0.0385	21.3

Origin-specific model

Parameter	*Mean*	*Min*	*Max*	*Std Dev*	*Quartiles*		
					25%	*50%*	*75%*
Jobs	1.0598	0.7645	1.2710	0.1032	1.0040	1.0599	1.1194
Distance	-1.8247	-2.7482	-0.8144	0.4180	-2.0740	-1.8291	-1.5395
Intercept	9.7697	3.9305	19.2064	2.9938	7.5179	9.9179	11.6682
t-values jobs	42.2	5.7	168.9	30.0	22.5	31.8	52.7
t-values distance	-32.9	-100.5	-7.8	16.8	-41.1	-29.5	-19.9
t-values intercept	28.4	7.7	69.5	13.3	19.4	24.9	34.3
p-values jobs	0.00	0.00	0.00	0.00	0.00	0.00	0.00
p-values distance	0.00	0.00	0.00	0.00	0.00	0.00	0.00
p-values intercept	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deviance	172.4	-182.7	1390.9	211.5	65.5	105.7	202.6
Pseudo R^2	0.8822	0.6365	0.9959	0.0740	0.8392	0.8949	0.9422

specific model. Comparing the min, max and standard deviation of the local parameters shows a clear variation between the values. The parameter for the number of jobs in the origin-specific model varies between 0.7645 and 1.2710 while the standard deviation is 0.1032. The standard deviation indicates the variation from the average across all communes. The local distance-decay parameters show higher variation from the mean, (i.e. standard deviation of 0.4180) and the estimates range from a minimum of -2.7482 to a maximum of -0.8144.

In order to facilitate further analysis of spatial variation in the model's results, we mapped the local parameter estimates and their t-values over communes (see figures 5.2 and 5.3). Comparing these maps with the overview maps of the Lausanne communes in figures 2.2 and 2.3 makes it easier to understand and analyse the source of these variations. For instance the distance-decay parameter in Lausanne city and the neighbouring industrial areas is less negative than average indicating the inhabitants of these areas consider distance to be a less important deterrent for their daily journey-to-work compared to people in other parts of the agglomeration where the distance-decay is more negative. This variation partly can be explained with the better transportation system in central Lausanne and the neighbouring regions. However, sometimes there are variations in parameter estimates that are not so easy to explain. For instance, the value of the job parameter in Lausanne city is high but a dramatic drop in the value occurs in some neighbouring communes. The t-values of the parameters are mapped in figure 5.3.

In general the t-values are considerably different from 0 indicating the significance of the local parameters.

5.4.2 Destination-specific spatial interaction model

Similar to the origin-specific model, a destination-specific model can be generated when a spatial interaction model is calibrated using flow data going to a specific destination j. See figure 5.4 for a general overview of the destination-specific model in an interaction system. The following formula shows a destination-specific version of the gravity model shown by the equation 5.18:

$$T_{ij} = \kappa_j \, {v_i}^{\alpha_j} \, d_{ij}^{\beta_j} \tag{5.21}$$

where the value of the destination attribute variable w_j^{γ} is a constant (as only one destination is considered in the model calibration) and is subsumed into the balancing parameter κ_j, and the parameters κ_j, α_j, and β_j are specific to destination j (Haynes and Fotheringham, 1984). As in the origin-specific model, the calibration process can be repeated for all destinations in the system and the resulting parameter estimates can be mapped over space.
As an example we apply a Poisson version of the destination-specific model on the journey-to-work dataset in the Lausanne agglomeration, considering the following formula:

$$\lambda_{ij} = \exp(\kappa_j + \alpha_j \ln P_i + \beta_j \ln d_{ij}) + \varepsilon_{ij} \tag{5.22}$$

where the variables are defined as before and the parameters κ_j, α_j, and β_j are specific to destination j. Repeating the calibration for each specific destination in the Lausanne agglomeration, we list the local results in table 5.2. Maps of the local parameters and their t-vales are shown in figures 5.5 and 5.6 respectively.

Although the local parameter estimates from the destination-specific model generally indicate the expected influence on the total interaction, i.e. negative for distance-decay and positive for active population, some unexpected values occur. For instance, the average value for the local active population parameter is 0.8830, higher than the global value of this parameter which is equal to 0.7908. However, its minimum value is -0.7594 which is completely counter intuitive. By checking the resulting local active population parameters for all destinations, it turns out that there is only one negative value between all destinations. This negative active population parameter is for Malapalud, the smallest commune in the agglomeration which has a total of 61 inhabitants according to census data 2000. There are only 30 people in the commune who are economically active and the number of jobs is 25 mainly in agriculture. In total only 1 person is commuting towards this destination, coming from another small rural commune which explains the negative parameter for the active population. Additionally, the t-value of this parameter

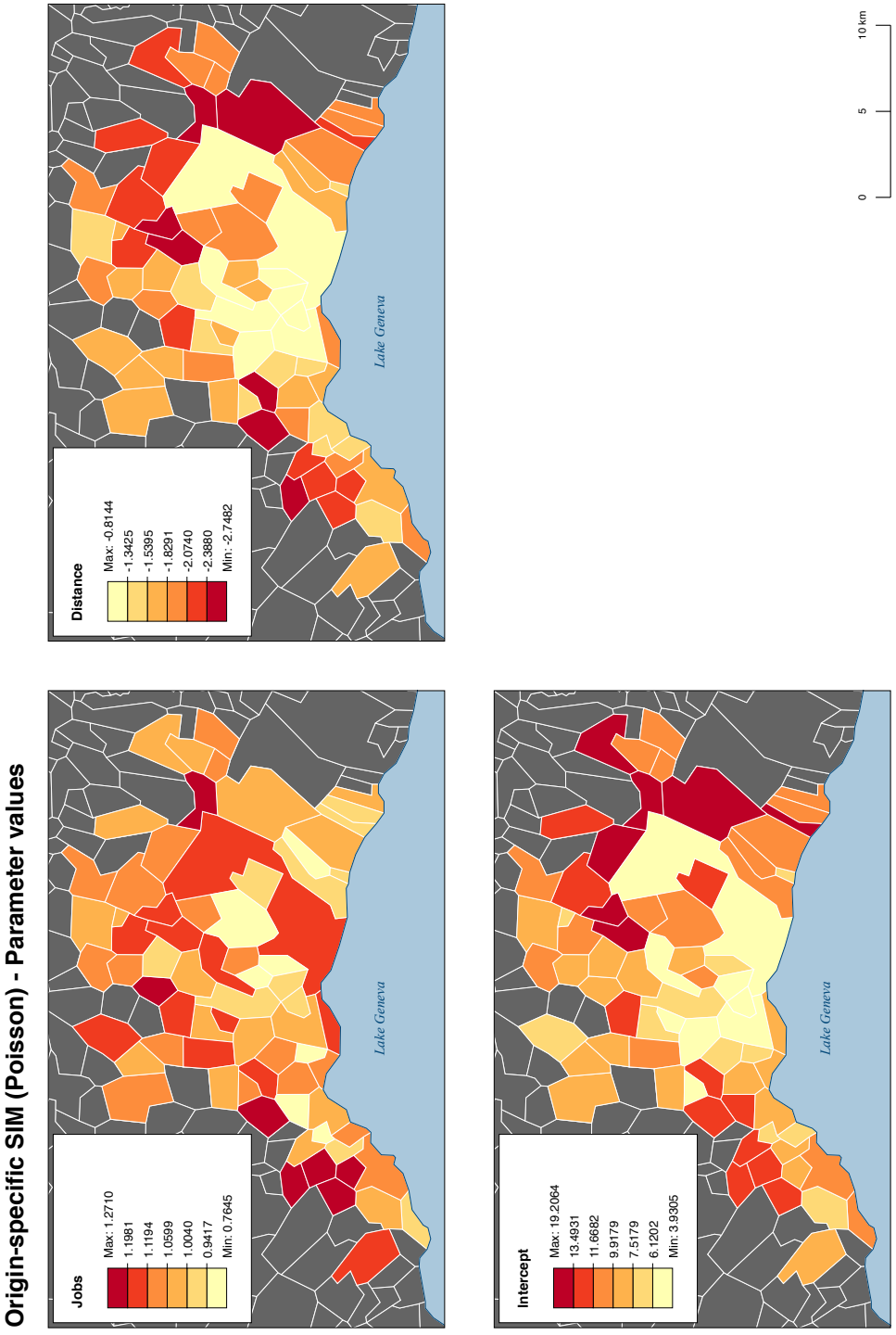


Figure 5.2: The parameter values for Poisson origin-specific model in agglomeration of Lausanne.

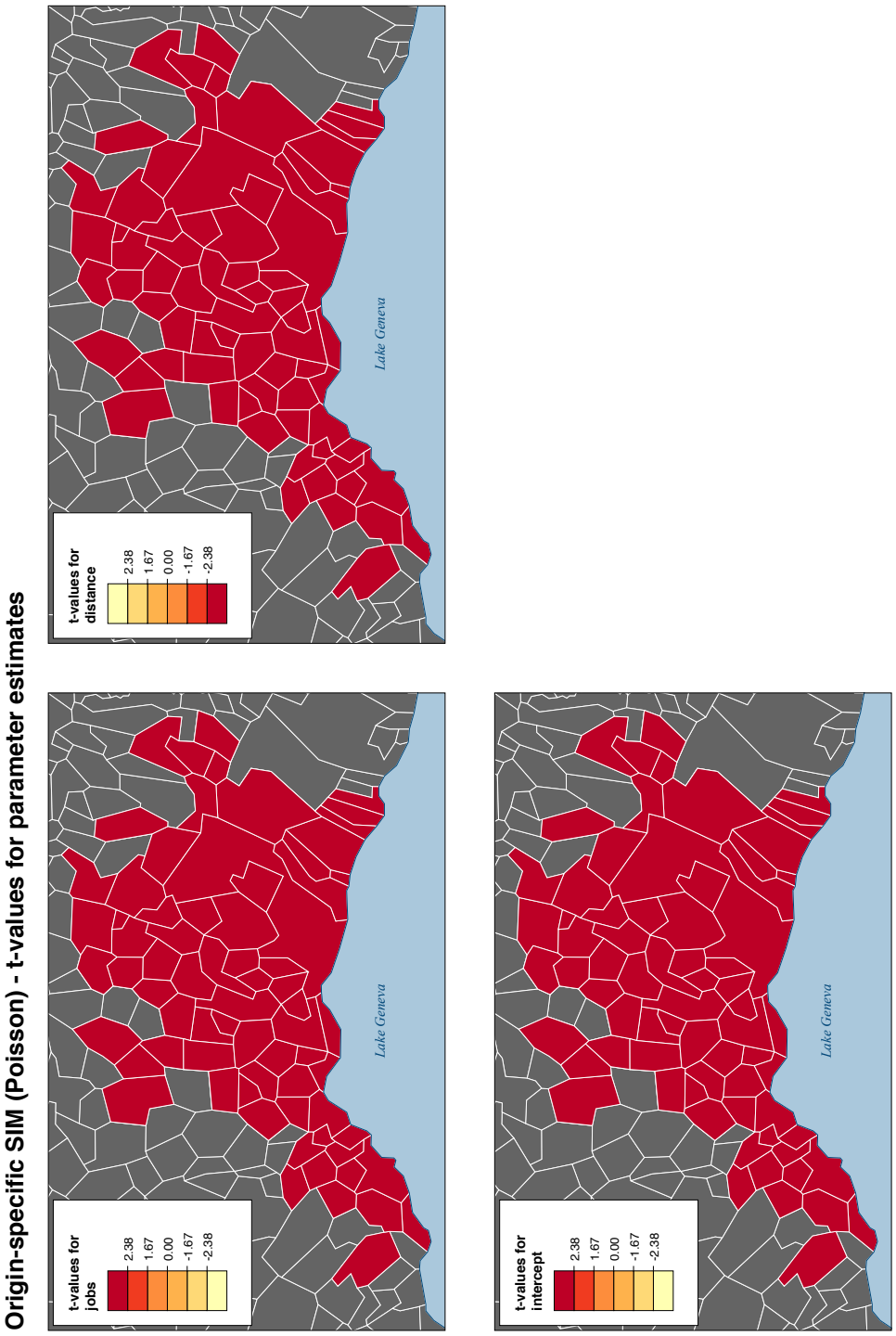


Figure 5.3: The t-values of the parameters for Poisson origin-specific model in agglomeration of Lausanne. These maps display absolute t-values, where values greater than 2.33 are significant at a level of 99%, and values greater than 1.65 are significant at a level of 95%. For negative parameter values (for the distance decay parameter), the negative t-values of -2.33 and -1.65 correspond to the significance levels of 99% and 95% respectively.

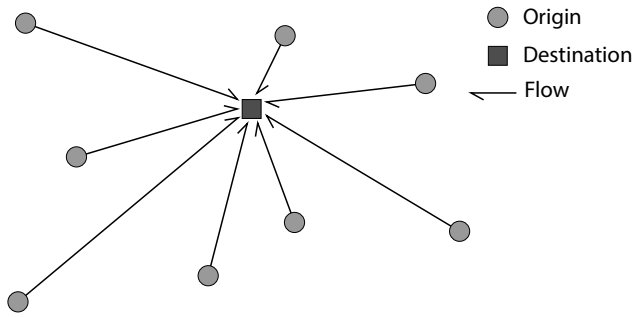


Figure 5.4: A simplified illustration of the destination-specific spatial interaction.

Table 5.2: The Poisson destination-specific model for journey-to-work in the agglomeration of Lausanne.

Destination-specific models							
Parameter	*Mean*	*Min*	*Max*	*Std Dev*	*Quartiles*		
					25%	*50%*	*75%*
Active population	0.8830	-0.7594	1.8510	0.3053	0.8000	0.8809	1.0068
Distance	-1.9847	-4.4205	-0.5854	0.6904	-2.4184	-1.8805	-1.4386
Intercept	11.5903	3.8253	23.3434	4.0070	8.4200	10.9907	14.4038
t-values act. pop.	26.4	-0.6	204.1	31.1	9.0	19.1	31.9
t-values distance	-30.0	-102.2	-1.4	19.3	-41.0	-24.4	-15.2
t-values intercept	26.7	4.5	70.8	12.0	17.8	24.9	33.7
p-values act. pop.	0.02	0.00	0.55	0.08	0.00	0.00	0.00
p-values distance	0.00	0.00	0.17	0.02	0.00	0.00	0.00
p-values intercept	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deviance	153.1	-62.9	1620.3	218.8	43.8	95.8	182.4
Pseudo R^2	0.8620	0.5427	0.9918	0.0808	0.8232	0.8771	0.9207

is -0.6 indicating that the parameter estimate for active population for this destination is not significantly different from 0.

For the local distance-decay parameters, the average estimate over all communes is -1.9847 compared to the value from the global model which is -1.2973. So, locally distance is perceived as a bigger barrier for daily travelling compared to globally. As was discussed earlier, a direct comparison of the local and global model is difficult as the number of variables and parameters are different. However, a general comparison of the models is useful as it shows how the locally disaggregated information provided by destination-specific model, e.g. spatial variation of the model's parameters, would be totally missed with a global calibration.

5.4.3 Local calibration of spatial interaction based on a GWR approach

The Lausanne journey-to-work spatial interaction results provide an example of spatial heterogeneity in the processes generating flows. The nature of this systematic spatial variation has been described by previous researchers (see for instance Curry, 1972; Curry et al., 1975; Sheppard, 1979, 1978; Fotheringham, 1981; Gordon, 1985; Thorsen

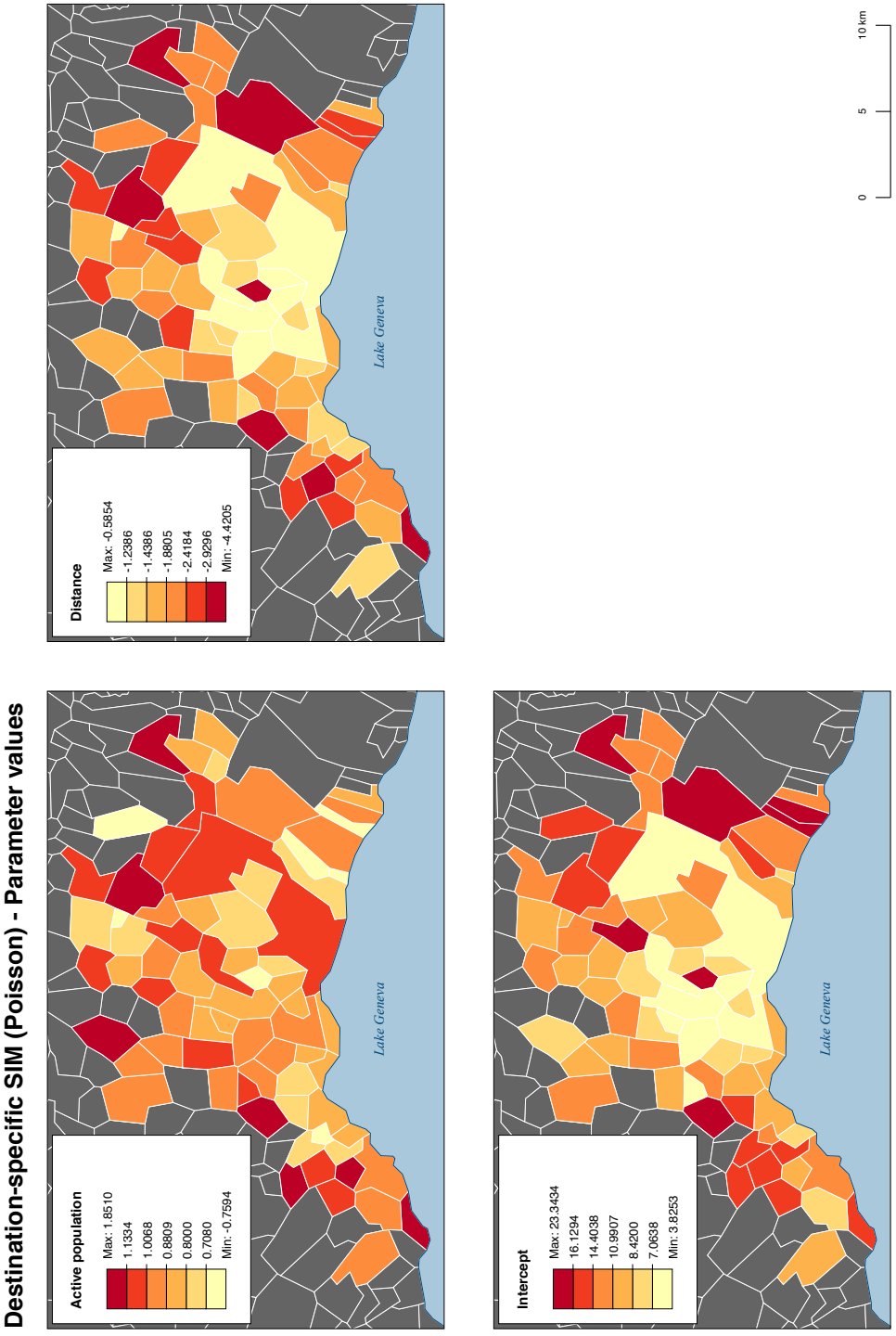


Figure 5.5: The parameters values for Poisson destination-specific model in agglomeration of Lausanne.

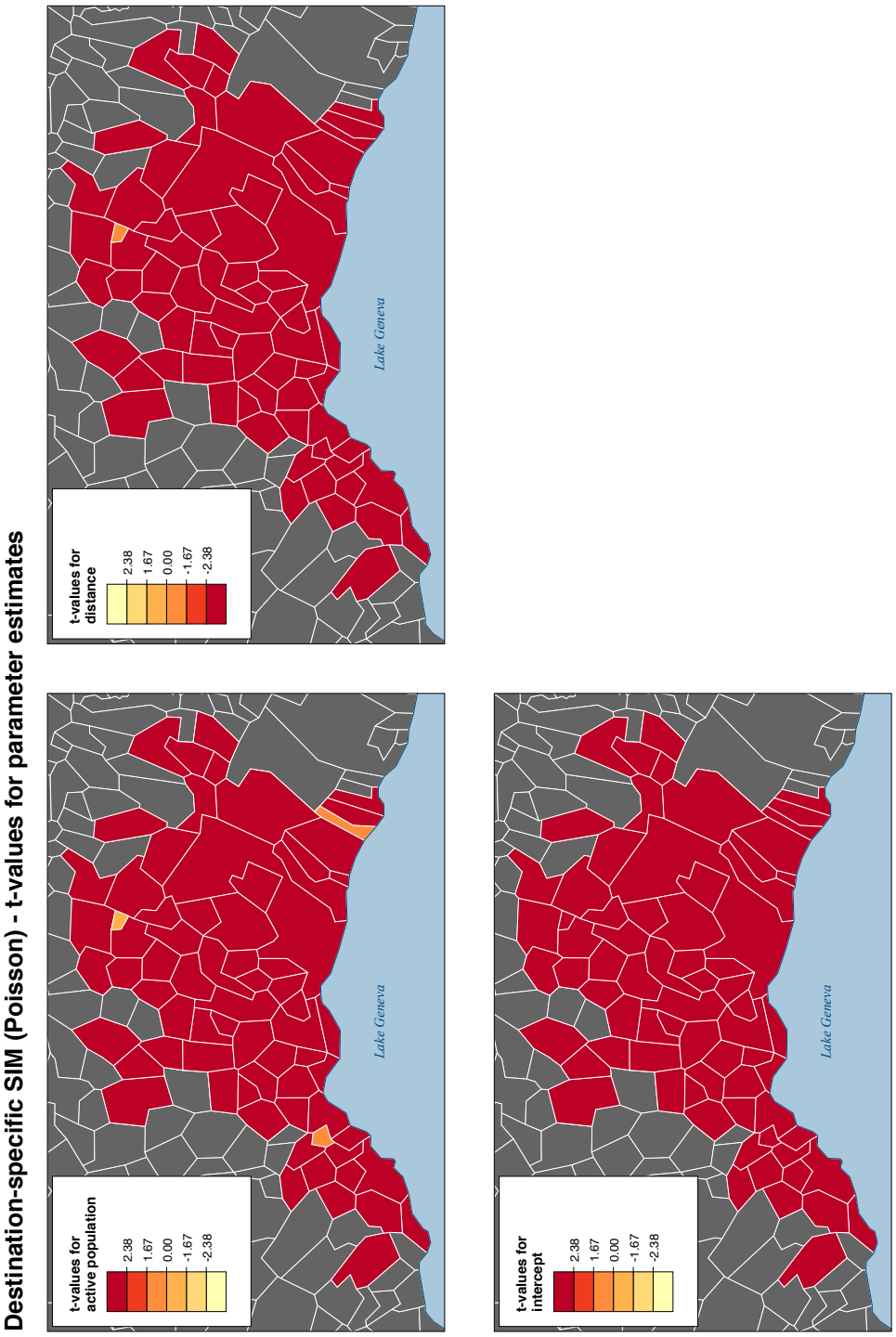


Figure 5.6: The t-values of parameters for Poisson destination-specific model in agglomeration of Lausanne. These maps display absolute t-values, where values greater than 2.33 are significant at a level of 99%, and values greater than 1.65 are significant at a level of 95%. For negative parameter values (for the distance decay parameter), the negative t-values of -2.33 and -1.65 correspond to the significance levels of 99% and 95% respectively.

and Gitlesen, 1998). As stated by Fotheringham and Webber (1980), this systematic spatial variation may be partly due to variation in the underlying spatial structure of the systems (see also Fotheringham, 1982b,a). For instance, if spatial configurations of origins and destinations in a spatial system influence the spatial pattern of parameter estimates (see Fotheringham and Webber, 1980; Fotheringham, 1981). Although modified spatial interaction models, such as the competing destinations model (see section 3.6), by incorporating relevant measures explicitly can remove the spatial structure effect from the interaction patterns and result in substantial improvements in model accuracy (Fotheringham, 1991; Yano et al., 2003; Fotheringham et al., 2001b; Pellegrini and Fotheringham, 2002), significant geographical variations of parameters in the interaction system can still remain in the model (Ishikawa, 1987; Yano et al., 2000; Nakaya, 2001, 2003). For instance, the results of Nakaya's study shows significant spatial heterogeneity of the accessibility parameter estimates when a competing destination origin-specific model was calibrated for a migration case study in Japan (see Nakaya, 2001, 2003).

Another attempt to capture spatial heterogeneity in spatial interaction models is to localise the models by applying the same local techniques for analysing normal spatial data on flow data. One of the earliest attempts in this regard is Fotheringham and Pitts (1995) where directional drifts of the spatial interaction model parameters are measured using the expansion method. Later Berglund and Karlström (1999) pointed out in their publication on local spatial association that GWR is, in its principle of exploring nonstationarity in relationships between different sets of variables over space, applicable to many different models including spatial interaction models. Nakaya (2001, 2003) investigated the capability of GWR for local calibration of flow models using a migration case study in Japan. As in GWR the regression model is calibrated repeatedly for each geographical location (i.e. calibration point) in space. Nakaya suggests calibrating the spatial interaction model for each specific flow in the interaction system. In his work, however, an origin-specific spatial interaction model is calibrated using the GWR technique in which a spatial kernel is considered around destination j of flow ij and flows from origin i to different destinations are weighted according to the distance between their destinations to the calibration point (i.e. destination j). So in Nakaya's work, for each specific origin in the system a local calibration of the model is obtained using GWR and the calibration points are destinations of the observed flows, not the observed flows themselves. This approach is described in more detail, following an empirical example, in the next chapter (see section 6.5).

Another example of using GWR for calibrating an origin-specific spatial interaction model is presented by Nissi and Sarra (2011) which is slightly different from Nakaya's approach. In this work, the authors introduced a modified version of the weighting function which considers both the distance and a new parameter of "strength of connection"

between destinations which is based on total interactions between destinations:

$$w_{k(j)} = exp\big(- \frac{d_{jk}^2}{b^2} \times (\text{strength of connection})\big) \tag{5.23}$$

where d_{jk} shows the distance between j and k, b is the bandwidth parameter, and the strength of connection represents a value of interaction between two destinations and is defined by the following formula:

$$\text{Strength of connection between k and j} = \frac{T_{kj}^2}{T_{ok} \times T_{oj}} \tag{5.24}$$

where the interaction between destinations k and j is shown as T_{kj}, T_{ok} and T_{oj} represent the total number of flows terminating at destinations k and j respectively. The original idea of this weighting function is to give higher weights to more connected destinations, (i.e. destinations that share more visitors); when k and j represent the same location, a unit weight is allocated to the destination. However, this formulation fails to address a situation when the destinations k and j are located in two different locations with no flows in between. In this case the strength of connection between k and j will be equal to zero and multiplication of zero to $\frac{d_{jk}^2}{b^2}$ in equation 5.23 gives rise to a unity weighting value due to the exponential function. In fact with this formulation, even a far destination k with no connections with destination j will have the highest weight of unity value which is totally opposite to the purpose of the GWR principle.

The GWR technique has been used in a wide range of applications such as analysing spatial variations in average rainfall and altitude relationship by Brunsdon et al. (2001), in school performance by Fotheringham et al. (2001a), in housing attribute price by Bitter et al. (2007), in crime patterns by Cahill and Mulligan (2007); Wheeler and Waller (2009), in real estate price by Huang et al. (2010), in mortality rates by Holt and Lo (2008), among others. The application of GWR can be found also in commuting analysis. For instance, Lloyd and Shuttleworth (2005); Shuttleworth and Lloyd (2005) use GWR to study the relationships between the average commuting distance and some socio-economic variables in Northern Ireland. However, these studies can not be considered in the category of local calibration of spatial interaction since the dependent variable is the average travel distance not the travel flows themselves.

Although the efficiency of GWR in local spatial data analysis has been shown in different applications, surprisingly there are not many applications of the geographically weighting concept in spatial interaction models. The reason is possibly due to the complexity of applying GWR on spatial flows since usually interaction models involve a huge amount of data covering information for all origins, destinations and flows between the regions. The existing software for GWR works fine for a reasonable size of spatial data in normal regression models but based on our experience has difficulties to run for spatial flow data. This might also to some extent explain the lack of GWR applications for

spatial flow analysis. In this thesis, we developed a simplified version of GWR software using Python that works for calibrating spatial interaction models[1].

In the following chapters we expand the idea of using the concept of geographically weighting for local calibration of spatial flows further. More specifically, we do not restrict the models to the origin- or/and destination-specific type of spatial interaction models but we develop this idea as a general local calibration method that can be applied to any type of spatial interaction model. Based on this idea, the spatial interaction models can be calibrated in any arbitrary location within the study region and the resulting parameters can be mapped over space to visualise any potential spatial heterogeneity in the system for further investigation. Also, based on the same idea of using the geographically weighted concept for localising spatial interaction model, a local calibration method will be described for calibrating each arbitrary flow (pair of locations) within the system.

[1] available at `https://github.com/mkordi/pygwr`

Chapter 6

Geographically weighted spatial interaction (GWSI)

6.1 Introduction

In the following two chapters we focus on localising spatial interaction models using geographical weighting in three different ways. In two of the three, we calibrate the interaction models for calibration points which are actual geographic locations within the study region (e.g. centroids of origins, destinations). The data are spatial flows between origins and destinations in the agglomeration of Lausanne in Switzerland. Considering a spatial kernel around the calibration point, the observed data flows are weighted according to their distance to the centre of the kernel. The distance estimation is categorised in two parts: 1) when the distance between the calibration point and the origin of the flows is considered (we name this an origin-focused approach), 2) when the distance is defined considering the destination of the observed flows (we term this a destination-focused approach). In these scenarios, we do not restrict the method to one of calibrating only origin- and/or destination-specific models, but we consider a broader approach which differs from the approach of Nakaya (2001). A comprehensive framework for local calibration of spatial interaction models is developed in which any type of spatial interaction model can be calibrated in any arbitrary location within the study region. The origin- and destination-focused approaches are discussed in more detail in this chapter while in a third approach for localising the spatial interaction models, we apply the geographically weighted concept with both observations and calibration data being spatial flows. The interaction model is then calibrated for each spatial flow, we call this a flow-focused approach. This model will be presented and described in details in the next chapter.

6.2 Origin-focused GWSI approach

As discussed before (see chapter 5), for a particular relationship in an origin-specific model, a parameter is estimated separately for each origin compared to a single average

estimate for the entire study area obtained in a global model. In a spatial interaction system with an m-by-n origin-destination matrix, the calibration process can be repeated for each of the m origins. The resulting parameters for all m origins can then be mapped to examine spatial variations in interaction determinants over origins. This can be used to visualise possible spatial heterogeneity when the parameter estimates vary across origins. However, an origin-specific model provides an average type of parameter estimates localised to the entire origin region. If the region representing the origin is relatively large, or when the relationships under study varies across different destinations, it would be more useful to estimate the parameters of the spatial interaction models at a more spatially disaggregated level, rather than only at the level of origin regions.

Additionally, an origin-specific model only considers flows that originate from a specific origin i to different destinations so that flows emanating from other origins in the system are ignored in the model calibration for origin i. However, sometimes information from other origin regions in the model can be useful, particularly when we are interested in calibrating the model's parameters for locations with no flow data available. In this situation, the information from neighbouring origins can be used to estimate the model parameters at the calibration origin. For instance, in our case study of journey-to-work, if workforces in nearby villages with similar conditions show similar behaviour in their daily commuting, (e.g. a similar level of interest in choosing a specific workplace or destination j), we can use this information for estimating the model parameters (level of interest of inhabitant for choosing destination j) for a nearby origin where there is no commuting information available.

In order to take into account information at a regional scale instead of a single origin, we can apply the principle of geographical weighting on the flows when a cluster of origins around the calibration point is considered in the calibration process of the model. The general principle of geographical weighting can be applied on the interaction flows as it is used in GWR. However, spatial interaction is a multidimensional phenomenon and applying the concept of geographical weighting to flows involves not only a single location for a calibration point and a set of surrounding data point as in normal GWR, but involves locations of origins, destinations and the interactions between them. Therefore applying the concept of the geographical weighting to spatial interaction is more complicated than applying GWR to non-flow spatial data. In applying the geographically weighted concept to the spatial interaction, observations (i.e. data points) are flows between origin and destination regions, while a calibration point can be either a specific flow (i.e. when the model is calibrated for each separate pair of origins and destinations in the system), or a single geographic location within the study area. In this chapter we focus on geographically weighted spatial interaction where the calibration points are absolute geographic points within the origin and destination regions and not flows. In this context, a geographic calibration point can be represented by the centroid of one of the existing origins or destinations, or an arbitrary location within the study area, when

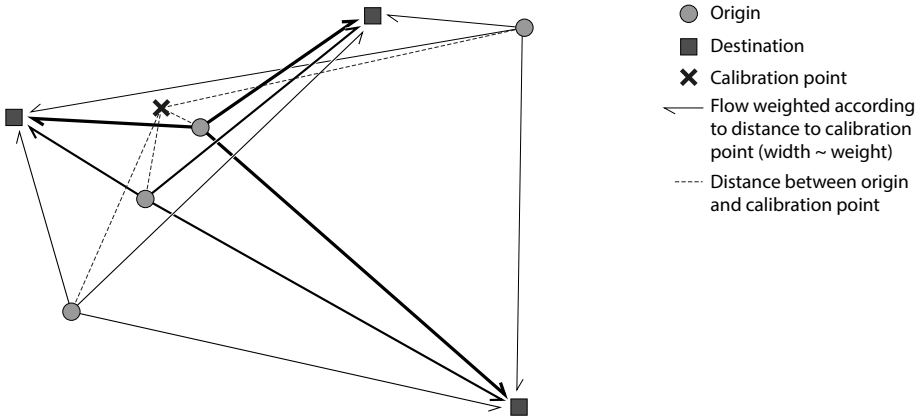


Figure 6.1: A schematic overview of the origin-focused GWSI approach.

we are interested in estimating flows in regions with no prior flow data.

In geographically weighted spatial interaction (GWSI) fitting a spatial kernel to the data involves defining a region around each calibration point. All the observed flows within this region are then weighted and used to calibrate the interaction model. The influence of the observed flows on the calibration process is based on a weighting scheme in which each observed flow is weighted by its distance from the centre of a kernel (i.e. distance to the calibration point); flows with greater weights have more influence on the calibration process. In this section, where we are interested in considering information on neighbouring origins when calibrating a spatial interaction model, we weight the observed flows around a calibration point based on proximity of the origins of the flows to the calibration point. Here, flows with closer origins to the centre of the spatial kernel have a greater weight and have a larger effect on the model calibration procedure. Weights decrease continuously as the distance between the calibration point and the observed origin increases. In this framework, when a calibration point is an existing origin, not only flows from the calibration point are involved in the model calibration but also, focusing on the origins of the neighbouring flows, a cluster of flows around the calibration point influence the local parameter estimation. If a calibration point and an existing origin share the same location (i.e. when the model is calibrated for an existing origin), the maximum weights are assigned to the flows originating from the calibration point, since the distance between that origin and the calibration point is a minimum. Figure 6.1 illustrates a schematic overview of GWSI when the focus is on the origins of the observed flows around the calibration point; this represents an *origin-focused* approach. In figure 6.1, the flows' widths represent their weights according to the distance between the flows' origin and the calibration point; these distances are shown with dashed lines. A wider flow represents a higher weight and a greater influence on the parameters' estimation for the calibration point.

Repeating the GWSI procedure for all calibration points, the observed flows are

weighted differently so that a unique local set of parameter estimates is achieved for each particular location across the region. The results of the local calibrations of the model can be visualised by mapping the local parameter estimates to show possible spatial variations in the determinants of interaction. The following equation shows the general formulation of the GWSI when the focus is on the origin of the flows (i.e. origin-focused approach):

$$T_{ij} = \kappa_{\{u(x,y)i\}}\ v_i^{\alpha_{\{u(x,y)i\}}}\ w_j^{\gamma_{\{u(x,y)i\}}}\ d_{ij}^{\beta_{\{u(x,y)i\}}} \tag{6.1}$$

where u is the calibration point, (i.e. one of the existing origins or any other point within the study region), v_i, w_j and d_{ij} are the model variables (i.e. the origin propulsiveness, attractiveness of destination and distance between origin i and destination j respectively). κ, α, γ and β are parameters of the model. In above formula, the parameters of the model are allowed to vary over space so we make them location-dependent in which the coordinates of the calibration point, $u(x, y)$, become a part of the model formula.

If the spatial interaction model is considered to be a Gaussian model, the relationships between the dependent variable (i.e. interaction) and the independent variables in equation 6.1 can be modelled using a linear regression approach by applying a logarithmic function to both sides of the equation. The model's parameters can then be estimated with weighted least square (WLS) method using the following formula:

$$b'_u = (X^T\ W_{u_{ij}}\ X)^{-1}\ X^T\ W_{u_{ij}}\ T_{ij} \tag{6.2}$$

where b'_u is a vector containing the local parameters of the model at location u, X is a matrix of the independent variables, including a column of ones for the intercept parameter, X^T is the transpose matrix of X, T_{ij} is the vector of dependent variable showing flows from i to j and $W_{u_{ij}}$ is a weighting matrix. The elements of the weighting matrix are defined using a weighting function of distance between the calibration point and origins of the flows (see equation 6.5). The same general principle is applied to calibrate equation 6.1 when the spatial interaction model follows a Poisson distribution. The Poisson origin-focused model then is formulated as follows:

$$\lambda_{ij} = \exp\ (\kappa_{\{ui\}} + \alpha_{\{ui\}} \ln v_i + \gamma_{\{ui\}} \ln w_j + \beta_{\{ui\}} \ln d_{ij}) \tag{6.3}$$

Equation 6.3 can be calibrated using the same geographically weighted likelihood principle described in GWPR (see section 5.3.3) in which the parameter estimates are calibrated in a point-wise way solving a set of equations to maximise the first derivation of the weighted log-likelihood of the model:

$$\ln L(\lambda_{ij}) = \sum_{ij}(-\lambda_{ij} + T_{ij} \ln \lambda_{ij} - \ln T_{ij}!)\ W_{u_{ij}} \tag{6.4}$$

where $W_{u_{ij}}$ indicates the weight of flow ij according to the proximity of its origin i to the calibration point u.

In both the Gaussian and Poisson models, different kernel types can be used for weighting the flows such as the Gaussian function. The Gaussian kernel for origin-focused spatial interaction model can be formulated as:

$$W_{u_{ij}} = exp\ [\ -\frac{1}{2}\ (\frac{d_{ui}}{b})^2\ ] \tag{6.5}$$

where d_{ui} is the geographical distance between the calibration point u and the centroid of origin i of the observed flow ij, and parameter b is bandwidth. If u and i coincide, (i.e. the calibration point is one of the existing origins), the weight of that observed flow with origin i will be unity and the weighting of other flows will decrease according to a Gaussian curve as the distance between i and other origins increases. Another possible weighting kernel is the squared Cauchy function which has the following formulation:

$$W_{u_{ij}} = (1 + \frac{d_{ui}^2}{b^2})^{-2} \tag{6.6}$$

where d_{ui} is again the geographical distance between the calibration point u and the origin of flow i, and b is a bandwidth parameter. Figure 6.2 shows a comparison of the squared Cauchy and the Gaussian kernel functions. The squared Cauchy kernel is similar to the Gaussian kernel but gives slightly less weight to close points and more weight to distant points. This second property is important if only a few data points are located within the kernel; the squared Cauchy kernel allows for fitting the model using the data outside of bandwidth distance. Using this function, the local model parameters are expected to tend towards the global parameters in the case of few data inside the kernel (Nakaya, 2001). Spatial interaction data are frequently sparse with only a few origin-destination pairs having considerable numbers of flows. This makes the squared Cauchy function an interesting kernel for geographical weighting of spatial interaction datasets; so the Cauchy spatial kernel will be used in this thesis in different GWSI approaches for obtaining weights for spatial interaction flows. For both Gaussian and squared Cauchy kernel, the bandwidth parameter can be estimated using the AICc criteria, as outlined in sections 5.3.2 and 5.3.3 (see equations 5.11 and 5.16 for the formal representation of the AICc).

6.2.1 Application of the origin-focused GWSI approach

In order to illustrate the origin-focused GWSI approach, we apply our approach to the journey-to-work dataset for the agglomeration of Lausanne. In the following equation, T_{ij} indicates total flows from origin i to destination j, including the intra-zonal flows. The model variables are defined as in previous chapters; P_i represents the active population living in origin i, N_j shows the number of jobs in destination j and d_{ij} is a variable indicating the distance between i and j calculated with the density-based scattering

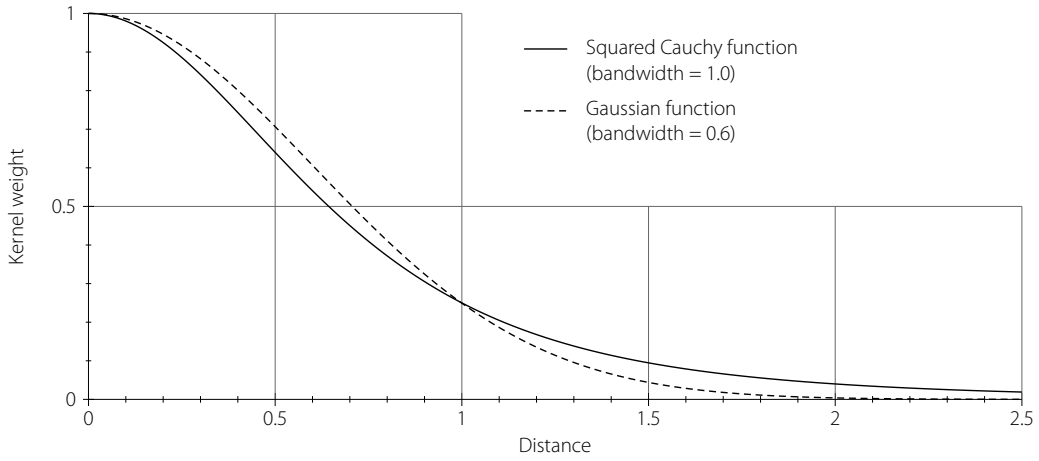


Figure 6.2: Comparison between squared Cauchy and Gaussian kernel functions.

method using population density data as explained in section 4.3.2, and κ_i, α_i, γ_i and β_i are parameters of the model at i to be estimated:

$$T_{ij} = \kappa_{\{i\}} \; P_i^{\alpha_{\{i\}}} \; N_j^{\gamma_{\{i\}}} \; d_{ij}^{\beta_{\{i\}}} \tag{6.7}$$

Considering equations 6.3 and 6.7, the Poisson version of the origin-focused model for our case study of journey-to-work in the agglomeration of Lausanne can be formulated as follows:

$$\lambda_{ij} = \exp \; (\kappa_{\{i\}} + \alpha_{\{i\}} \ln P_i + \gamma_{\{i\}} \ln N_j + \beta_{\{i\}} \ln d_{ij}) \tag{6.8}$$

where λ_{ij} indicates the flow between i and j and parameters κ_i, α_i, γ_i and β_i will be calibrated for each origin commune i considering a cluster of nearby origins. Following the considerations in the previous section, the squared Cauchy function is used as a weighting function (see equation 6.6). The first step in applying the geographically weighted approach to spatial interaction is to find an optimal bandwidth to be used in the weighing function. For this we apply the AICc approach (see equation 5.16) for selecting an optimal fixed bandwidth for the agglomeration of Lausanne when the model is Poisson and origin-focused. Figure 6.3 shows the plotted values of the bandwidth in metres against the deviance, AICc and BIC scores. The optimal bandwidth occurs where the AICc score is minimum; in this case equal to 100 metres. Using BIC as criteria, the optimal bandwidth would be 500 metres.

Following the GWSI methodology, the selected bandwidth is used to set a spatial kernel around each calibration point (i.e. each origin) to weight the observed flows within the kernel. Repeating this procedure for all origins within the agglomeration of Lausanne, table 6.1 (bottom) shows the results of the geographically weighted Poisson origin-focused approach using the selected bandwidth (100 metres). The parameters, t-

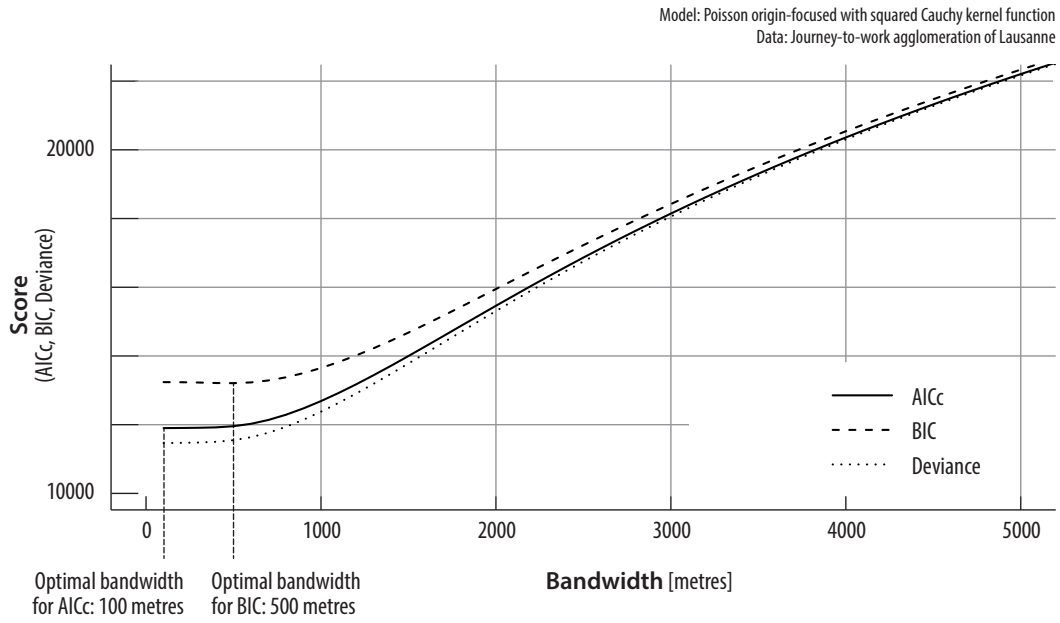


Figure 6.3: Bandwidth value (fixed) against AICc, BIC and deviance for Poisson origin-focused GWSI in Lausanne agglomeration, using a squared Cauchy weighting function.

values and p-values are illustrated with a mean, standard deviation, minimum, maximum and quartiles 25%, 50% (= median) and 75%. In order to compare the results of the local origin-focused model with a global one, results of the Poisson origin-specific model calibrated in the previous chapter are listed at the top of table 6.1. The deviance and Pseudo R^2 of the models are calculated as measures of goodness-of-fit. It should be noted that since the number of observations and variables in the two models are different, a direct comparison of the models cannot be considered. For instance, the deviance value is expected to increase in origin-focused model which has more variables than the origin-specific model. A reminder here is needed to clarify the difference between the origin-focused and origin-specific models. As was shown in section 5.4.1, an origin-specific model for our journey-to-work case study can be formulated as follows:

$$\lambda_{ij} = \exp\ (\kappa_i + \gamma_i \ln N_j + \beta_i \ln d_{ij}) \tag{6.9}$$

where any variable associated with the origin i (e.g. in this case P_i) is a constant and can integrated into the balancing factor. In other words, in the calibration process of an origin-specific model for origin i, only information on destinations and the travel cost (distance) are considered and information on origins is ignored. However, in the origin-focused GWSI considering a kernel around i, a cluster of nearby origins are involved in the calibration process so that the equation of the origin-focused model covers all the variables in the model and can provide an estimated parameter value for the origin propulsiveness variable (i.e. here the number of active population in each origin).

Table 6.1: Poisson origin-specific and origin-focused GWSI models

Origin-specific model							
Parameter	*Mean*	*Min*	*Max*	*Std Dev*	*Quartiles*		
					25%	*50%*	*75%*
Jobs	1.0598	0.7645	1.2710	0.1032	1.0040	1.0599	1.1194
Distance	-1.8247	-2.7482	-0.8144	0.4180	-2.0740	-1.8291	-1.5395
Intercept	9.7697	3.9305	19.2064	2.9938	7.5179	9.9179	11.6682
t-values jobs	42.2	5.7	168.9	30.0	22.5	31.8	52.7
t-values distance	-32.9	-100.5	-7.8	16.8	-41.1	-29.5	-19.9
t-values intercept	28.4	7.7	69.5	13.3	19.4	24.9	34.3
p-values jobs	0.00	0.00	0.00	0.00	0.00	0.00	0.00
p-values distance	0.00	0.00	0.00	0.00	0.00	0.00	0.00
p-values intercept	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deviance	172.4	-182.7	1390.9	211.5	65.5	105.7	202.6
Pseudo R^2	0.8822	0.6365	0.9959	0.0740	0.8392	0.8949	0.9422

Local origin-focused GWSI							
Bandwidth	100 metres						
Parameter	*Mean*	*Min*	*Max*	*Std Dev*	*Quartiles*		
					25%	*50%*	*75%*
Active pop.	0.6647	-0.4964	1.2266	0.2739	0.4865	0.7160	0.8331
Jobs	1.0598	0.7646	1.2707	0.1027	1.0026	1.0599	1.1206
Distance	-1.8235	-2.7482	-0.8145	0.4152	-2.0754	-1.8291	-1.5351
Intercept	5.2193	-6.0886	17.6987	4.8164	1.7744	4.9432	8.2181
t-values active pop.	0.5	-0.0	6.0	1.0	0.1	0.2	0.4
t-values jobs	85.6	1.5	2019.9	277.5	12.6	24.8	47.2
t-values distance	-71.8	-1519.1	-2.1	232.7	-41.7	-20.1	-10.8
t-values intercept	0.2	-3.0	1.5	0.6	0.1	0.2	0.4
p-values active pop.	0.76	0.00	0.99	0.25	0.68	0.87	0.93
p-values jobs	0.00	0.00	0.12	0.01	0.00	0.00	0.00
p-values distance	0.00	0.00	0.03	0.00	0.00	0.00	0.00
p-values intercept	0.76	0.00	0.99	0.22	0.70	0.81	0.92
Deviance	-12373.3	-417006.2	528320.6	149530.9	-98029.4	8212.7	76164.2
Pseudo R^2	0.8447	-0.2735	0.9888	0.2025	0.8597	0.9008	0.9396

Maps of local parameter estimates of the Poisson origin-focused GWSI model over all communes in Lausanne are given in figure 6.4. Estimated parameters from both the origin-focused and origin-specific models are extremely similar for both distance-decay and number of jobs parameters (see figure 5.2 for origin-specific model). The average for the distance parameter is -1.8235 for the origin-focused model, and -1.8247 for the origin-specific model. The spatial distribution of the estimated distance-decay parameters across Lausanne shows a clear pattern with a concentric gradient. The communes around Renens, west of the city of Lausanne, have the least negative values, (i.e. the active population living in these communes consider distance to be less of a deterrent for their daily travel to work than people farther away). It is interesting to note that the deterrence of distance on interaction is higher in peripheral communes than in the more central zones in the west of Lausanne. One possible explanation for this can be better accessibility, which is especially so in the west of Lausanne with a high proximity of all three motorway branches. A good transportation infrastructure allows workforces to travel longer distances in the same amount of time to different destinations. In both the origin-focused and origin-specific models, the average for the number of jobs parameter is 1.0598. The city of Lausanne and some communes at the border of the agglomeration have higher jobs parameters. Interestingly, important commercial centres and branches of international companies are located at the western end of the agglomeration which might explain the higher parameter values for the number of jobs in this area. The eastern end of the agglomeration, bordering the protected zone of Lavaux, has the lowest parameter values for the number of jobs. The Lavaux is one of the UNESCO world heritage sites, whose vineyards are protected from further development. These communes are known for being mainly residential with higher house prices than average.

The reason for these similar values for the distance-decay and number of jobs parameters can be the fact that the origin-focused model has a very small bandwidth of only 100 metres. In this situation, the flows from the same origin have maximum weight of 1 and all other flows smaller weights, making the model similar to an origin-specific model. One difference between the two models is that the origin-focused model contains a parameter for the active population while in the origin-specific model this parameter is omitted. Interestingly, the p-values for this parameter (with mean p-value of 0.76) for most origins indicate that the estimated parameter values are not significantly different from 0, showing no significant contribution of this parameter in the model. This can be also seen in the map of t-values of the parameters in figure 6.5. In this case, the origin-focused model tends towards an origin-specific model, both through the selection of a small optimal bandwidth and non significant parameter estimates for the active population parameter. However, it should be noted that comparing the Pseudo R^2 of the origin-focused and the origin-specific models shows a better score for the origin-focused model with a value of 0.9008 compared to 0.8949 in median (quartiles 50%). This indicates that in most origins, the local origin-focused approach shows improvement over the

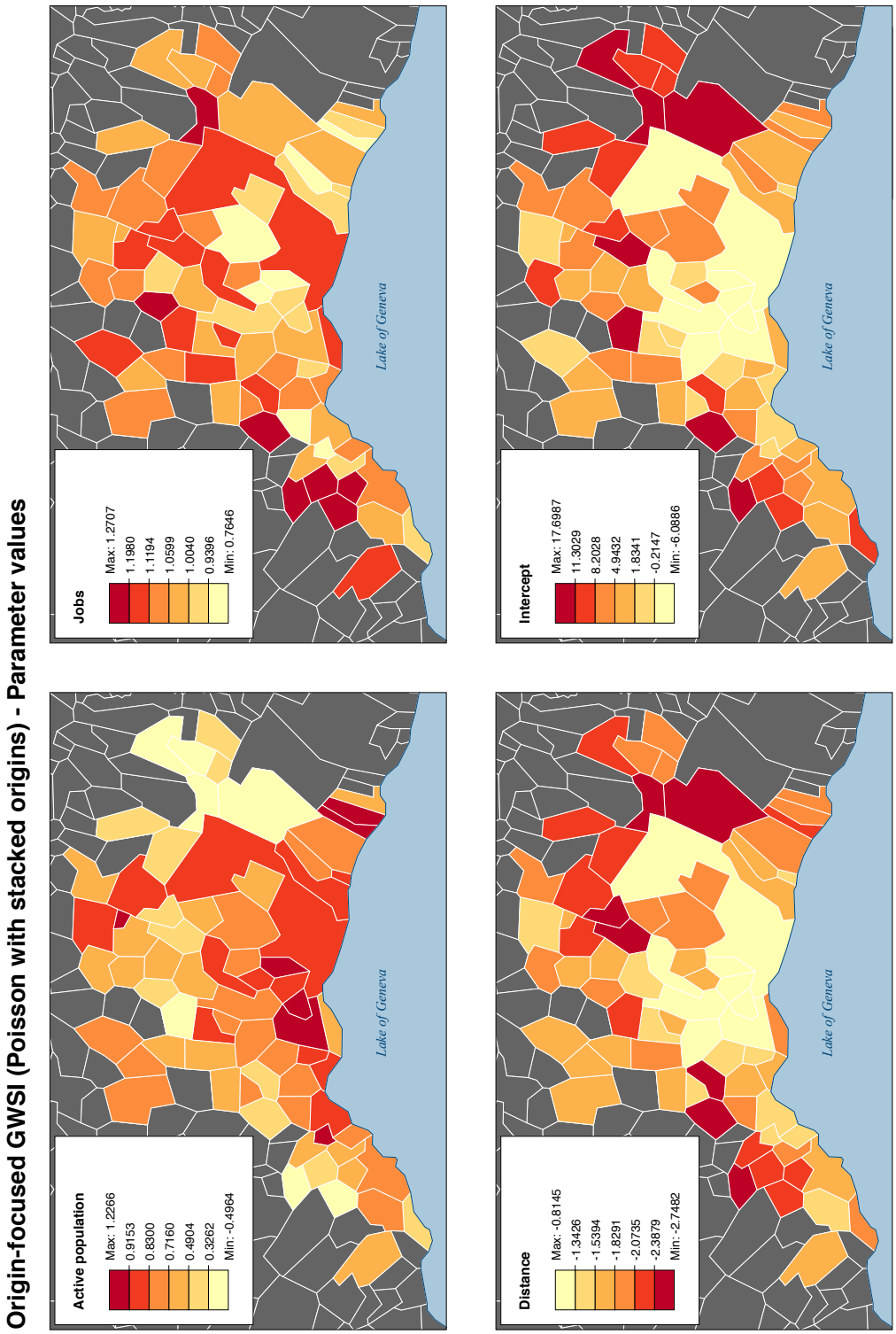


Figure 6.4: The parameter estimates of the Poisson origin-focused GWSI model over all communes of the Lausanne agglomeration.

origin-specific model.

6.3 Destination-focused GWSI approach

As discussed in section 6.2, by using the concept of geographical weighting for calibrating spatial interaction models, a local set of parameter estimates can be obtained for each arbitrary calibration location within the region. In GWSI, a spatial kernel is defined around each calibration point, all observed flows (i.e. data points in regression) falling within this kernel will contribute in the model calibration process. In this section we focus on the destination of the flows. The main approach of GWSI for destinations is similar to the origin-focused approach, however here a cluster of destinations of the observed flows around the calibration point are considered within the kernel. The observation flows are weighted according to the proximity of their destinations to the calibration point in which greater weights are assigned to the flows with closer destinations to the calibration point. As for each calibration point, the nearby observations will be weighted differently so the results of the model calibration are unique to the particular location. We call this scenario the destination-focused approach. The destination-focused GWSI procedure can be repeated for all calibration points within the region, possible variability in the calibrated parameter estimates can be visualised when these values are mapped over space. Figure 6.6 illustrates the destination-focused scenario where observed flows are weighted according to distance of their destinations to the calibration point. Again here the width of the flows represents the influence of that flow on the calibration process, i.e. wider flows have larger weights and higher effect on the calibration process. The distances between the calibration points and destinations are shown with dashed lines.

The general mathematical formula for a gravity destination-focused GWSI model can be written as follows:

$$T_{ij} = \kappa_{\{u(x,y)j\}}\ v_i^{\alpha_{\{u(x,y)j\}}}\ w_j^{\gamma_{\{u(x,y)j\}}}\ d_{ij}^{\beta_{\{u(x,y)j\}}} \tag{6.10}$$

where (x, y) are the coordinates of the calibration point u, (i.e. one of the existing destinations or any other point within the study region), v_i, w_j and d_{ij} are the model variables (i.e. the origin propulsiveness, attractiveness of destination and distance between origin i and destination j respectively), and κ, α, γ and β are parameters of the model to be calibrated. Similar to equation 6.1, the above formula can be calibrated using the WLS method as shown in equation 6.2 and using the iteratively re-weighted least squares when the model is Poisson to arrive at a maximum likelihood model fit (see equations 6.3 and 6.4). A weighting function can be used to assign weights to the observation flows. The Cauchy kernel for the destination-focused GWSI approach can be formulated as follows:

$$W_{u_{ij}} = (1 + \frac{d_{uj}^2}{b^2})^{-2} \tag{6.11}$$

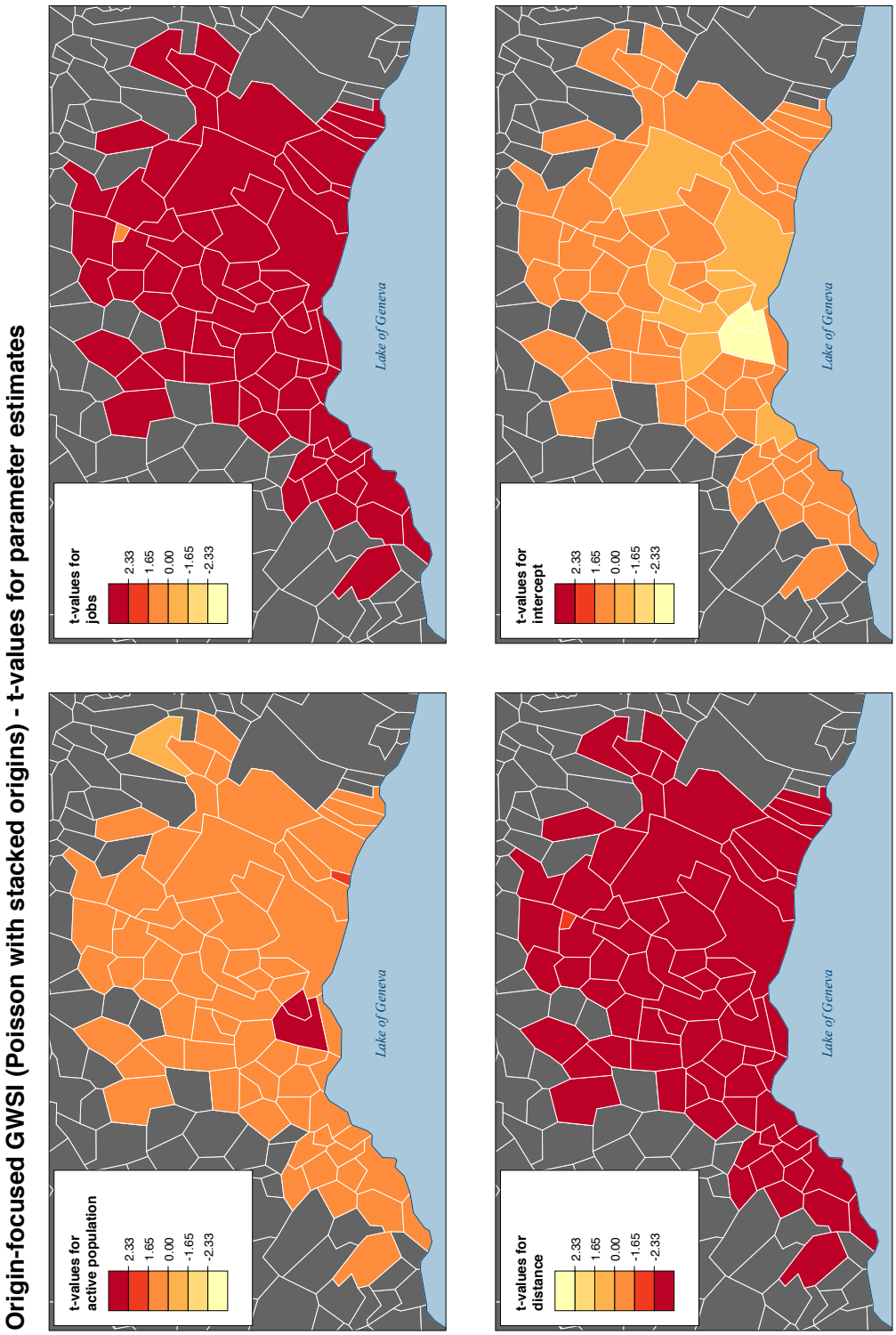


Figure 6.5: The t-values of parameters of the Poisson origin-focused GWSI model over all communes of the Lausanne agglomeration. These maps display absolute t-values, where values greater than 2.33 are significant at a level of 99%, and values greater than 1.65 are significant at a level of 95%. For negative parameter values (for the distance decay parameter), the negative t-values of -2.33 and -1.65 correspond to the significance levels of 99% and 95% respectively.

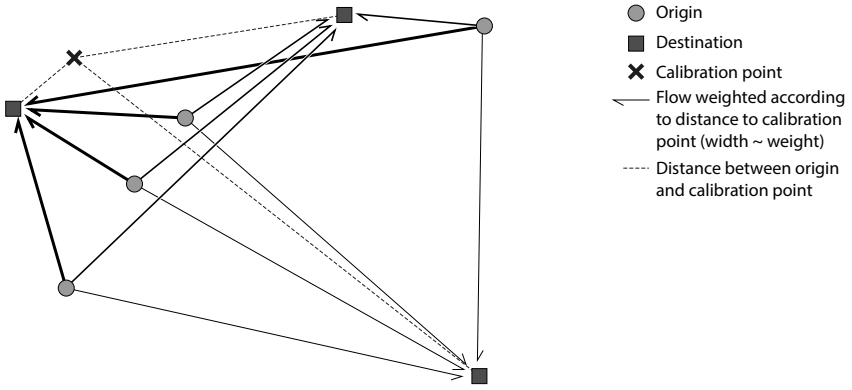


Figure 6.6: A schematic overview of destination-focused GWSI approach.

where d_{u_j} is the geographical distance between the calibration point u and destination j of the observed flow ij. The bandwidth parameter, b, can be estimated with the AICc or BIC methods as discussed in section 5.3.3.

6.3.1 Application of the destination-focused GWSI approach

In this section we apply the destination-focused GWSI approach on the journey-to-work data in Lausanne. The following formula shows the Poisson destination-focused spatial interaction model that we calibrate for each destination j within the agglomeration of Lausanne:

$$\lambda_{ij} = \exp(\kappa_{\{j\}} + \alpha_{\{j\}} \ln P_i + \gamma_{\{j\}} \ln N_j + \beta_{\{j\}} \ln d_{ij}) \tag{6.12}$$

where λ_{ij} are the flows between i and j, P_i represents the number of active population in each origin, N_j indicates the number of jobs in each destination, d_{ij} is the distance between i and j calculated with density-based scattering method using the population density (see section 4.3.2), and κ_j, α_j, γ_j and β_j are parameters of the model to be estimated for each calibration commune j considering a cluster of nearby destinations.

The optimal bandwidth has been estimated using the AICc and BIC approaches. The calculated bandwidth values against the AICc, BIC and deviance scores for the destination-focused approach are shown in figure 6.7. The optimal bandwidth value that could be calculated with both AICc and BIC is equal to 100 metres. The results of the destination-focused approach for all communes in Lausanne using the calculated optimal bandwidth are shown in table 6.2. The results of a destination-specific model are also shown in the top part of table 6.2 for a comparison with local results of GWSI. The parameters of the models show the expected effect on the overall interaction in which the distance shows a negative effect and the active population and the number of jobs are related to the interaction in a positive relationship. Similarly to section 6.2.1, a comparison between destination-specific and destination-focused models needs to be done

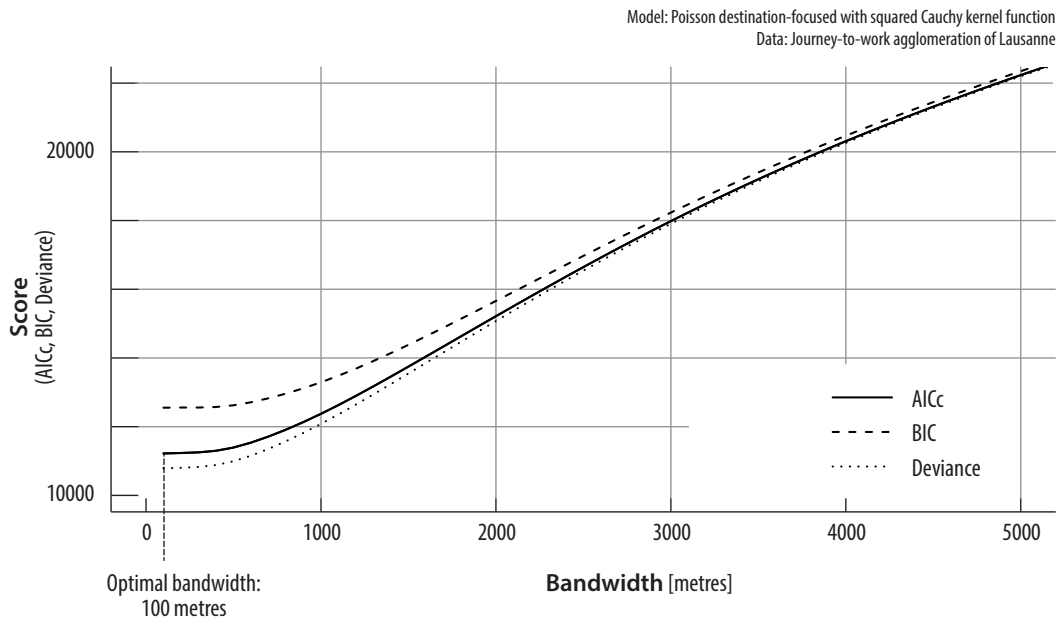


Figure 6.7: Bandwidth value (fixed) against deviance, AICc and BIC score for Poisson destination-focused model in Lausanne agglomeration, using a squared Cauchy weighting function.

carefully as both models do not include the same number of variables. The destination-specific model ignores the destination attractiveness information and only origins and distance attributes are considered in the model's formula.

While table 6.2 shows the summary statistics for the destination-specific and local destination-focused models, a map of the spatial variation of the model's parameters is shown in figure 6.8 for the destination-focused GWSI. This map can be compared with one of the destination-specific model presented in figure 5.5. The t-values of the model's parameters are mapped in figures 6.9 and 5.6 for the destination-focused and the destination-specific models respectively. The t-values of the parameters of the destination-focused GWSI, except the intercept parameter, show higher values than equivalents from the destination-specific model; in the former, parameter values seem to be significantly different from 0, while in the latter this is not always the case.

For both destination-specific and destination-focused GWSI models, all the distance-decay parameters are negative but the destination-specific model shows more negative parameter value for communes on average. The spatial pattern of distance-decay parameters in the destination-focused model is similar to the origin-focused GWSI (see section 6.2.1), in the sense that we can find a gradient from the central communes of the agglomeration showing the least negative values towards the periphery with larger negative values. This pattern indicates that the deterrence of distance to interaction is less for people commuting to the central communes. This can be explained by a better transportation system in the central communes (e.g. Lausanne city or Renes) which

Table 6.2: Poisson destination-specific and destination-focused GWSI models

Destination-specific models							
Parameter	*Mean*	*Min*	*Max*	*Std Dev*	*Quartiles*		
					25%	*50%*	*75%*
Active population	0.8830	-0.7594	1.8510	0.3053	0.8000	0.8809	1.0068
Distance	-1.9847	-4.4205	-0.5854	0.6904	-2.4184	-1.8805	-1.4386
Intercept	11.5903	3.8253	23.3434	4.0070	8.4200	10.9907	14.4038
t-values active pop.	26.4	-0.6	204.1	31.1	9.0	19.1	31.9
t-values distance	-30.0	-102.2	-1.4	19.3	-41.0	-24.4	-15.2
t-values intercept	26.7	4.5	70.8	12.0	17.8	24.9	33.7
p-values act. pop.	0.02	0.00	0.55	0.08	0.00	0.00	0.00
p-values distance	0.00	0.00	0.17	0.02	0.00	0.00	0.00
p-values intercept	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deviance	153.1	-62.9	1620.3	218.8	43.8	95.8	182.4
Pseudo R^2	0.8620	0.5427	0.9918	0.0808	0.8232	0.8771	0.9207

Local destination-focused GWSI							
Bandwidth	100 metres (fixed)						
Parameter	*Mean*	*Min*	*Max*	*Std Dev*	*Quartiles*		
					25%	*50%*	*75%*
Active pop.	0.7852	0.6614	0.8651	0.0468	0.7646	0.7935	0.8208
Jobs	0.9136	0.7912	1.0324	0.0568	0.8769	0.9083	0.9564
Distance	-1.3775	-1.7612	-1.0372	0.1724	-1.4957	-1.3943	-1.2526
Intercept	1.8216	-2.5062	6.7314	2.2585	0.2119	2.0878	3.5218
t-values active pop.	126.4	36.2	281.3	62.7	75.6	109.0	170.7
t-values jobs	150.3	54.6	313.7	66.6	95.9	135.1	196.4
t-values distance	-114.5	-157.2	-63.9	23.6	-130.5	-113.6	-97.3
t-values intercept	10.8	-33.0	31.5	17.3	2.4	17.3	23.6
p-values active pop.	0.00	0.00	0.00	0.00	0.00	0.00	0.00
p-values jobs	0.00	0.00	0.00	0.00	0.00	0.00	0.00
p-values distance	0.00	0.00	0.00	0.00	0.00	0.00	0.00
p-values intercept	0.00	0.00	0.22	0.03	0.00	0.00	0.00
Deviance	41562.4	276.2	77733.4	24281.6	22563.4	39911.9	63838.8
Pseudo R^2	0.9532	0.9335	0.9721	0.0064	0.9508	0.9537	0.9563

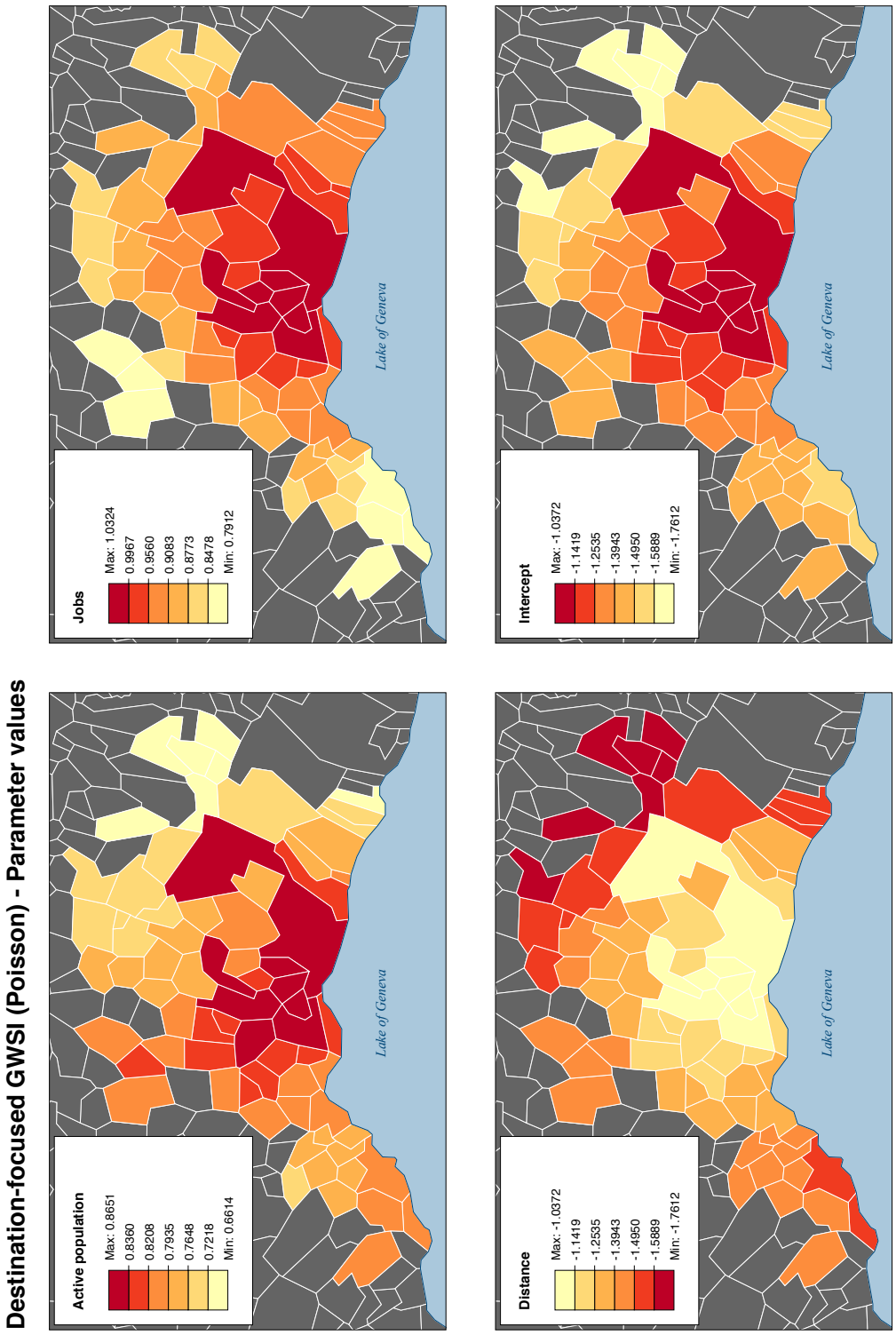


Figure 6.8: The parameter estimates for Poisson destination-focused GWSI model in agglomeration of Lausanne.

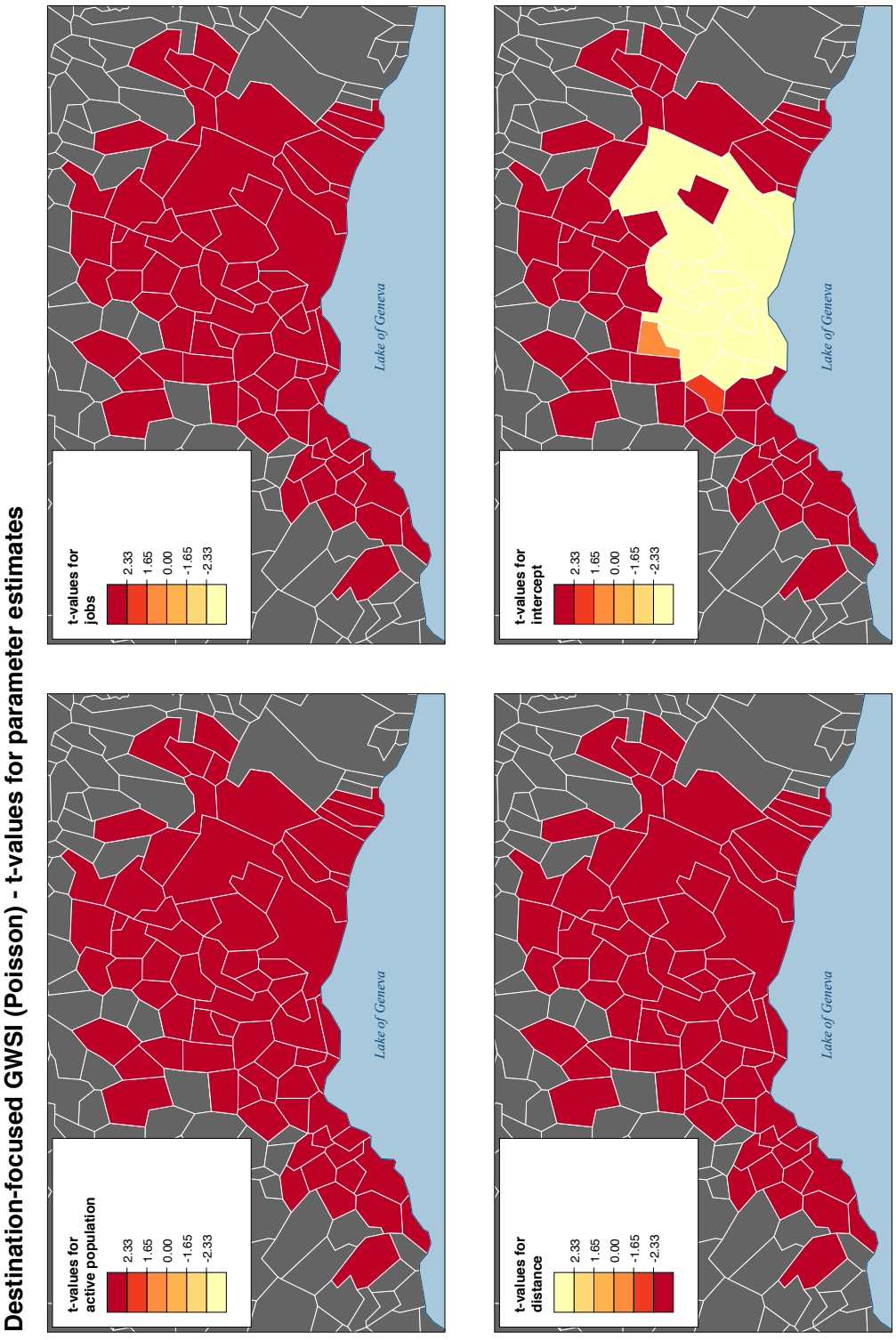


Figure 6.9: The t-values of parameters for Poisson destination-focused GWSI model in agglomeration of Lausanne. These maps display absolute t-values, where values greater than 2.33 are significant at a level of 99%, and values greater than 1.65 are significant at a level of 95%. For negative parameter values (for the distance decay parameter), the negative t-values of -2.33 and -1.65 correspond to the significance levels of 99% and 95% respectively.

workforces can access more easily from different communes compared with less accessible destinations located in high elevation areas (e.g. Servion or Carrouge). In comparison with the destination-specific model, distance-decay parameters of destination-focused GWSI show smoother variations in space (i.e. in the destination-specific model, the variation of distance parameters is larger between neighbouring communes resulting in some places exhibiting a salt-and-pepper pattern with abrupt changes). For instance, the distance-decay parameter from the destination-specific model for Jouxtens-Mézery, a central commune located west of Lausanne city, shows a large difference compared with other neighbouring communes. Comparison of t-values of the distance-decay parameter between models shows a higher range of values in destination-focused GWSI (with a mean of -114.5) indicating more significance of the distance-decay parameter estimates obtained with this model in comparison with those from a destination-specific model (with a mean of -30.0). The t-value of distance-decay parameter ranges between -157.2 and -63.9 in destination-focused GWSI while a much lower range is shown in the destination-specific model between -102.2 and -1.4.

The local values of the active population parameter vary between 0.66 and 0.87 in the destination-focused GWSI model, with an average value of 0.79. The spatial pattern shows higher values for the central communes. Some smaller secondary centres at the periphery, such as Cossonay in the north-west and Aubonne in the west, also have slightly higher parameter values for the active population compared to their respective neighbours. This shows that although active population has a general positive effect on total interaction, this effect is larger in the central communes of the agglomeration. Overall, the destination-focused GWSI yields a smooth spatial pattern without large variations between neighbours, with a relatively small range in the resulting parameter values. The parameter values for the active population obtained using the destination-specific model are less homogenous. They vary between -0.76 and $+1.85$. Negative parameter values for the active population are rather unusual. Some of the smallest parameter values occur in the communes of Villette, Jouxtens-Mézery, Villars-Tiercelin, Malapalud or Chigny. The t-values for these communes in the destination-specific model are close to 0, indicating that these parameter values are not significant. This result can be explained by the fact that these communes are typically small and rather residential. In order to obtain a better result for these communes, we typically need to pull in information from the neighbouring communes; this is achieved through the use of the destination-focused GWSI where a cluster of destinations is considered in the calibration process. The improvement of t-values of the active population parameter in destination-focused GWSI shows more significance of this parameter estimate in comparison with those obtained with the destination-specific model.

The destination attractiveness variable, (i.e. number of jobs), is only considered in the destination-focused GWSI model since in the calibration of a destination-specific model information on the neighbouring destination attributes is ignored (see the general

formula of a destination-specific model in equation 5.21). In comparison to a destination-specific model, in a destination-focused GWSI where a cluster of destinations around the calibration communes is considered, an estimated parameter value for the destination attractiveness (i.e. number of jobs) variable *can* be obtained. The parameter values for this variable for the Lausanne agglomeration vary between 0.79 and 1.03 with an average value of 0.91. The spatial pattern shows a clear gradient from the central communes where the highest values are located, towards the periphery with the lowest values. The t-values of the number of jobs parameter of the destination-focused GWSI show significantly higher value from 0 indicating significance of the local parameters (see figure 6.9).

Goodness-of-fit tests for the models are undertaken based on Pseudo R^2 and deviance values listed in table 6.2. The Pseudo R^2 is on average higher for the destination-focused GWSI model showing a better fit compared to the destination-specific model. For the destination-specific model, the Pseudo R^2 has an average of 0.86 and median of 0.88 compared to average of 0.95 and median of 0.95 for the destination-focused variant. Comparison of deviance of the models however is not straightforward because deviance depends on the number of observations and number of parameters in the model. The deviance is higher in the destination-focused GWSI, probably due to the fact that this model includes more variables, parameters and numbers of observations in each regression compared with the destination-specific model. In general, comparing results of the models shows the destination-focused GWSI provides stable and smooth regression results even in locations where enough information for a destination-specific model is lacking, preventing large variations in parameter values between neighbouring communes.

The results of the above sections in this chapter have shown that the origin- and destination-focused GWSI models can provide more local information compared with conventional origin- and destination-specific models. This is because they can be calibrated in any arbitrary location within the study region and provide local information about the situation in that location. In the following sections, we further disaggregate our analysis by applying the origin- and destination-focused approaches to each separate destination and origin in the region.

6.4 Destination-specific origin-focused GWSI approach

The origin-focused GWSI model can be calibrated for each specific destination in the system, providing a destination-specific version of the origin-focused GWSI model. In this scenario, a cluster of origins is considered around the calibration point, as in the origin-focused approach. The calibration point can be an existing origin interacting with the destination j or any arbitrary location within the study area. The weighting scheme is similar to the one in the origin-focused approach in which a kernel is considered around the calibration point u and observation flows within this kernel are weighted according to the proximity of their origins to the calibration point. However, in comparison to the

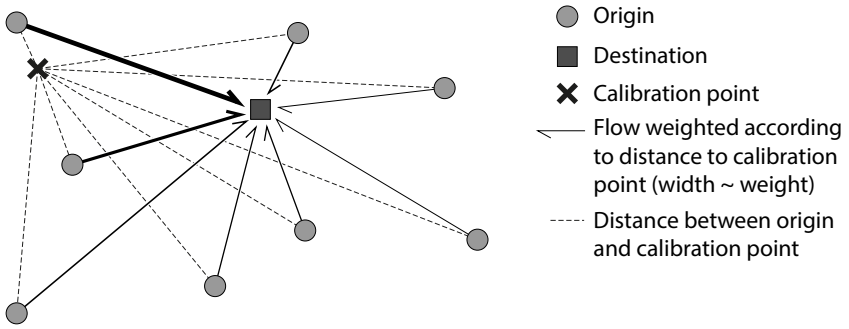


Figure 6.10: A simplified illustration of the destination-specific origin-focused GWSI approach.

origin-focused GWSI approach, in the destination-specific origin-focused approach, only flows *terminating in a specific destination j* are considered in the model. If the calibration process with this approach is repeated for all arbitrary locations within the study area, a surface of local parameters specific to all destinations (i.e. origin propulsiveness and distance parameters) will be obtained. Figure 6.10 shows a simplified illustration of the destination-specific origin-focused approach where the width of the flows represents their weights and therefore their influence on the calibration process.

The following formula shows a general equation for the proposed destination-specific origin-focused GWSI in the form of a gravity-type spatial interaction model:

$$T_{ij} = \kappa_{(\{ui\},j)}\ v_i^{\alpha_{(\{ui\},j)}}\ d_{ij}^{\beta_{(\{ui\},j)}} \tag{6.13}$$

where T_{ij} represents the number of flows from different origins i to destination j, v_i and d_{ij} are variables of origin propulsiveness and distance between regions respectively, $\kappa_{(\{ui\},j)}$, $\alpha_{(\{ui\},j)}$ and $\beta_{(\{ui\},j)}$ are parameters of the model to be estimated, specific to destination j by considering a cluster of origins around the calibration point u. This model equation can be calibrated similarly to origin- and destination-focused GWSI models through linear regression for a Gaussian model and as a GLM when the interaction model is considered to be Poisson. If the model is considered to follow a Poisson distribution, the calibration process can be undertaken similarly to the origin-focused GWSI explained in section 6.2.1 using geographically weighted Poisson regression in which the maximum likelihood estimates are obtained by iteratively re-weighted least squares using the following equation:

$$\ln L(\lambda_{ij}) = \sum_{ij}(-\lambda_{ij} + T_{ij}\ln\lambda_{ij} - \ln T_{ij}!)\ W_{u_{ij}} \tag{6.14}$$

where $\lambda_{ij} = \exp(\kappa_{uj} + \alpha_{uj}\ln v_i + \beta_{uj}\ln d_{ij})$. To obtain the flow weights in matrix $W_{u_{ij}}$

in above equation 6.14, a Cauchy function can be used as a spatial kernel:

$$W_{u_{ij}} = \left[1 + \left(\frac{d_{(ui)j}}{b}\right)^2\right]^{-2} \tag{6.15}$$

where $W_{u_{ij}}$ indicates the weight of flow ij according to the proximity of its origin i to the calibration point u, (i.e. parameter $d_{(ui)j}$ shows the distance between u and origin i connected to a specific destination j). Parameter b is the kernel bandwidth and can be calculated using the AICc or BIC methods as before.

6.4.1 Application of the destination-specific origin-focused GWSI approach

To illustrate the use of the proposed destination-specific origin-focused GWSI approach, we apply this model on the journey-to-work dataset of Lausanne considering the following Poisson gravity model:

$$\lambda_{ij} = \exp(\kappa_{(\{i\},j)} + \alpha_{(\{i\},j)} \ln P_i + \beta_{(\{i\},j)} \ln d_{ij}) \tag{6.16}$$

where the variables P_i and d_{ij} are defined as before (i.e. P_i represents active population living in origin i and d_{ij} is distance between regions i and j), parameters $\kappa_{(\{i\},j)}$, $\alpha_{(\{i\},j)}$ and $\beta_{(\{i\},j)}$ are calibrated for origin communes in the agglomeration that are interacting with a specific destination j. In order to provide a local set of parameter estimates specific to each destination within the agglomeration, the general origin-focused GWSI approach should be repeated for each commune. This requires the estimation of the bandwidth parameter for each destination separately. We have calculated the optimum bandwidth values for each commune using the AICc method. In order to visualise the spatial variation of the bandwidth values across the agglomeration, the resulted bandwidth values are mapped for each specific destination commune, as shown in figure 6.11. In this figure, the minimum, maximum and percentiles 10%, 25%, 50%, 75% and 90% of the bandwidth are shown in the legend of the map. The bandwidth values vary between 300 metres for central communes to 12 kilometres for some small bordering communes. This pattern of bandwidth values is interesting since it shows how the GWSI model considers a smaller bandwidth in central area where the flow data are dense and the spatial kernel opens to have a bigger bandwidth when the flow data are sparse in the small bordering communes.

In a destination-specific origin-focused GWSI where the calibration process is repeated for all m origin communes connected to the specific destination j, m sets of local parameter estimates result for each destination. Using the selected optimum bandwidth for each commune, we have computed the local parameter estimates of the model for all 70 destinations within the agglomeration. To facilitate visualisation and analysis of the results, we have plotted each of these parameters in a 70×70 matrix. Figures 6.12 and 6.13 show the values of the distance-decay and active population parameters respectively. The origin communes are listed in the bottom columns of the matrices while

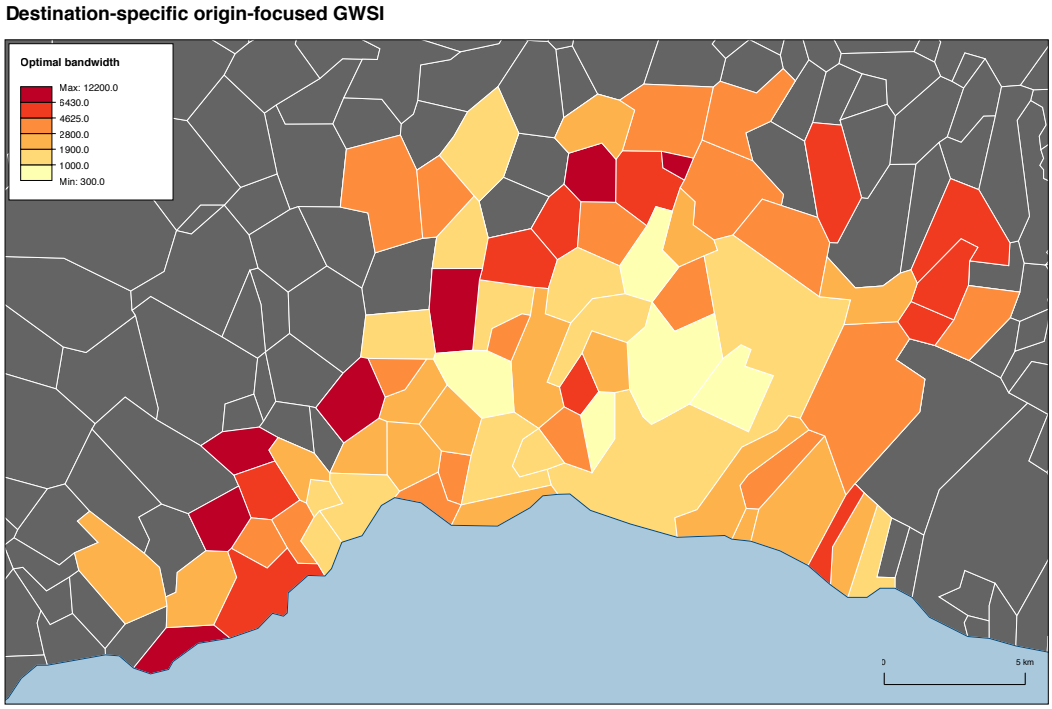


Figure 6.11: Optimal bandwidth for destination-specific origin-focused GWSI models.

destination communes are shown in rows. Both origins and destinations are placed in decreasing order of the total number of outgoing flows; as a result, the bigger communes are at the top and left of the matrices (e.g. see position of the city of Lausanne as the biggest commune). As mentioned in section 3.12, different approaches exist for ordering the rows and columns of the matrix visualisation giving raise to different visual patterns. In order to simplify the identification of the different communes for interpretation purposes, we have chosen to keep the order between all matrix visualisation constant, in decreasing order of size. As a result, important flows appear at the top of the matrix visualisation. For each specific destination j in the rows of the matrices, different values in columns show the estimated parameter values at corresponding origins. For instance in the distance-decay matrix shown in figure 6.12 consider the destination Lutry (10th destination from top column), variation of parameter values in different origins shows how people living at those origins consider distance as a barrier for commuting to destination Lutry.

The variations in the distance-decay parameter values range from around -3 to -1 with less negative values generally found for the larger destinations with better transportation systems and higher number of jobs. In general in the destination-specific origin-focused GWSI model for this agglomeration, distance-decay parameters of origin communes do not show much variations for most larger destinations such as Lausanne, Renens or Crissier. This is visible through the appearance of horizontal lines having nearly identical parameter values for these destinations. However, for some of the smaller

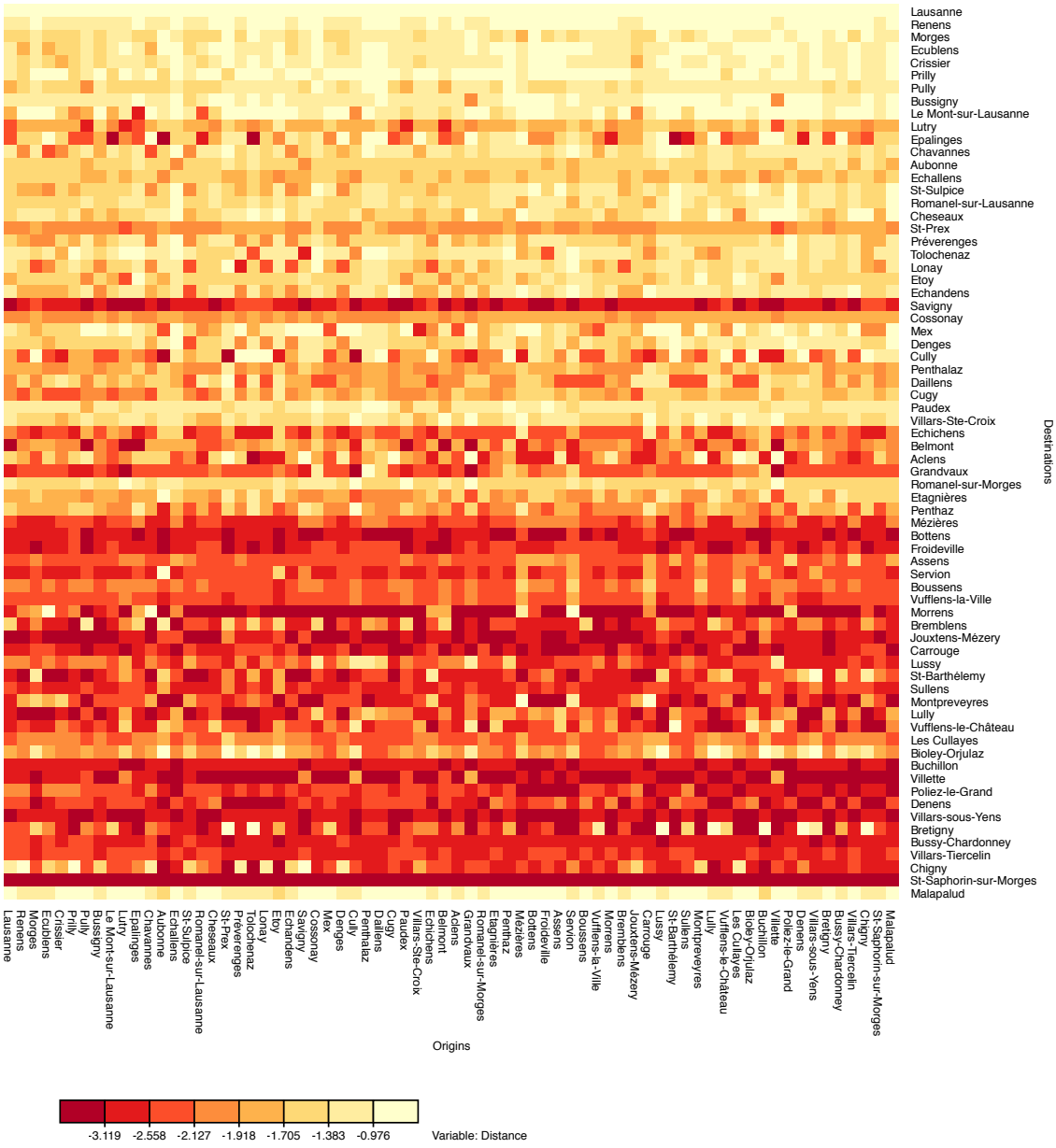


Figure 6.12: Distance-decay parameter of the destination-specific origin-focused GWSI for Lausanne agglomeration. Each row (destination) represents one destination-specific origin-focused model, calibrated separately for each origin (column), resulting in 4900 different parameter estimates.

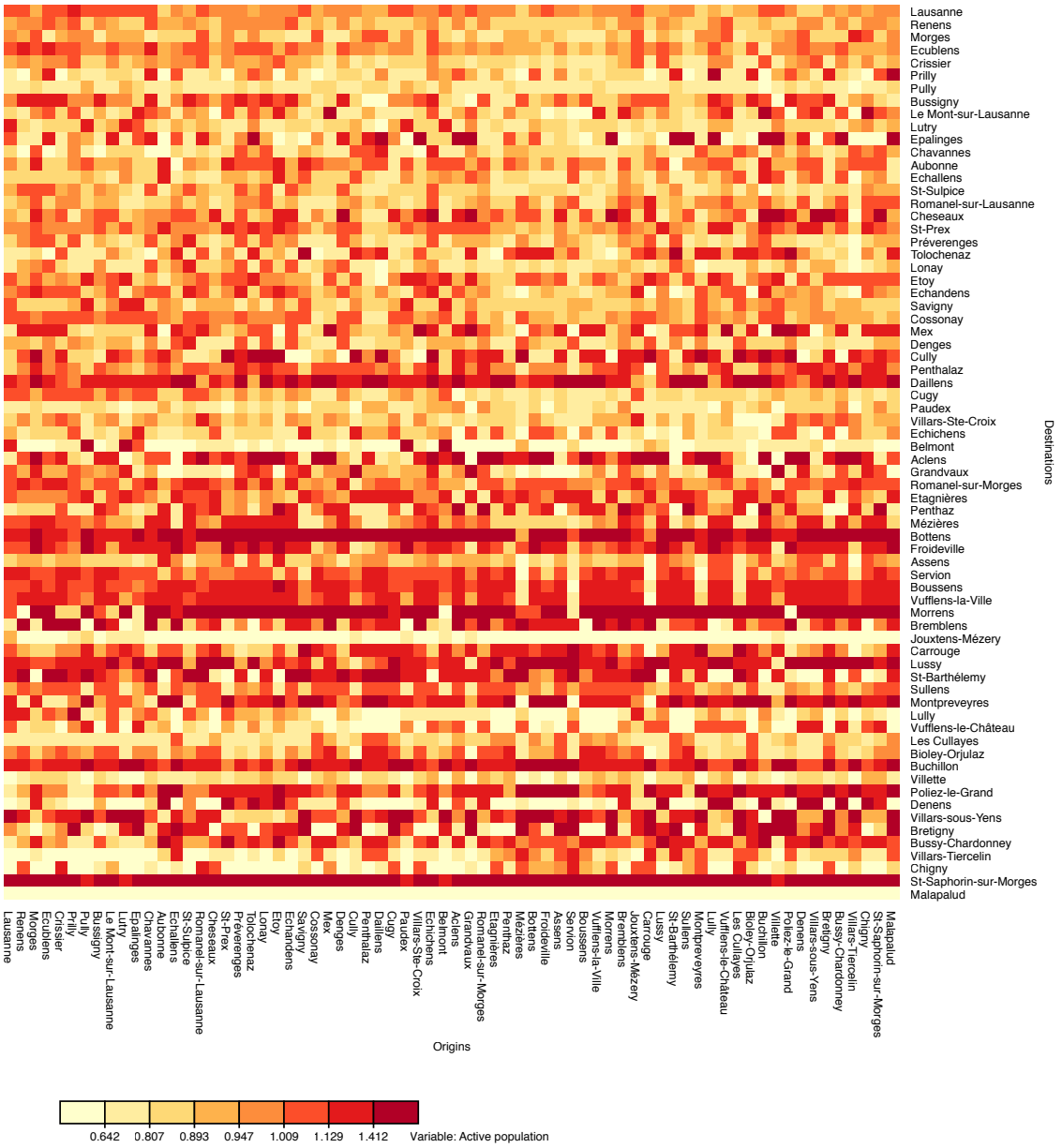


Figure 6.13: Active population parameter of the destination-specific origin-focused GWSI for Lausanne agglomeration. Each row (destination) represents one destination-specific origin-focused model, calibrated separately for each origin (column), resulting in 4900 different parameter estimates.

destinations located further down the matrix's rows, variations across the different origins exist, such as for destinations of Belmot, Mézières or Assens. The significance of the distance-decay parameters are tested by calculating t-values which are shown in figure 6.14 for all communes. Higher t-values of the distance-decay parameter for mainly medium size destination communes in the the matrix indicate the greater significance of them over parameters of the smaller and some of the larger communes in the bottom rows and top part of the matrix respectively. However, there are some exceptions such as Mex, Cully or Morrens where the t-values are close to zero. We have also calculated the p-values of the parameters. The p-values of distance-decay parameter of more than 70% of the models are still in a acceptable range, (70th percentile = 0.1), which shows the significance of the distance-decay parameters of these models with 90% confidence.

The active population parameter matrix in figure 6.13 shows variations of these parameters across different origins for each specific destination. The values of the parameters vary roughly between 0.6 and 1.4 with an overall positive effect on total interaction. In general, this parameter shows more variations over different communes compared to distance-decay. This indicates that the active population variable does not always have the same importance for each destination across all origins. However, there are some exceptions such as Saint-Saphorin-sur-Morges or Malapalud where the values of the parameters are very similar across all origins. Figure 6.15 illustrates the matrix of t-values for the active population parameter estimates of the model. In this matrix a clear gradient from the larger destinations to the smaller ones can be found. For some of the models, the t-values indicate that some active population parameters are not significantly different from 0 including ones for above examples of Saint-Saphorin-sur-Morges and Malapalud. The p-values of the active population parameter show that this parameter in more than 68% of the local models is significant (68th percentile of p-values for the active-population parameter = 0.1).

In order to compare goodness-of-fit of the destination-specific origin-focused GWSI models over different communes, Pseudo R^2 values for all models are calculated. As it is illustrated in figure 6.16, on average the models show reasonable fit with R^2 values range between roughly 0.7 and over 0.95. In most destinations in the matrix, R^2 values of the origin-focused models show few variations over different origins. Considering all results, the origin-focused GWSI provides spatially disaggregated information when the specific parameter estimates are obtained for each destination within the region, considering a cluster of origins around each calibration commune.

6.5 Origin-specific destination-focused GWSI approach

An origin-specific version of the destination-focused GWSI model can be developed when in the model calibration process only observation flows with origin i to different destinations are considered. Similar to the destination-focused approach, a weighting kernel

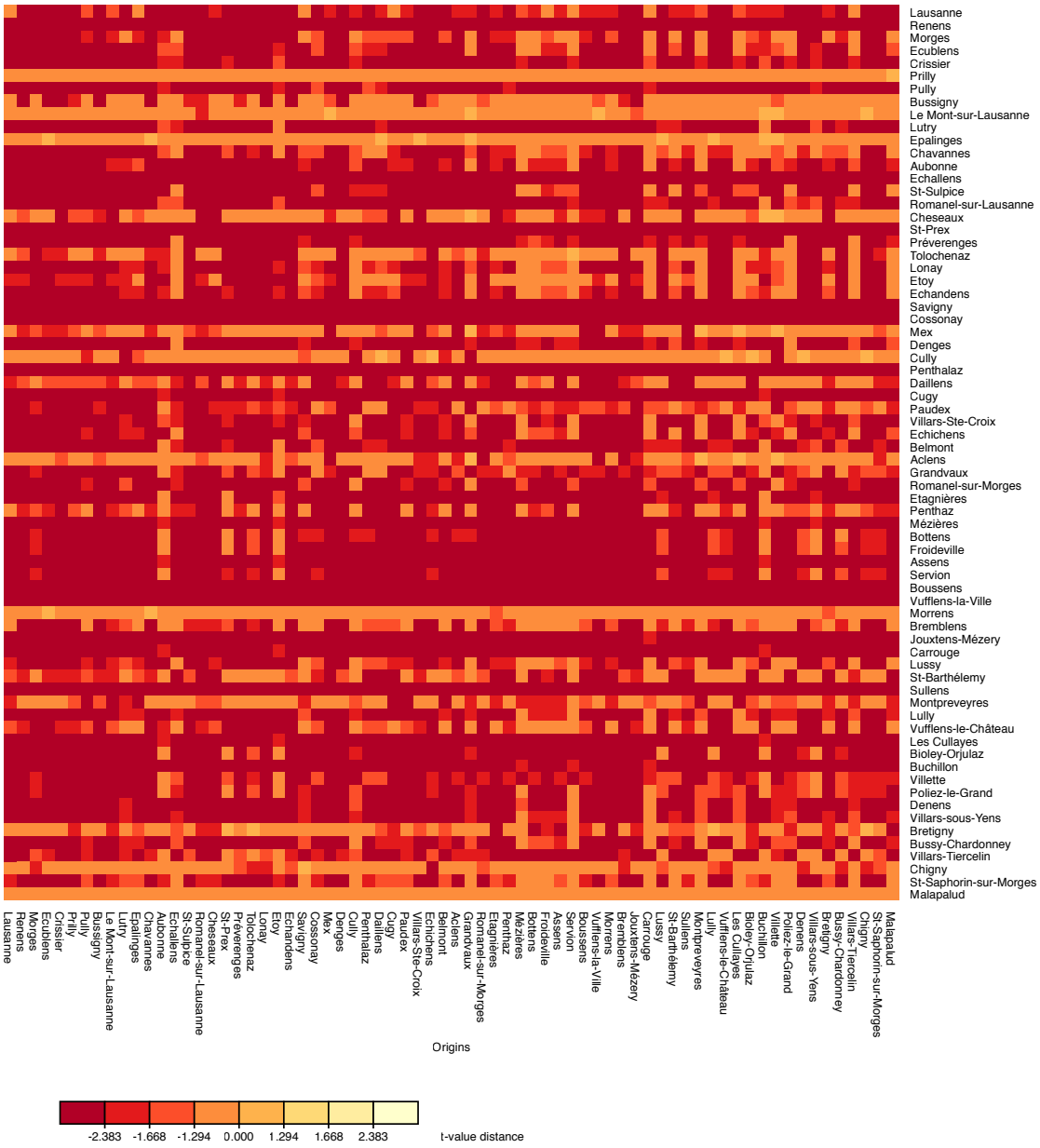


Figure 6.14: The t-values of distance-decay parameter of the destination-specific origin-focused GWSI for Lausanne agglomeration. This map displays absolute t-values, where values smaller than -2.383 are significant at a level of 99%, and values smaller than -1.668 are significant at a level of 95%, and values smaller than -1.294 are significant at a level of 90%. Each row (destination) represents one destination-specific origin-focused model, calibrated separately for each origin (column), resulting in 4900 different t-values.

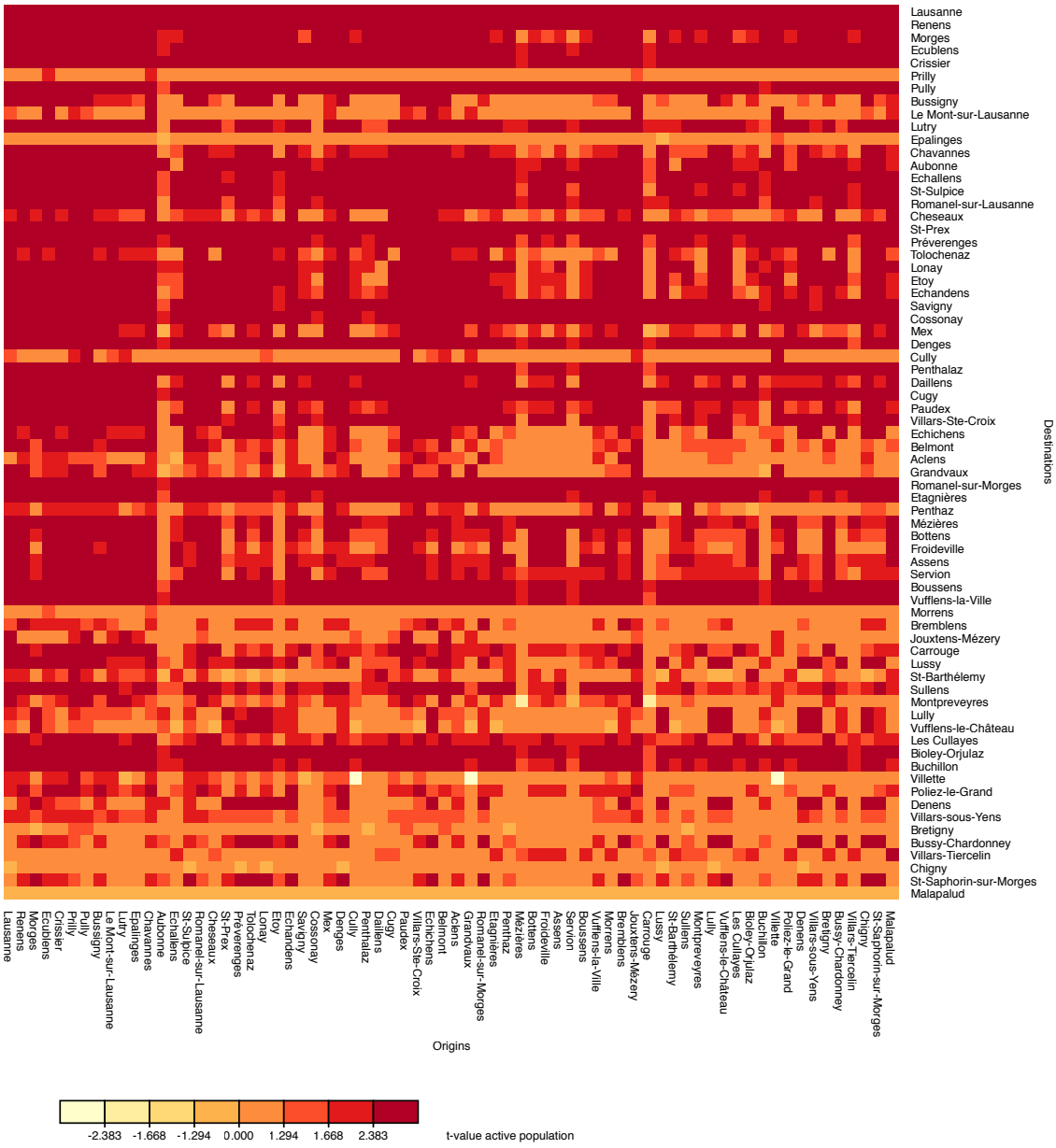


Figure 6.15: The t-values for active population parameter of the destination-specific origin-focused GWSI for Lausanne agglomeration. This map displays absolute t-values, where values greater than 2.383 are significant at a level of 99%, and values greater than 1.668 are significant at a level of 95%, and values greater than 1.294 are significant at a level of 90%. Each row (destination) represents one destination-specific origin-focused model, calibrated separately for each origin (column), resulting in 4900 different t-values.

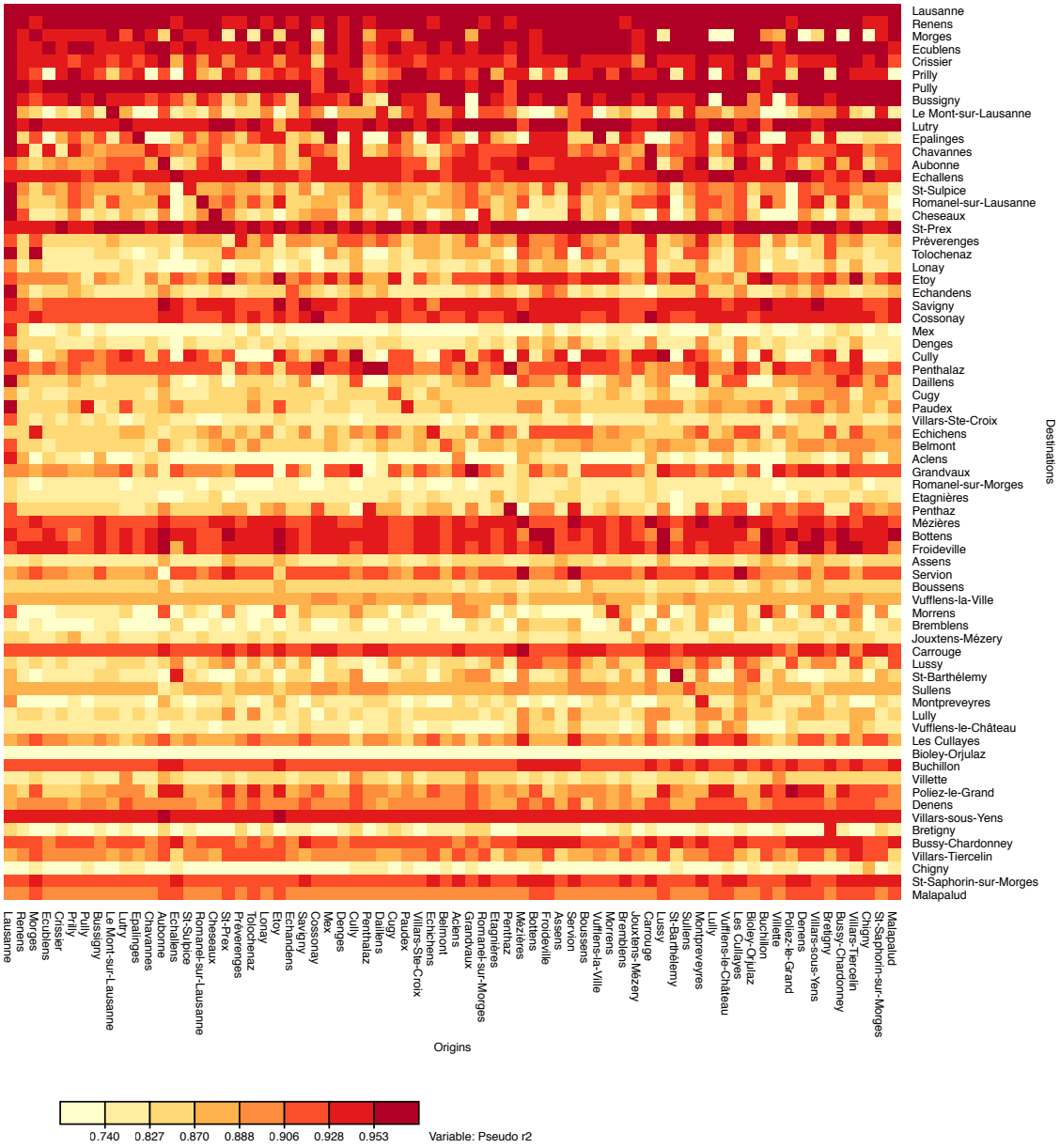


Figure 6.16: The Pseudo R^2 of the destination-specific origin-focused GWSI models for Lausanne agglomeration. Each row (destination) represents one destination-specific origin-focused model, calibrated separately for each origin (column), resulting in 4900 different Pseudo R^2 values.

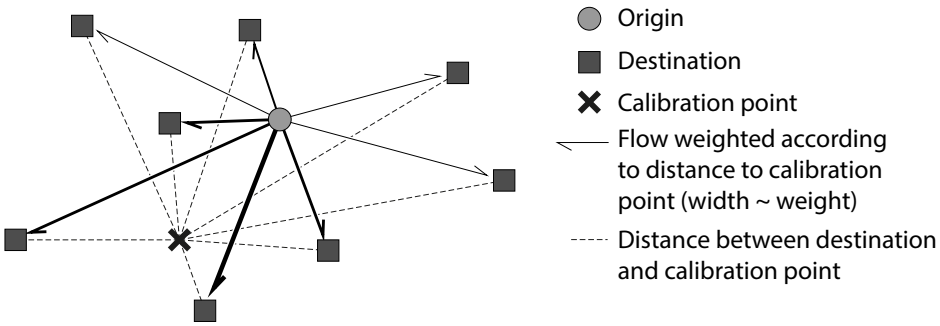


Figure 6.17: An overview illustration of the origin-specific destination-focused GWSI approach.

is defined around any calibration point u and observation flows are weighted based on proximity of their destinations to the centre of the kernel u. For an illustration of the model see figure 6.17 where observed flow weights are represented by their width. In this figure, flows with destinations closer to the calibration point are wider, representing higher weights and a larger effect on the model calibration. The following equation shows a general mathematical formula for the origin-specific destination-focused approach:

$$T_{ij} = \kappa_{(i,\{uj\})}\ w_j^{\gamma_{(i,\{uj\})}}\ d_{ij}^{\beta_{(i,\{uj\})}} \tag{6.17}$$

where T_{ij} represents flows between specific origin i and different destination j, w_j is the number of jobs in destination j, d_{ij} represents the distance between regions and $\kappa_{(i,\{uj\})}$, $\gamma_{(i,\{uj\})}$ and $\beta_{(i,\{uj\})}$ are parameters of the model to be estimated for origin i when a cluster of destinations is involved in the calibration point u. The weighting function and the bandwidth selection can be undertaken in a similar way to the destination-focused approach; for example, a Cauchy kernel with the following formula can be used to assign weights of $W_{u_{ij}}$ to the observed flows:

$$W_{u_{ij}} = \left[1 + (\frac{d_{i(uj)}}{b})^2\ \right]^{-2} \tag{6.18}$$

where $d_{i(uj)}$ is the geographical distance between the calibration point u and destination j of the observed flow ij. As before, the bandwidth parameter b can be estimated with the AICc method.

6.5.1 Application of the origin-specific destination-focused GWSI approach

We apply the proposed origin-specific destination-focused GWSI approach on the journey-to-work dataset of the Lausanne agglomeration with a Poisson gravity-type spatial in-

teraction model as:

$$\lambda_{ij} = \exp(\kappa_{(i,\{j\})} + \gamma_{(i,\{j\})} \ln N_j + \beta_{(i,\{j\})} \ln d_{ij}) \tag{6.19}$$

where as before N_j indicates the number of jobs in destination j and d_{ij} is distance between i and j calculated with a density-based scattering method using the population density (see section 4.3.2 of chapter 4), the parameters of the model $\kappa_{(i,\{j\})}$, $\gamma_{(i,\{j\})}$ and $\beta_{(i,\{j\})}$ are specific to origin i such that a cluster of destinations around the calibration commune j is considered in weighting the flows. The origin-specific destination-focused GWSI is equivalent to the origin-specific model introduced in Nakaya (2001, 2003). As explained before, an origin-destination specific model was introduced in (Nakaya, 2001) to be calibrated using GWR. The main idea of the method was to estimate a set of local parameter estimates for each specific flow (origin-destination specific) in the system. However, to avoid the challenge of calculating weights between flows in his paper, Nakaya (2001) uses a simplified origin-specific version of the proposed flow-specific model where the model restriction helps to apply geographical weighting only for destination locations (see Nakaya, 2001, 2003). It should be noted here that when the calibration points and destinations share the same geographical locations (i.e. when the model is calibrated for each destination in the study region), Nakaya's origin-specific model for flows is equivalent to the origin-specific destination-focused GWSI. However, the latter considers a broader view of the geographically weighted spatial interaction in the sense that this model can be calibrated in any location within the region even where no flow data are available. In fact, the fundamental theory behind the origin-specific destination-focused GWSI models is to enable us to provide more disaggregated local information at any arbitrary location of the study region rather than only at each origin. As noted before in section 6.1, two possible solutions for calculating distances between flows will be presented in the next chapter, where the calibration points are spatial flows rather than geographical locations.

In order to evaluate the optimum bandwidth for different origins when applying the destination-focused GWSI on the Lausanne dataset, the same AICc approach is used. Figure 6.18 shows a map of optimum bandwidth results for origin communes. The values illustrate a similar pattern to the bandwidth values for destinations in the destination-specific origin-focused approach (see figure 6.11) in which bandwidths are smaller for central communes and bigger for the smaller origins in the bordering regions of the agglomeration. However, here in the origin-specific destination-focused approach, bandwidths are bigger in general, ranging between 1800 metres to more than 31 kilometres. The value of 31 kilometres and above is almost the global bandwidth value which covers nearly the whole study area. This may shows that for these communes, the optimal model is close to a pure origin-specific model and perhaps the combination with the origin-focused approach does not lead to an improvement. However, this is only the case of a few number of small bordering communes.

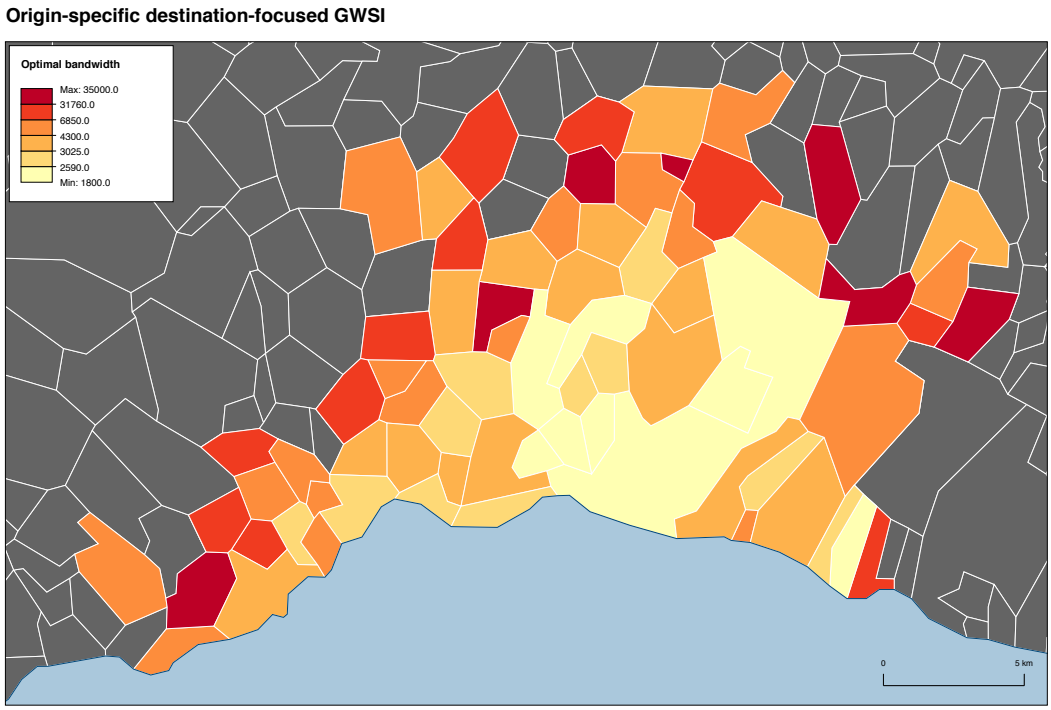


Figure 6.18: Optimal bandwidth for the origin-specific destination-focused GWSI models.

We have used the calculated optimal bandwidth values shown in figure 6.18 to further illustrate the application of the origin-specific destination-focused GWSI on the journey-to-work dataset in the Lausanne agglomeration. The calibration procedure is repeated for all 70 destinations to provide local parameter estimates for each specific origin within the agglomeration. Again, for each specific origin, there are 70 local values for each parameter corresponding to different destinations. Similarly to section 6.4.1, we provide 70×70 matrices to illustrate the local parameter estimates of the model. As before, origins and destinations are placed in decreasing order of the total number of outgoing flows, this time however the rows of the matrices represents origins and the columns show the parameter estimates at the corresponding destinations. For instance, figure 6.19 represents the distance-decay parameter of the models, the values in each column show how distance is perceived by people living in a specific origin (located in one of the rows of the matrix) when commuting to each of the destinations. In general, the estimated distance-decay values display the expected negative effect on interactions. However, the negative effect of distance is less for people living in larger communes as their associated distance-decay values are lower compared to the smaller communes shown at the bottom part of the matrix. This can be explained again with the better transportation systems in larger cities. The t-values of the distance-decay parameters illustrated in figure 6.20 do not show a clear spatial pattern for larger or smaller communes. In general, although some of the t-values shown in the matrix are close to zero, the p-values of the distance-decay parameter in more than 90% of the models are significant with 98% confidence

(90th percentile of p-values for the distance-decay parameter = 0.02).

The matrix for the number of jobs parameter estimates is shown in figure 6.21. This matrix presents more variation of the parameter estimates for different origins compared to the distance-decay estimates. The values of the number of jobs parameter for each model shows that how would increasing the number of jobs in different destinations affect the commuting behaviour of the residence of each specific origin. In other words, how total interaction from each origin would be affected by changing the number of jobs in different destinations. For instance, consider Lausanne city located in the first row, the results show how people living in Lausanne will behave in terms of commuting to the different destinations if the active population increases. Generally, the number of interactions to all destinations will increase, but the amount of increase will vary for the separate destinations. Furthermore, the values of the number of jobs parameter in figure 6.21 seems to be slightly higher in the lower part of the matrix where smaller communes are listed. This indicates that increasing number of jobs in destinations affects more smaller communes in the agglomeration than bigger origins. This might be because bigger origins have more number of jobs in their own communes with a higher internal flows compared to outgoing flows so that increasing the number of jobs in other communes do not change much the commuting pattern in these communes.

The t-values of the number of jobs parameters for all models are illustrated in figure 6.22. For most of the models the t-values are significantly higher than zero with few exceptions such as the small commune of Villette. It is also interesting that for most communes there are nearly no variations in the t-values across different destinations indicating the significance of all the number of jobs parameters over different destinations. The p-values of the number of jobs parameters shows that this parameter in more than 95% of the local models is significant with 99% of confidence (95th percentile of p-values for the number of jobs parameter = 0.008). Finally, the values of Pseudo R^2 of all the origin-specific destination-focused models are calculated and presented in figure 6.23. The R^2 values range roughly between 0.8 to over 0.97 showing a satisfactory goodness-of-fit for the models, with a better fit in larger communes. This could be due to more flow data available in larger communes in comparison with smaller residential places.

6.6 Summary

The key idea of this chapter was localising spatial interaction models based on a geographically weighted approach. The general concept of geographical weighting, known as GWR, has been applied to spatial interaction models with the calibration points being geographical locations within the study region. The same case study of journey-to-work flows in the agglomeration of Lausanne used in other chapters was used to illustrate the proposed local geographically weighted spatial interaction (GWSI). Considering a spatial kernel around each calibration point, all flows within this kernel have been weighted

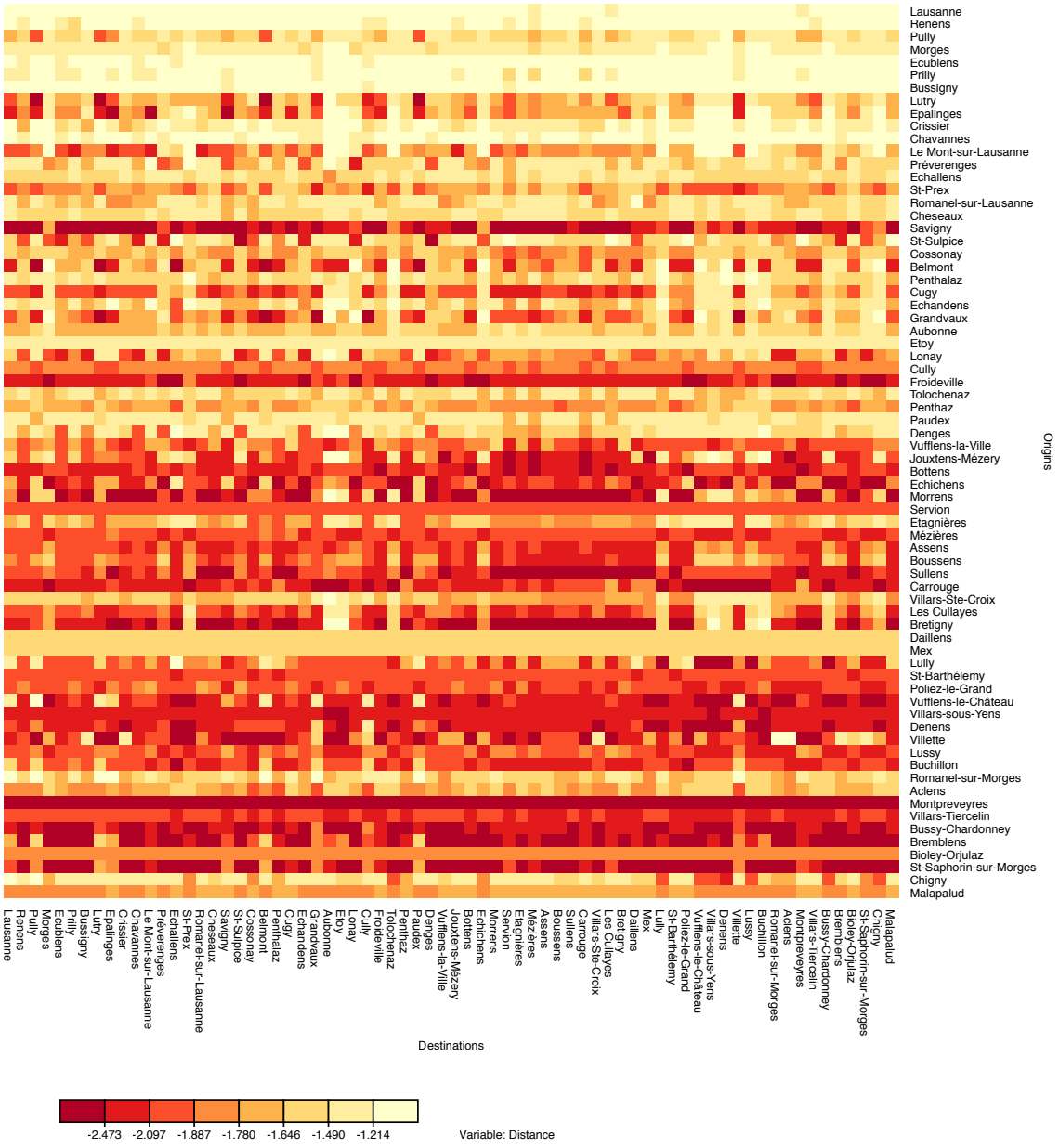


Figure 6.19: Distance-decay parameter of the origin-specific destination-focused GWSI for Lausanne agglomeration. Each row (origin) represents one origin-specific destination-focused model, calibrated separately for each destination (column), resulting in 4900 different parameter estimates.

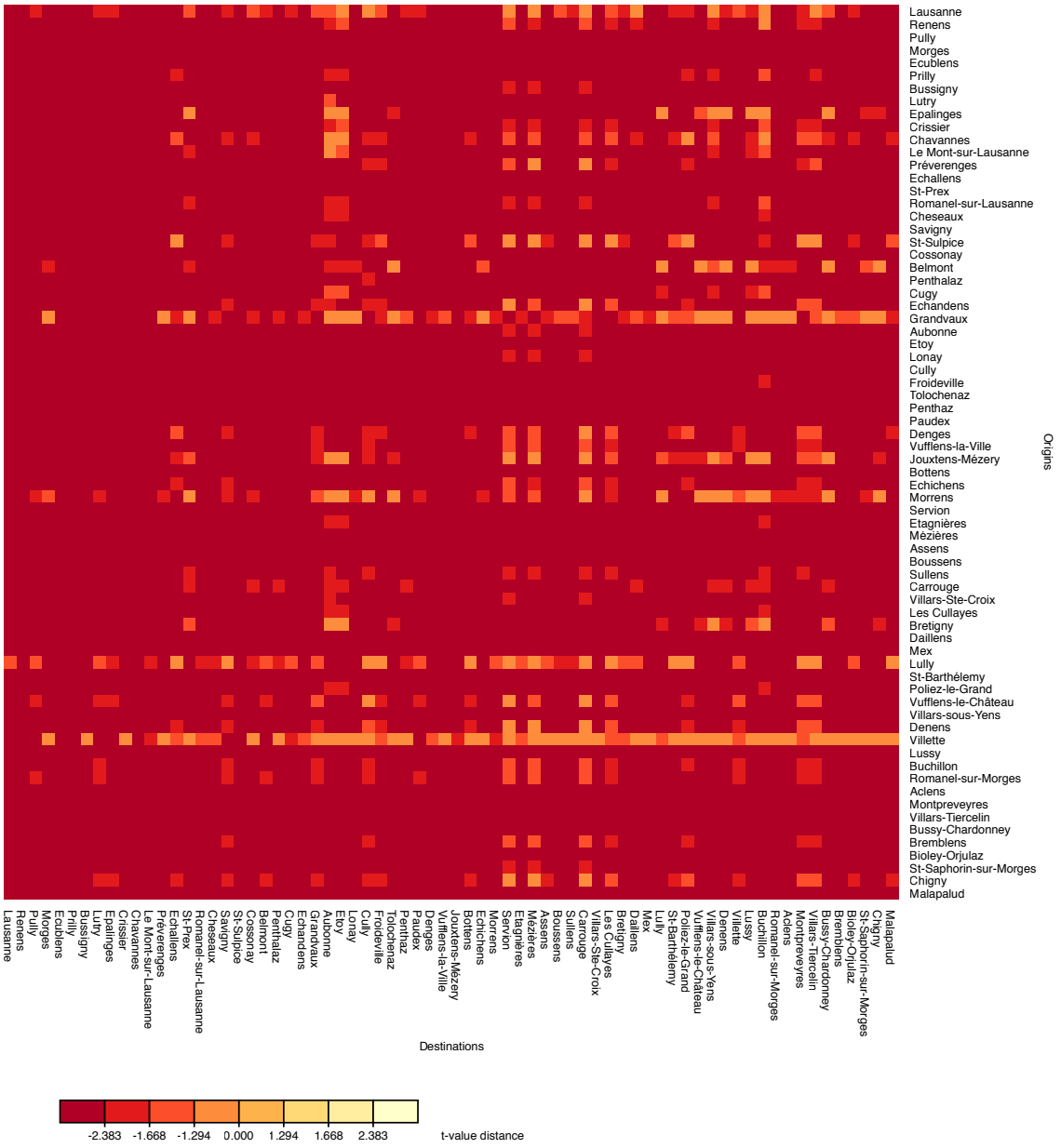


Figure 6.20: The t-values for the distance-decay parameter of the origin-specific destination-focused GWSI for Lausanne agglomeration. This map displays absolute t-values, where values smaller than -2.383 are significant at a level of 99%, and values smaller than -1.668 are significant at a level of 95%, and values smaller than -1.294 are significant at a level of 90%. Each row (origin) represents one origin-specific destination-focused model, calibrated separately for each destination (column), resulting in 4900 different t-values.

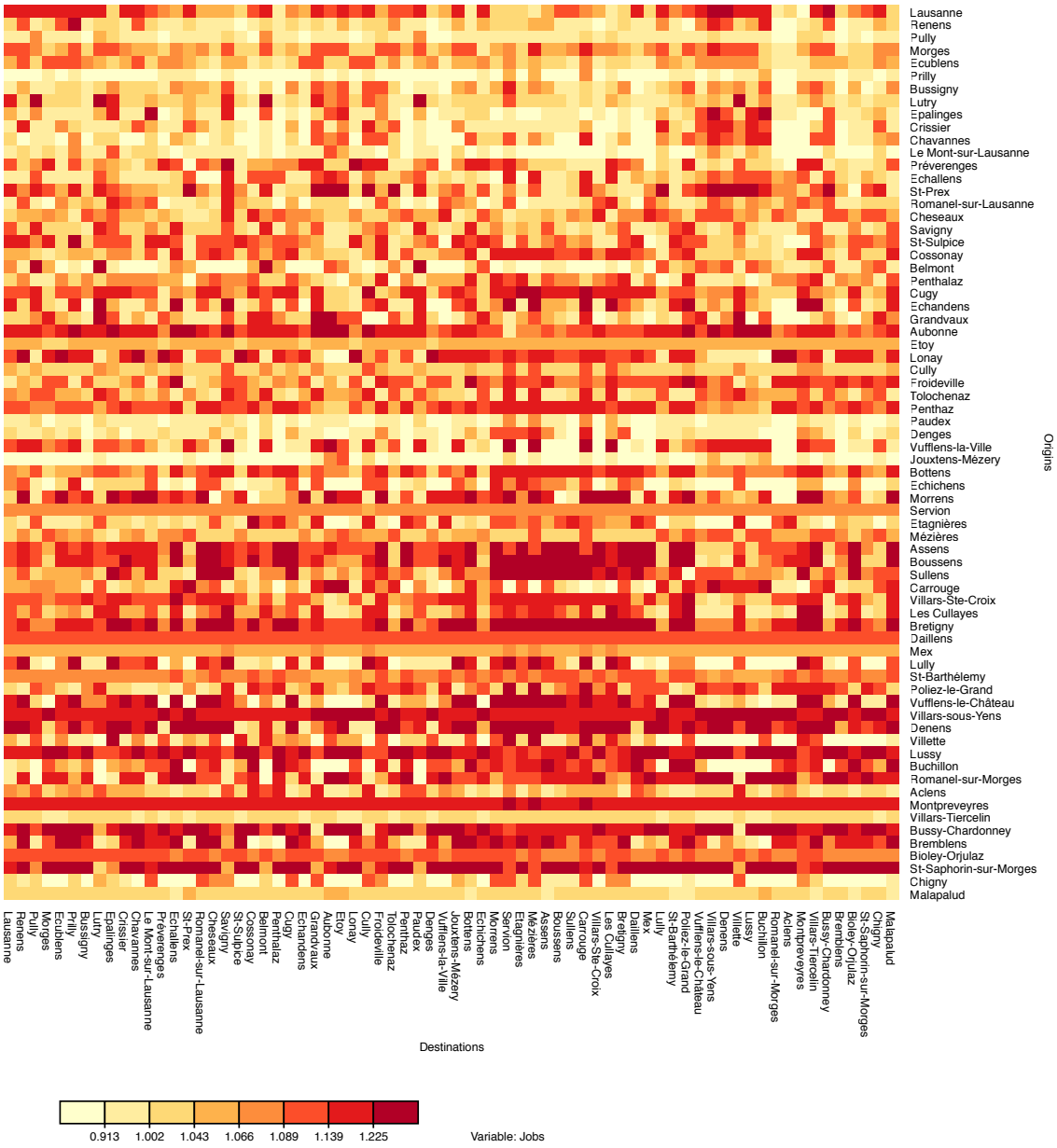


Figure 6.21: Number of jobs parameter of the origin-specific destination-focused GWSI for Lausanne agglomeration. Each row (origin) represents one origin-specific destination-focused model, calibrated separately for each destination (column), resulting in 4900 different parameter estimates.

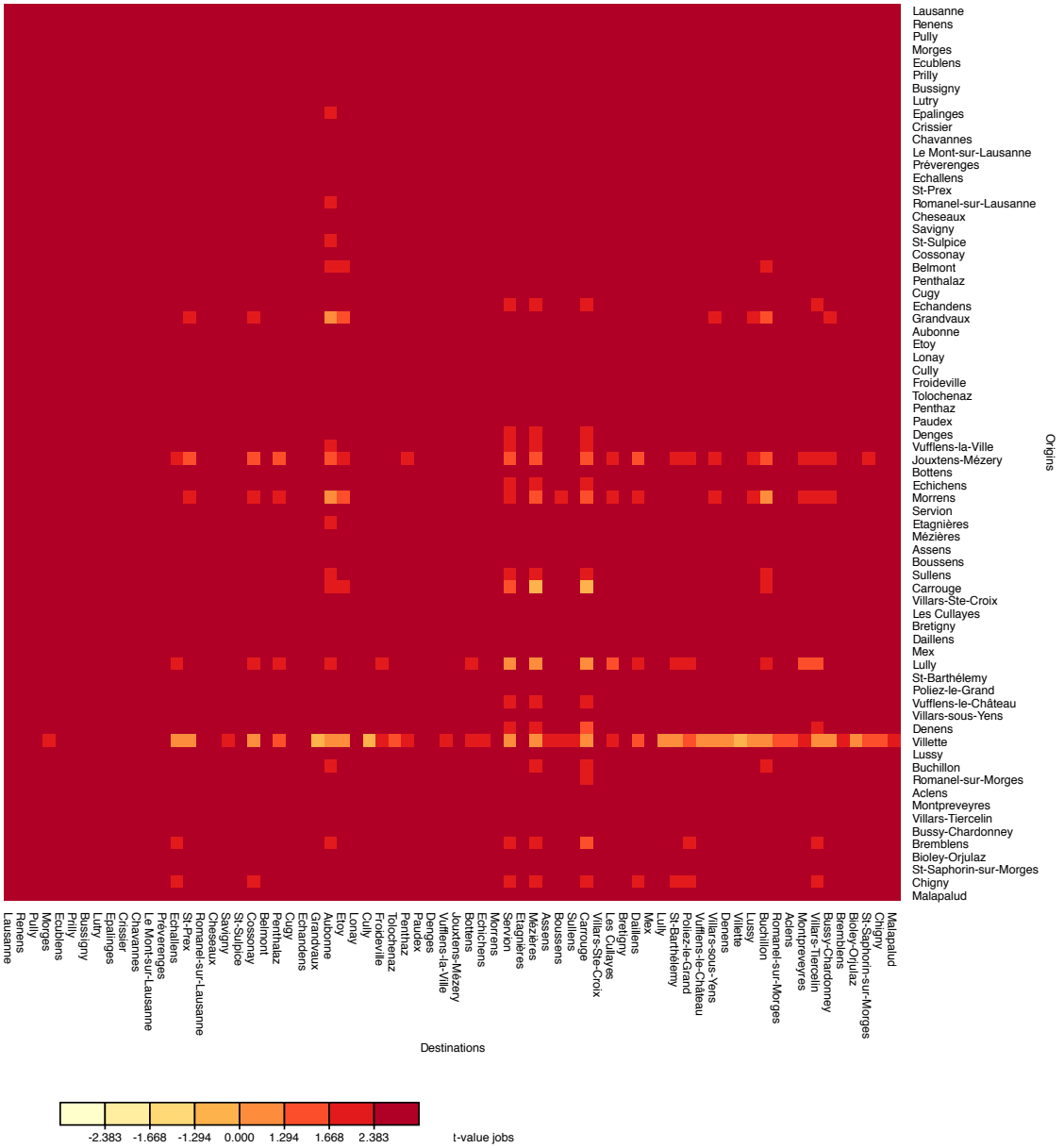


Figure 6.22: The t-values for number of jobs parameter of the origin-specific destination-focused GWSI for Lausanne agglomeration. This map displays absolute t-values, where values greater than 2.383 are significant at a level of 99%, and values greater than 1.668 are significant at a level of 95%, and values greater than 1.294 are significant at a level of 90%. Each row (origin) represents one origin-specific destination-focused model, calibrated separately for each destination (column), resulting in 4900 different t-values.

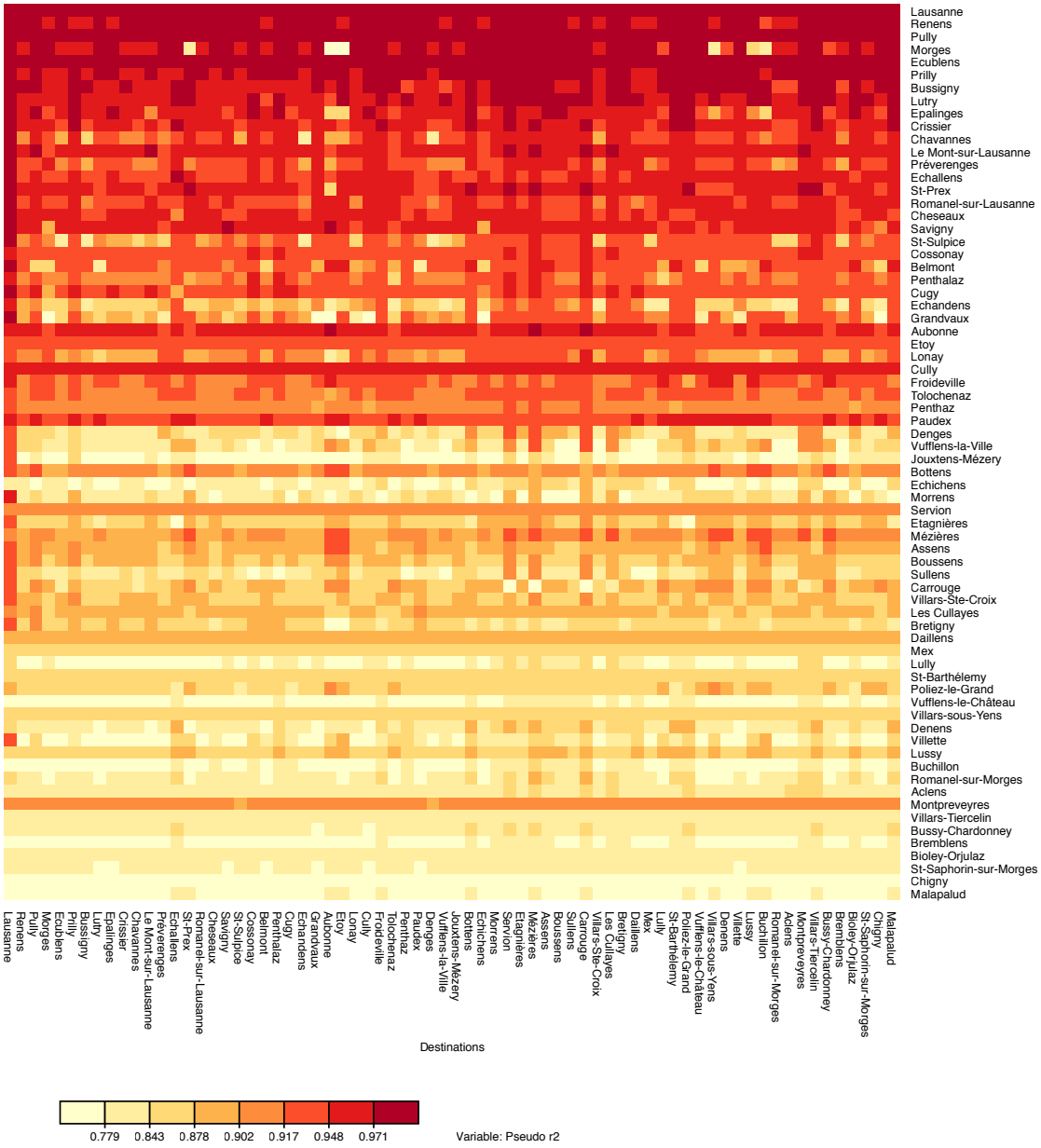


Figure 6.23: The Pseudo R^2 of the origin-specific destination-focused GWSI models for Lausanne agglomeration. Each row (origin) represents one origin-specific destination-focused model, calibrated separately for each destination (column), resulting in 4900 different Pseudo R^2 values.

according to their distance to the calibration point. The distance between observed flows and the calibration point was obtained using two approaches; one based on the origins of the flows (origin-focused approach) and another based on destinations (destination-focused approach). These models have been taken a step further to provide more spatial disaggregated information by calibrating a destination- and origin-specific version of the origin- and destination-focused GWSI models respectively. For each model, the optimal bandwidths and local parameters were obtained and analysed. All the presented GWSI models in this chapter allow for localising spatial interaction models with geographical weighting of the flows and the calibration points being geographical location within the region. The calibration locations can be any arbitrary location where no interaction information is available. Considering a cluster of origins or destinations around the calibration point, GWSI models use information of neighbouring regions for model calibration.

In the following chapter we present another variant of a GWSI model where observations and calibration points are spatial flows. In this scenario, each flow is represented with both its origin and destination and two different approaches based on a four dimensional kernel and spatial trajectories are considered for calculating the distances between flows.

Chapter 7

GWSI: Flow-focused approach

7.1 Introduction

In order to localise spatial interaction models using the geographically weighted (GW) concept we have taken three approaches differing mainly in the way calibration points (or calibration flows) are defined and spatial separation (distance) between flows estimated. In the first two approaches the calibration points are actual geographical locations within the study region, e.g. any existing origin or destination, or any arbitrary location with no flow data available. These models, (i.e. the origin- and destination-focused GWSI models), have been presented in the previous chapter along with some empirical examples.

In a third approach, we consider the calibration points to be spatial flows (i.e. pairs of origin-destinations) within the study region. Therefore, the interaction models can be calibrated locally for each flow and by moving the calibration flows across the region, a surface of local parameter estimates can be generated across the region. We name this approach flow-focused GWSI and will present this model in more detail below.

7.2 Flow-focused GWSI approach

In the origin- and destination-focused models, a local set of parameter estimates is calibrated for each existing origin and destination within the study area. Analysing these parameter estimates provides local information about the interaction behaviour of the calibration locations considering their neighbouring regions. These models could also be used for forecasting purposes when they are calibrated for locations with no interactions or with no flow information. It is also interesting to localise spatial interaction models over spatial flows rather than over a single location. These spatial flows can be any arbitrary interaction flows between the existing origins and destinations in the system, or the model can be calibrated for a pair of origins and destinations with no flow between them. In this case the localised interaction model can be used for prediction purposes.

The main methodology in this version of GWSI models (i.e. flow-focussed GWSI) is similar to the origin- and destination-focused GWSI in which the GW concept is used for

calibrating the spatial interaction models. However here in the flow-focused approach, the calibration points are spatial flows between two geographical locations i and j. In this method, a spatial kernel will be placed around each calibration flow and observed flows within this kernel will be weighted based on their proximity to the calibration flow ij. The higher weights are given to the flows closer to the flow ij and the maximum weight of unity is at the centre of the kernel. The following equation shows the general formulation of the flow-focused GWSI model:

$$T_{ij} = \kappa_{ij}\ v_i^{\alpha_{ij}}\ w_j^{\gamma_{ij}}\ d_{ij}^{\beta_{ij}} \tag{7.1}$$

where T_{ij} is the flow between i and j, v_i is the origin propulsiveness variable, w_j represents the destination attractiveness, d_{ij} is the distance between i and j; κ, α, γ and β are model parameters to be calibrated localised to the flow ij .

The calibration of a Gaussian flow-focused model, similar to an origin- or a destination-focused model can be done using the WLS method (see chapter 6 for more information) with the following formulation:

$$b'_{ij} = (X^T\ W_{ij}\ X)^{-1}\ X^T\ W_{ij}\ T_{ij} \tag{7.2}$$

where b'_{ij} is a vector containing the model parameter estimates local to the flow ij, X is a matrix of the independent variables including a column of ones for the intercept, X^T is the transpose matrix of X, T_{ij} is the vector of the dependent variable showing flows from origin i to destination j and W_{ij} is a weighting matrix with elements calculated with a weighting function based on the distance between the calibration flow ij and neighbouring observation flows. In a similar way, when the flow-focused GWSI is formulated as a Poisson spatial interaction model as below:

$$\lambda_{ij} = \exp(\kappa_{ij} + \alpha_{ij} \ln v_i + \gamma_{ij} \ln w_j + \beta_{ij} \ln d_{ij}) \tag{7.3}$$

the calibration of the model can be done using the weighted log-likelihood method (see section 6.2) with the following formulation:

$$\ln L(\lambda_{ij}) = \sum_{ij} (-\lambda_{ij} + T_{ij} \ln \lambda_{ij} - \ln T_{ij}!)\ W_{ij} \tag{7.4}$$

As explained in chapter 6, different weighting functions can be used for calculating the elements of the weighting matrices. In a Gaussian kernel for instance, the weighting function for a flow-focused GWSI model can be formulated as:

$$W_{ij} = \exp\ [\ -\frac{1}{2}\ (\frac{d}{b})^2\ ] \tag{7.5}$$

where the spatial weights depend on two parameters: (i) d the spatial separation (e.g. distance) between flows, and (ii) b, the bandwidth parameter that should be estimated

by an appropriate method such as minimising the AICc (see section 6.2 for more information). In GWR, the spatial separations between the locations of the observed data and the calibration point are often estimated by calculating the geographical (Euclidean) distance between them. However, in the flow-focused GWSI model the observed data and the calibration points are all spatial interactions (flows). As stated in Nakaya (2001, 2003), estimating the spatial separation (distance) between two flows containing two geographical locations each, can be highly complicated. In the following sections we propose two different ways of calculating the distance between spatial flows based on (*i*) a four-dimensional distance calculation and (*ii*) a spatial trajectory distance measure.

7.2.1 A four-dimensional kernel approach

The term *"spatial interaction"* conceptually can be defined as the *"flow"* of goods, people, information or units of any kind in motion (Bavaud, 2010b) between geographical places (Fotheringham and O'Kelly, 1989; Haynes and Fotheringham, 1984; Fischer, 2000; Fotheringham, 2001). Each spatial flow is associated with two geographical locations; the origin and destination points. Traditionally, spatial flows are illustrated as *"arrows"* between these geographical locations, e.g. flow between origin i and destination j is usually shown as: $i \rightarrow j$. Considering the fact that each spatial flow contains two points in space, we can illustrate each flow as a pair of origin and destinations, e.g. flow ij can be shown as $(\overrightarrow{i,j})$. In this form, each flow is represented as a directed spatial vector with magnitude equal to the number of interactions between i and j and then the spatial separation between different flows can be represented as the *"dissimilarity"* between vectors.

Dissimilarity measures or coefficients are defined in mathematics by the following properties (see for instance Webb, 2002; Greenacre, 2008): If d_{ab} is the dissimilarity of an object (e.g. vector) a from object b, then:
1) $d_{ab} \geq 0 \ (= 0 \ if \ a = b)$
2) $d_{ab} = d_{ba}$
3) $d_{ab} \leq d_{ac} + d_{cb}$
The first item is trivial; the distance between flows is always a positive value and it is zero for the distance between a vector and itself. The second condition represents the symmetrical attribute of the measure and in most cases is satisfied (e.g. when the dissimilarity is measured with Euclidean distance). However, as stated in Webb (2002), the symmetry condition might not be always respected; for instance if the road distance is considered between two places when in one direction the communication is longer than the other way because of one-way streets. The third and last property is known as *triangle inequality*. If this condition is satisfied, the dissimilarity measure is said to be a metric dissimilarity or a distance (Webb, 2002).

In a Euclidean space R^n, the most widely known metric for dissimilarity between two points (vectors) is the Euclidean distance. Consider point A with coordinate $(x_1, x_2, \ldots, x_n)$

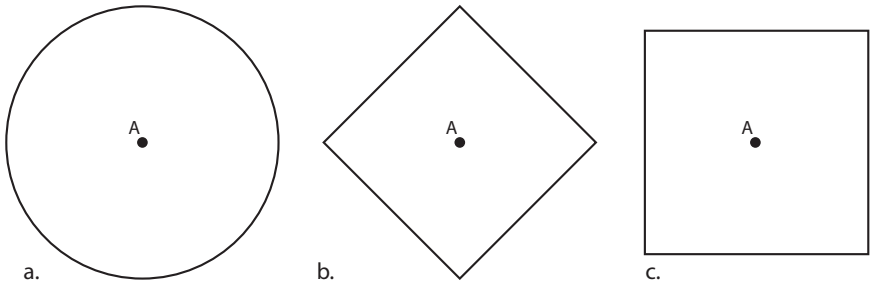


Figure 7.1: The set of points at equal distance r from a given point A where r is a: (a.) Euclidean distance, (b.) city-block distance and (c.) Chebyshev distance.

and point B with coordinate $(y_1, y_2, \ldots, y_n)$, the Euclidean distance between A and B is calculated by:

$$d_{Euclidean}(A, B) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{7.6}$$

The set of points at equal distance r from a given point A, where r is a Euclidean distance, form the circumference of the circle in the two-dimensional space (see figure 7.1 a.), and the surface of a sphere in three-dimensional space.

While the Euclidean distance is the best known and most widely used distance measure, other distance measures exist for space points (vectors). For instance, the city-block distance (also known as Manhattan or taxicab distance) between two points A and B is calculated by the sum of the absolute differences of their coordinates:

$$d_{Manhattan}(A, B) = \sum_{i=1}^{n}|x_i - y_i| \tag{7.7}$$

The connections between points in city-block distance are grid lines. As the name of the metric suggests, the grid can be considered as a net of streets (square blocks) between the points in a city (e.g. Manhattan). In this metric, the set of points at equal distance r from a given point A, where r is a city-block distance, is the outline of a square with sides oriented at a 45° angle to the coordinate axes (see figure 7.1 b.). Both Euclidean distance and Manhattan distance can be seen as special cases of the Minkowski distance ($m \geq 1$):

$$d_{Minkowski}(A, B) = \big(\sum_{i=1}^{n}|x_i - y_i|^m\big)^{1/m} \tag{7.8}$$

In the case of city-block distance, $m = 1$, and for the Euclidean distance m is 2. While the value of m takes typically values of 1 or 2, it can in theory take values bigger than 2. If m tends to infinity, the distance tends to a metric called Chebyshev or maximum

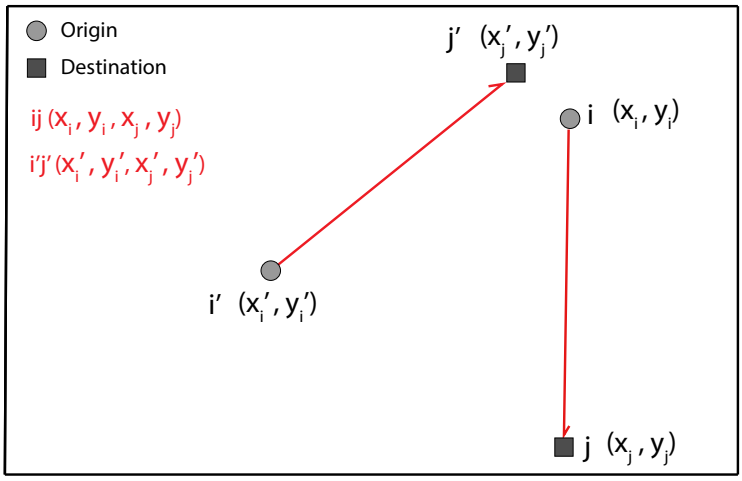


Figure 7.2: The flow (ij) and $(i'j')$ represented as a four-dimensional vector in Euclidean space.

value distance:

$$d_{Chebishev}(A, B) = max_i \ |x_i - y_i|. \tag{7.9}$$

The set of points at equal distance r from a given point A, where r is a Chebyshev distance, form the outline of a square where the sides are parallel to the coordinate axes (see figure 7.1 c.). Minkowski distances with a value of m bigger than 2 are abstract distances with no direct representation in Euclidean space. However, from a computational perspective, the Chebyshev distance is the fastest distance to be computed and is often used when the execution time is a critical factor (Webb, 2002).

Defining spatial flows as pairs of origins and destinations, a flow between origin i with coordinates (x_i, y_i) to destination j with coordinates (x_j, y_j) can be considered as a four-dimensional vector (x_i, y_i, x_j, y_j) (see figure 7.2.1). Given that all the distance measures presented above can be applied on any pair of vectors of n dimensions, we can use them for calculating the distance between two flows as a four-dimensional distance measure. In the case of the four-dimensional Euclidean distance, we can compute the distance between flow ij with origin-destination coordinates (x_i, y_i, x_j, y_j) and the flow $i'j'$ with coordinates $(x_{i'}, y_{i'}, x_{j'}, y_{j'})$ with the following formulation:

$$d_{(ij)(i'j')} = \sqrt{(x_i - x_{i'})^2 + (y_i - y_{i'})^2 + (x_j - x_{j'})^2 + (y_j - y_{j'})^2} \tag{7.10}$$

In the flow-focused GWSI model, each flow ij is weighted using a function like the one specified in equation 7.5, with respect to a calibration flow $i'j'$. The distance d in formula 7.5 becomes in this case the four-dimensional distance between the calibration flow and flow ij, which can be calculated using equation 7.10 or potentially any other

distance presented above. We will refer to this kind of weighting function with a four-dimensional distance as a *four-dimensional kernel* function. As the visualisation of 4-dimensional objects is not trivial, here no figure is presented of such a four-dimensional kernel. When using a four-dimensional kernel, we will write the flow-focused GWSI model as:

$$T_{ij} = \kappa_{\{ij\}}\ v_i^{\alpha_{\{ij\}}}\ w_j^{\gamma_{\{ij\}}}\ d_{ij}^{\beta_{\{ij\}}} \tag{7.11}$$

where notation $\{ij\}$ indicates that a cluster of flows is considered around each calibration flow ij and the separations between flows are calculated using the four-dimensional distance.

7.2.2 Spatial trajectories approach

A trajectory can be defined as a directed trace or a path generated by a moving object in geographical space that defines a link between two locations (Wood et al., 2009; Zheng and Zhou, 2011). Sometimes a trajectory is a simple path, e.g. the shortest Euclidean distance between a start and an end point but it can also have more complex traversal objectives, e.g. to be a function of time. In this case, a trajectory is a time-stamped series of spatial coordinates: $p_1 \rightarrow p_2 \rightarrow \cdots \rightarrow p_n$, with every $p_i = (x_i, y_i, t_i)$.

Measuring the similarity or dissimilarity of two trajectories is a common problem and several approaches have been suggested for this purpose. If two trajectories have the same number of points corresponding to each other, the sum-of-pairs distance (SOP) suggested by Agrawal et al. (1993), can be computed by simply summing up the distances between the corresponding points:

$$SOP(p_1 \cdots p_n, q_1 \cdots q_n) = \sum_{i=1}^{n} d(p_i, q_i) \tag{7.12}$$

where $d(p_i, q_i)$ is the Euclidean distance between points p_i and q_i (see Zheng and Zhou, 2011). Figure 7.3 illustrates this principle. Obviously, the Euclidean distance $d(p_i, q_i)$ could be replaced by any distance measure defined between points. For trajectories of different lengths, more sophisticated dissimilarity measures are known (see e.g. Zheng and Zhou, 2011). Some of these measures do not meet the triangle equality condition and therefore not all can be considered as distances. These alternative measures are not discussed here, as they are not relevant for spatial interaction models.

The sum-of-pairs distance for trajectories can also be applied to flows in spatial interaction. In this case, the distance between flow ij and $i'j'$ becomes:

$$SOP(ij, i'j') = d(i, i') + d(j, j') \tag{7.13}$$

where $d(\cdot)$ can be again the Euclidean distance. Using the presented trajectory distance

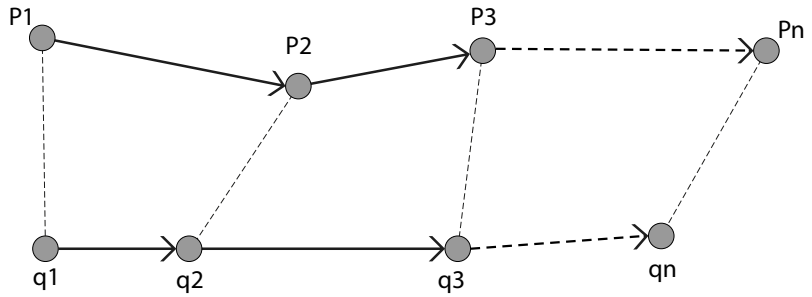


Figure 7.3: $Distance(p_1 \cdots p_n, q_1 \cdots q_n) = \sum_{t=1}^{n} \|p_t - q_t\|$

approach, the formula for the flow-focused GWSI model can be written as:

$$T_{ij} = \kappa_{\{i\}\{j\}}\ v_i^{\alpha_{\{i\}\{j\}}}\ w_j^{\gamma_{\{i\}\{j\}}}\ d_{ij}^{\beta_{\{i\}\{j\}}} \tag{7.14}$$

where the notation $\{i\}\,\{j\}$ indicates that a cluster of flows is considered around the calibration flow ij and the distance between flows is calculated based on distances between origins and destinations using the sum-of-pairs trajectory distance approach. It should be noted here that the main methodology for localisation of the flow-focused GWSI is the same in both the four-dimensional kernel (equation 7.11) and the trajectory approach (equation 7.14) but the calculation of distances between flows in the weighting function is different.

7.3 Application of the local flow-focused model to Lausanne journey-to-work data

In this section we apply the flow-focused GWSI model to the Lausanne journey-to-work dataset. A Poisson flow-focused GWSI model can be formulated as:

$$\lambda_{ij} = \exp(\kappa_{\{ij\}} + \alpha_{\{ij\}} \ln P_i + \gamma_{\{ij\}} \ln N_j + \beta_{\{ij\}} \ln d_{ij}) \tag{7.15}$$

where the model variables and parameters are defined as in the previous chapters (e.g. see chapter 6).

7.3.1 Bandwidth selection

The four-dimensional kernel and the spatial trajectories approaches presented in sections 7.2.1 and 7.2.2 provide two possible new alternatives for estimating the separations (distances) between spatial flows. In the calibration process of the above flow-focused model, we can use any of these spatial kernel approaches for weighting the spatial flows. However, the main principle of geographically weighting of the flows is the same in both

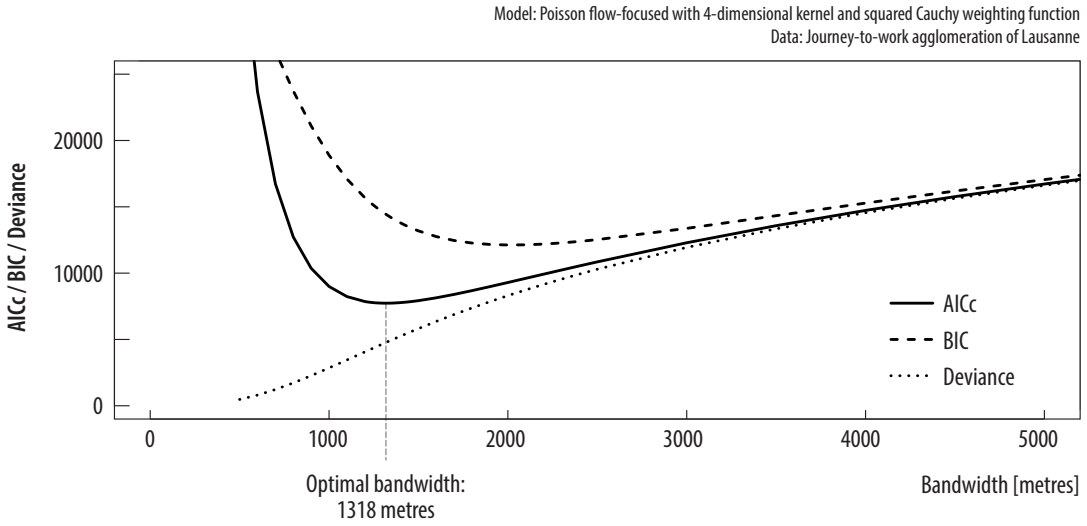


Figure 7.4: Bandwidth value (fixed) against AICc score for the Poisson flow-focused model in Lausanne agglomeration. Deviance and BIC are also shown.

spatial kernels, even though these approaches might result in a slightly different weight for each flow. Therefore, the result of the calibrated flow-focused model using these spatial kernel approaches would be similar and so as an example, we calibrate this model for each flow within the agglomeration of Lausanne using the four-dimensional kernel approach only.

The optimal bandwidth for the flow-focused model can be found using the same techniques as for other GW models. In the present case, the optimal bandwidth has been estimated using the AICc approach. The calculated bandwidth values against AICc scores for the flow-focused GWSI model when the squared Cauchy kernel has been used as weighting function are shown in figure 7.4. Using AICc, the optimal bandwidth is found to be 1318 metres. The plot also contains BIC scores and deviance.

7.3.2 Adaptive spatial kernels

Instead of using a fixed bandwidth in the flow-focused GWSI models, it would also be possible to apply an adaptive kernel. The adaptive kernel is sensitive to the density of data and would vary spatially so that the kernel is smaller in regions where many data flows are available while the kernel has larger bandwidth where the flows are sparse (see Fotheringham et al., 2002). As mentioned in section 5.3.1, there are different methods for producing spatially varying kernels. One of the kernels that is related to the Nth nearest neighbours is the bi-square function with following formula:

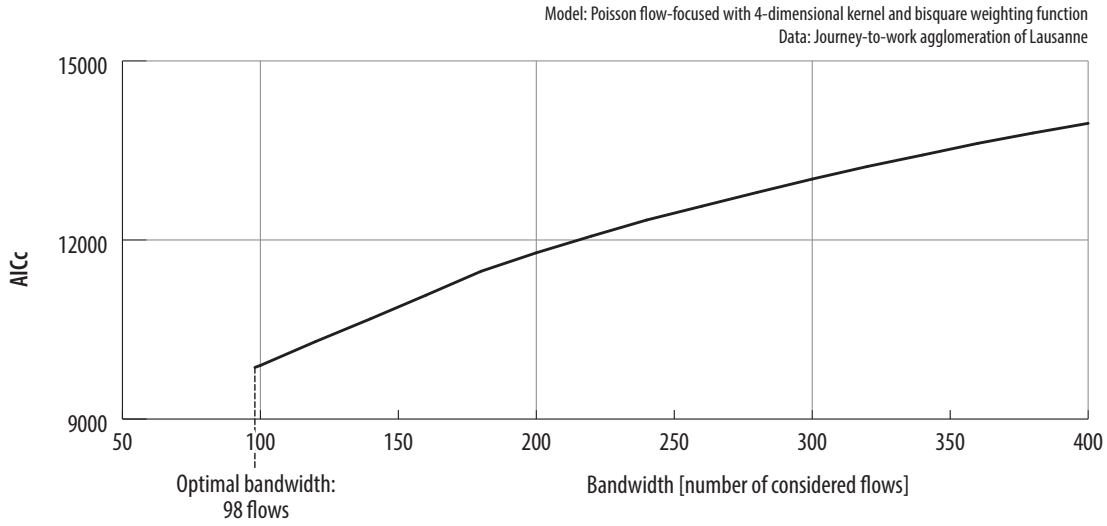


Figure 7.5: Adaptive bandwidth value (number of flows considered) against AICc score for the Poisson flow-focused model in Lausanne agglomeration (using a bi-square kernel).

$$W_{ij} = \begin{cases} [1 - (\frac{d_{(ij)(i'j')}}{b})^2\ ]^2 & \text{if } ij \text{ is one of the Nth nearest} \\ & \text{neighbours of the calibration} \\ & \text{flow and } b \text{ is the distance} \\ & \text{to the Nth nearest neighbour.} \\ 0 & \text{otherwise} \end{cases} \tag{7.16}$$

In this method, the adaptive bandwidth is therefore measured in terms of the number of flows to be considered in the model calibration instead of a distance as in the fixed kernel. The procedure to find the optimal adaptive bandwidth is the same as with flexible bandwidth, i.e. the AICc score is computed for a set of different bandwidth values, for example by using a golden section search approach. Figure 7.5 shows the bandwidth plot for different number of neighbouring flows considered, for the Poisson flow-focused GWSI model, using a bi-square kernel as the weighting function. In this case, the optimal bandwidth is 98 neighbouring flows, which is the minimum bandwidth that can still be reliably calculated.

7.3.3 Analysis of the model results

This section illustrates and discusses the results of a flow-focused model. For this demonstration, a Poisson flow-focused GWSI model with a fixed bandwidth of 1318 metres and four-dimensional kernel is used. The squared Cauchy kernel has been used as a weighting function. Table 7.1 shows a summary of the results of the flow-focused model for the agglomeration of Lausanne along with some summary statistics. The model has been

Table 7.1: Poisson flow-focused GWSI model, with fixed bandwidth of 1318 metres, and four-dimensional squared Cauchy kernel

Parameter	*Mean*	*Min*	*Max*	*Std Dev*	*Quartiles*		
					25%	*50%*	*75%*
Active pop.	0.8216	0.3969	1.1918	0.0695	0.7808	0.8264	0.8642
Jobs	0.9583	0.0150	1.2543	0.0804	0.9204	0.9727	1.0091
Distance	-1.3100	-2.4916	-0.6021	0.2197	-1.4338	-1.2696	-1.1647
Intercept	0.4554	-6.2670	12.1462	2.4992	-1.2571	-0.1074	1.9504
t-values active pop.	6.5	1.3	71.1	4.7	3.6	5.1	7.9
t-values jobs	8.5	0.1	80.5	5.5	5.0	7.0	10.1
t-values distance	-5.8	-30.7	-1.0	3.8	-7.3	-4.7	-3.2
t-values intercept	0.4	-7.5	15.9	1.7	-0.4	-0.0	0.6
p-values active pop.	0.00	0.00	0.21	0.02	0.00	0.00	0.00
p-values jobs	0.00	0.00	0.96	0.02	0.00	0.00	0.00
p-values distance	0.01	0.00	0.33	0.02	0.00	0.00	0.00
p-values intercept	0.55	0.00	1.00	0.34	0.24	0.62	0.86
Deviance	56081.1	-147672.5	569893.4	38270.5	38395.4	53466.1	70196.8
Pseudo R^2	0.9544	0.6049	0.9996	0.0164	0.9501	0.9569	0.9632

estimated separately for each of the 4900 flows. For each flow, one set of local parameter estimates has been estimated.

The Pseudo R^2 varies between 0.605 and 0.999 with an average of 0.954. Half of the models have a Pseudo R^2 between 0.950 and 0.963, showing that most of the 4900 local models have an excellent fit. The parameter α for active population varies between 0.397 and 1.192, with an average of 0.822, and half of the models have values between 0.781 and 0.864. The parameter γ for the number of jobs shows has a range of values from 0.015 to 1.254 with an average of 0.958. The minimum parameter value of 0.015 is for the internal flow of the small village of Carrouge located at the north-eastern border of the agglomeration; this commune has only 400 workforce and roughly 130 jobs and an internal flow of 87 people. The corresponding p-value of 0.96 also shows that the parameter value is not significantly different from 0. The standard deviation for this parameter is low (0.08), which shows that there is not a lot of variation over all local models.

The distance-decay parameter β ranges from -2.492 to -0.602 with an average of -1.31. The standard deviation of β is higher compared to the α and γ parameters (0.22 against 0.07 and 0.08); half of the models have values in the range of -1.434 to -1.165. The intercept in most models is close to 0, but can be as low as -6.267 or as high as 12.146. The t-values and p-values for the different parameters show that the parameters are in most models significantly different from 0, except the intercept in some cases. Overall, the flow-focused model successfully detects and takes into account spatial variations in relationships. The relatively small bandwidth of 1318 metres, and the significant variation of the model parameters for the different flows shows the presence of spatial heterogeneity in the interaction behaviour in this agglomeration.

Figures 7.6, 7.7 and 7.8 show the complete set of parameter estimates for all 4900 calibration flows using matrix visualisation. The origins and destinations are ordered

descending by size. Figure 7.6 shows parameter α for the active population. While there is no clear pattern visible, bigger communes seem to have slightly higher values for the α parameter, indicating that the size of the active population has a slightly bigger impact on the number of interactions in bigger communes. Also, there does not seem to be different parameter values for the internal flows. For the γ parameter for the number of jobs (figure 7.7), there is also no clear visible pattern, except for the mid-size and smaller communes where the parameter value for the internal flows (diagonal in figure 7.7) has smaller parameter values compared to most other flows. The number of jobs seems to have a smaller effect on the number of intrazonal flows, especially outside the central locations. The distance-decay parameter (β) in figure 7.8 is mostly more negative for internal flows, and also for some mid-size and smaller rural communes, e.g. the communes of Servion, Sullens, or Carrouge. These communes are typical residential communes in more rural parts of the agglomeration. Distance seems to be a bigger obstacle to commuting for flows with one of these communes as an origin or a destination. Distance-decay seems to be closer to 0 (i.e. higher or less negative) for flows towards bigger communes, and also some mid-size towns, e.g. Paudex where Nespresso has a major office.

Figures 7.9 to 7.14 show the values for the β, α and γ parameters for the in- and outflows from/to five selected communes in the agglomeration. The number of commuters is represented by the line width, while the colour corresponds to the parameter value. The five selected communes are the central city of Lausanne, the smaller city of Morges in the western part of the agglomeration, the suburban towns of Lutry (East of Lausanne), Epalinges (North of Lausanne) and Bussigny (West of Lausanne). Figure 7.9 shows the distance-decay parameters for the inflows into the five selected communes. The parameter estimates for the flows to Lausanne are all very similar and relatively high (close to 0), indicating that distance is not important for many commuters to Lausanne, probably due to the attraction of the city, high number of jobs and good transportation system. Some exceptions are the neighbouring towns of Lausanne, where the distance-decay parameter is slightly more negative. Also for flows towards other locations such as Morges or Lutry, the neighbouring towns have generally slightly more negative β values. The reason might be that for some people, distance to work is an important factor for residential choice. The flows from the south-eastern communes of the agglomeration towards Lutry or Epalinges show more negative distance-decay parameter values than average. These communes are known to have a higher percentage of well educated and wealthy population; these people are also able to pay higher rent for homes and can afford to live closer to work, resulting in more negative β values.

Figure 7.10 shows the active population parameters (α) for the inflows. Flows from the outer parts of the agglomeration show generally smaller estimates for α. This is especially the case for the flows towards Lausanne, Epalinges and Morges, but less for flows to Bussigny. The reason behind this pattern might be that for commuting to a city like Lausanne or Morges, it is easier to use public transportation, while it is easy

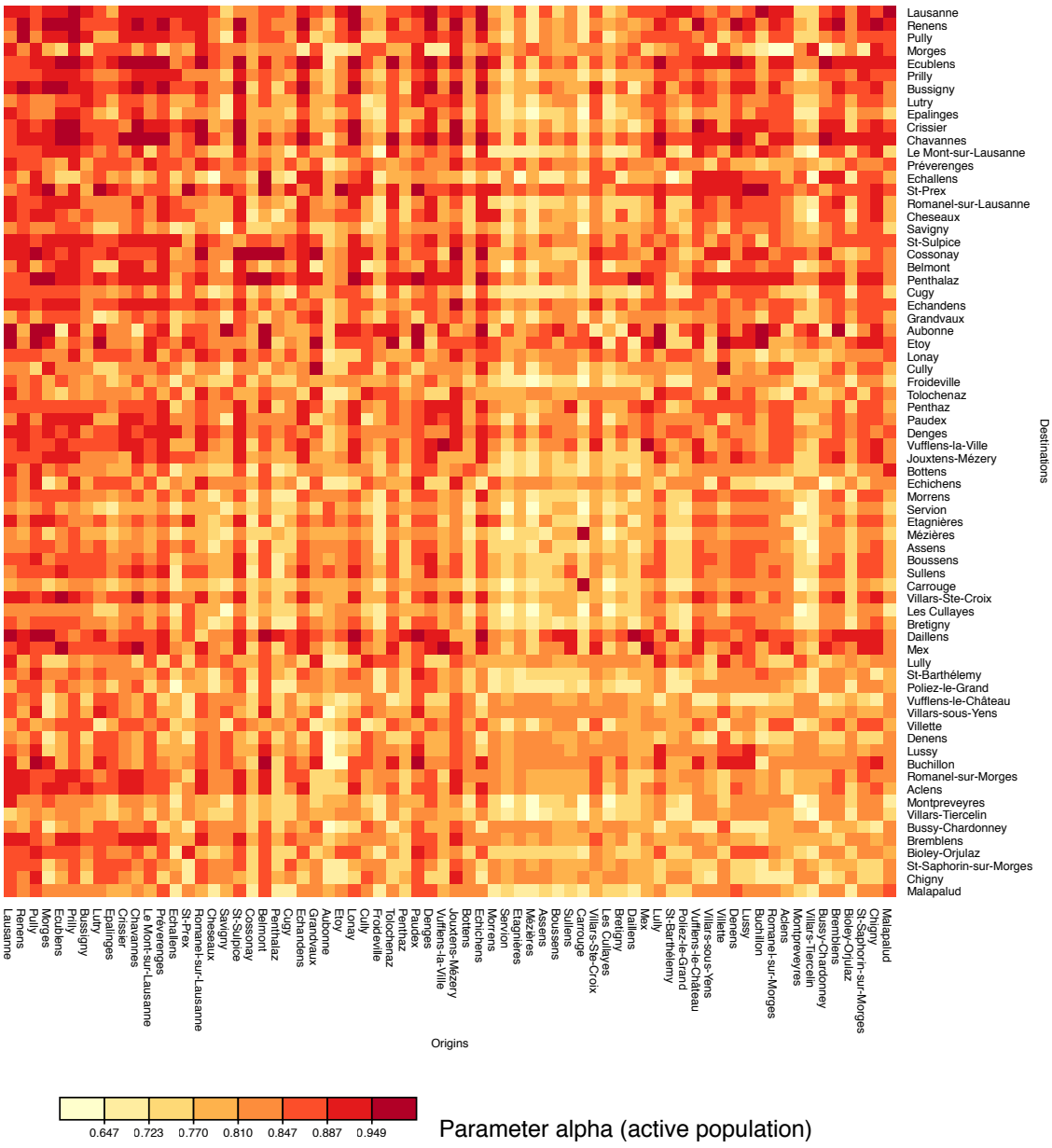


Figure 7.6: Parameter α, active population, using the bandwidth of 1318 metres. Each cell shows the parameter value of one flow-focused model, corresponding to the destination (row) and the origin (column) of each flow. The resulting matrix visualisation shows the parameter values of all 4900 calibrated models.

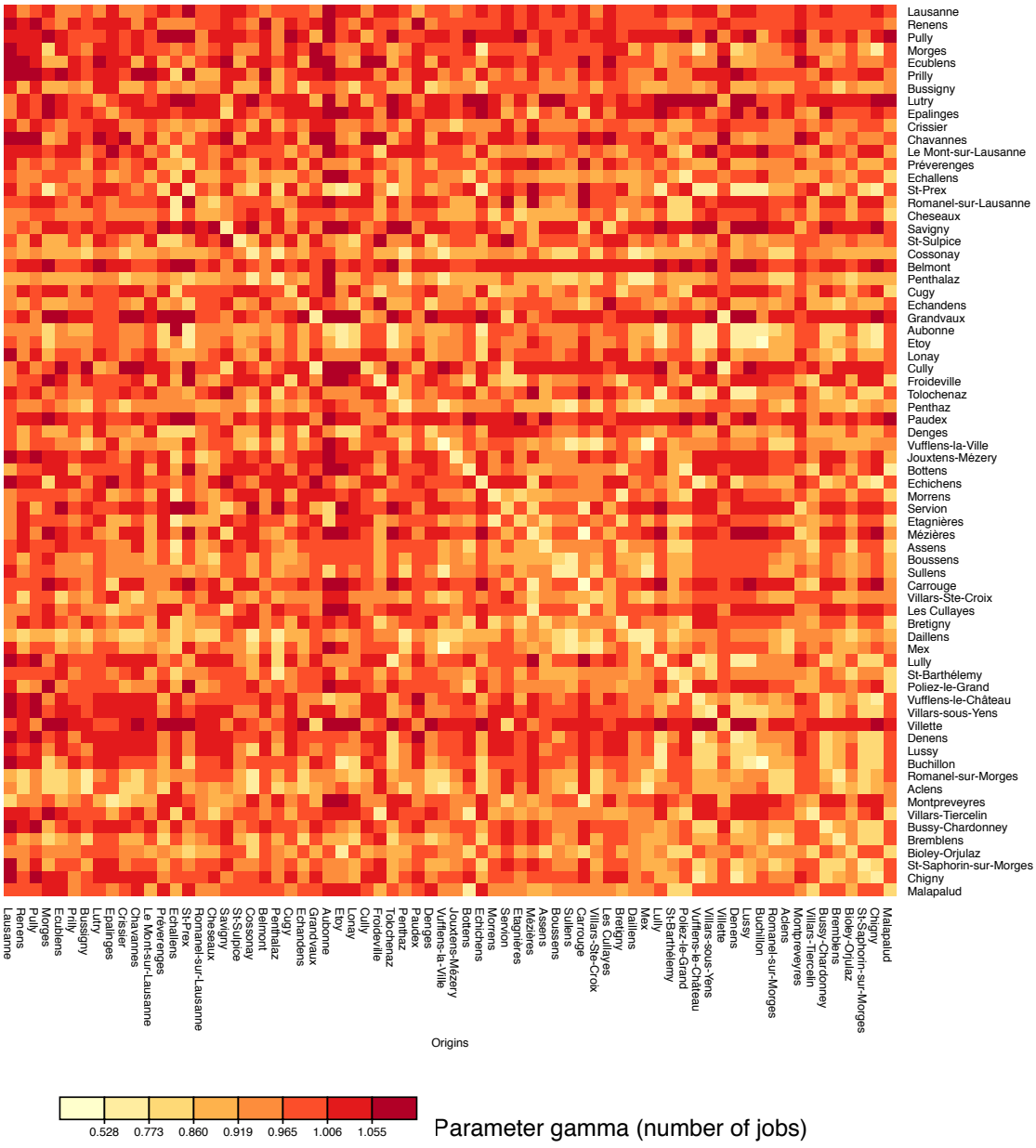


Figure 7.7: Parameter γ, number of jobs, using the bandwidth of 1318 metres. Each cell shows the parameter value of one flow-focused model, corresponding to the destination (row) and the origin (column) of each flow. The resulting matrix visualisation shows the parameter values of all 4900 calibrated models.

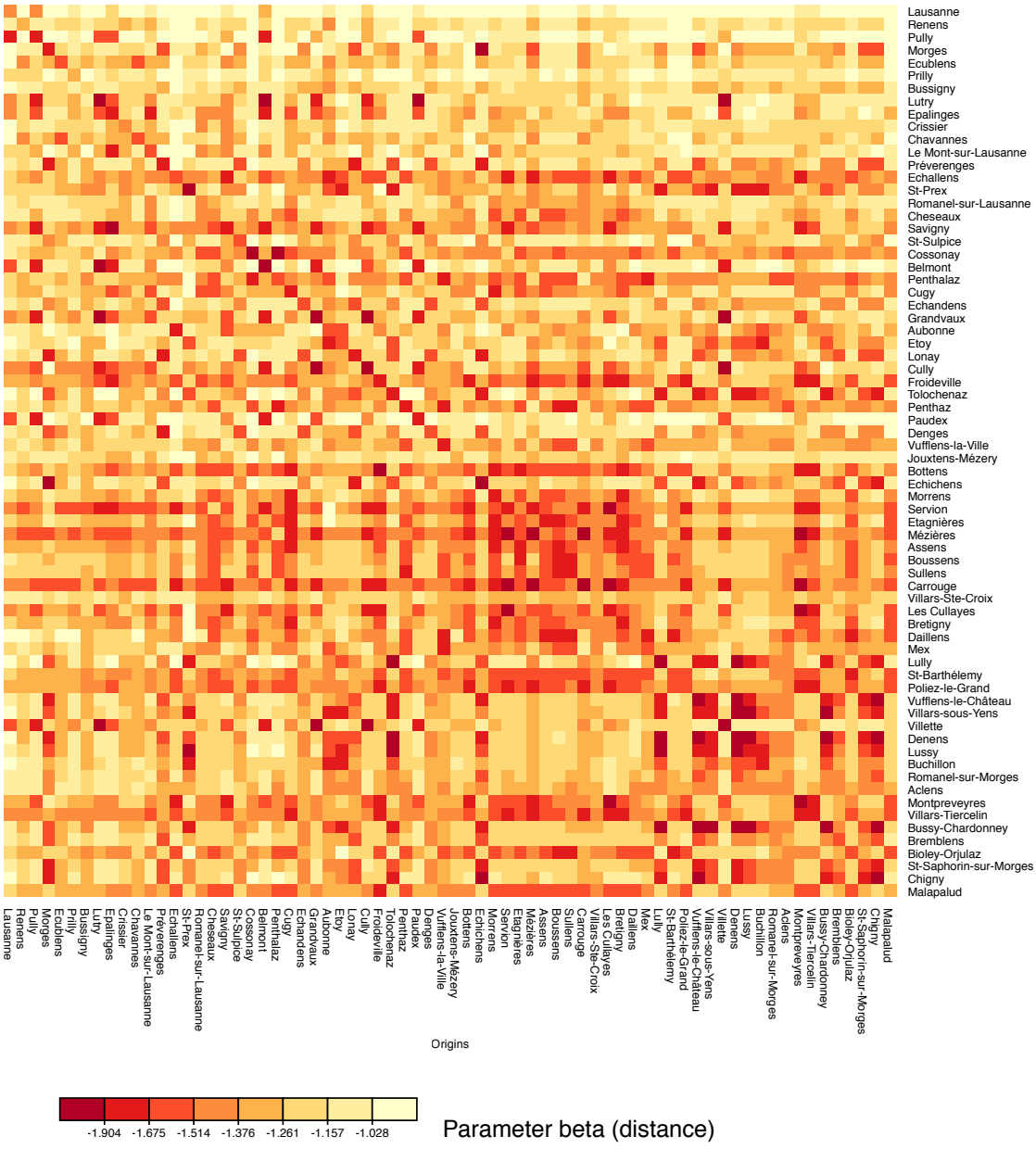


Figure 7.8: Parameter β, distance-decay, using the bandwidth of 1318 metres. Each cell shows the parameter value of one flow-focused model, corresponding to the destination (row) and the origin (column) of each flow. The resulting matrix visualisation shows the parameter values of all 4900 calibrated models.

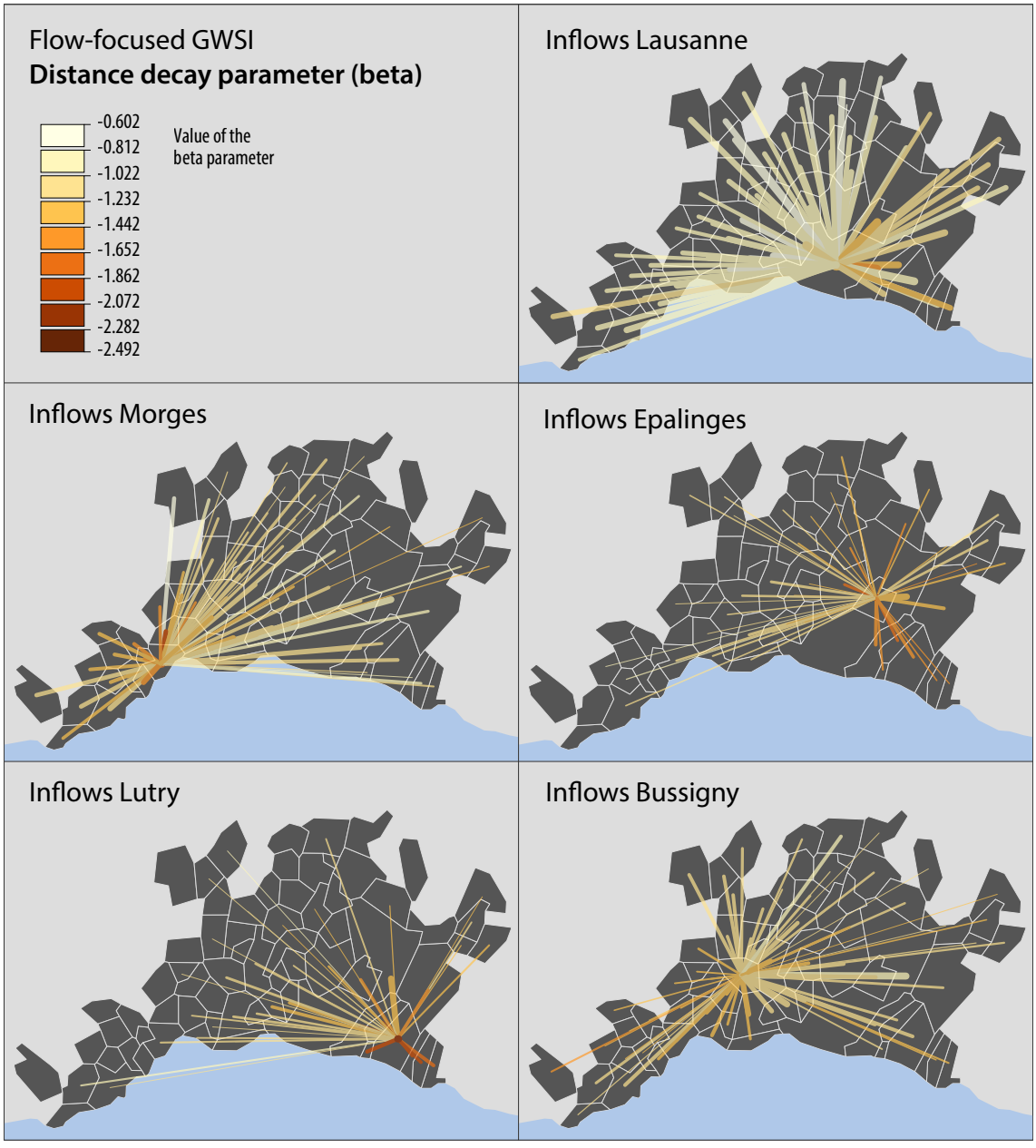


Figure 7.9: Distance decay parameters for inflows in 5 communes in the agglomeration of Lausanne. The value of the parameter estimates are represented by different colours and the width of the lines shows original flow data values (flow size).

to commute by car to Bussigny which is close to the highway. For people living in the smaller communes in the outer parts of the agglomeration, public transportation is less well developed, and people tend to commute less to the crowded cities.

Figure 7.11 shows the number of jobs parameters (γ) for the inflows. The γ values are mostly smaller for closer origins and destinations. There also seems to be a gap from east to west, visible especially for the flows to Lutry and Epalinges. It might be that relatively few people cross the city of Lausanne for work, even though a highway by-passes the city.

Figure 7.12 shows the distance-decay parameter (β) for flows going out of the five selected communes. Again, shorter flows tend to have more negative β values, and there are also some differences between east and west. Figure 7.13 shows the active population parameter for the outflows of the five communes but does not give additional insight into the commuting patterns. The spatial pattern for the number of jobs parameter (γ) for the outgoing flows in figure 7.14 shows again a difference between the flows within the central zone of the agglomeration with slightly higher estimates for γ and communes at the outer parts of the agglomeration having smaller values, indicating that the number of jobs is a less important factor for people commuting from the central part of the region towards the outer zones.

More analysis of the spatial patterns and their meaning would be possible, especially if the model result are combined with other socio-economic data. But this short analysis shows the potential of the flow-focused GWSI method to find regional differences in commuting behaviour. Also again should be emphasised here that the flow-focused GWSI, similar to other variants of GWSI, can also be used for forecasting purposes when the model is calibrated for a pair of origin and destination with no flow in between.

In the following chapter we will discuss some issues related to GWSI models, e.g. adaptive bandwidth selection in the origin- and destination-focused models, along with some empirical examples using the GWSI models.

Figure 7.10: Active population parameters (alpha) for inflows in 5 communes in the agglomeration of Lausanne. The value of the parameter estimates are represented by different colours and the width of the lines shows original flow data values (flow size).

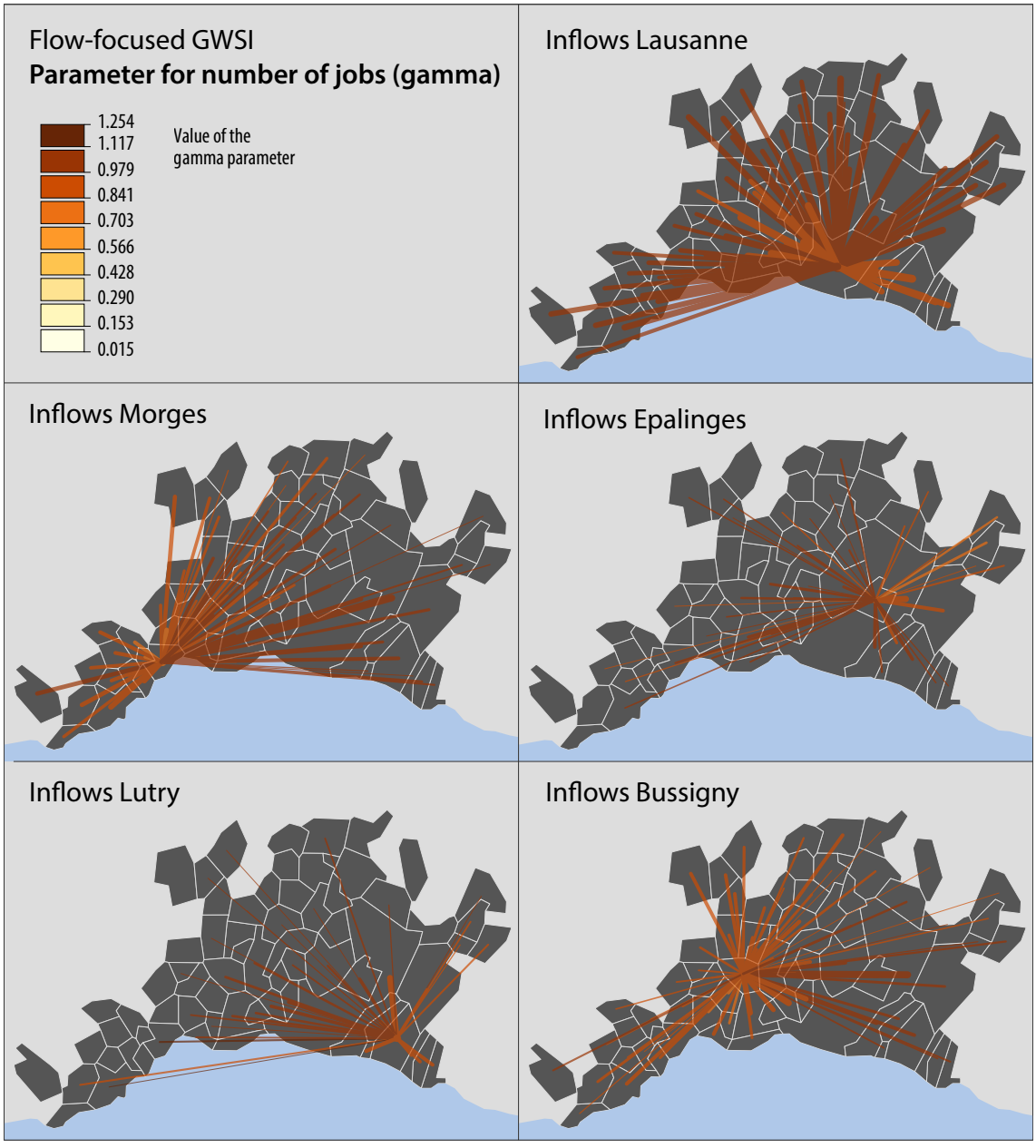


Figure 7.11: Number of jobs parameters (gamma) for inflows in 5 communes in the agglomeration of Lausanne. The value of the parameter estimates are represented by different colours and the width of the lines shows original flow data values (flow size).

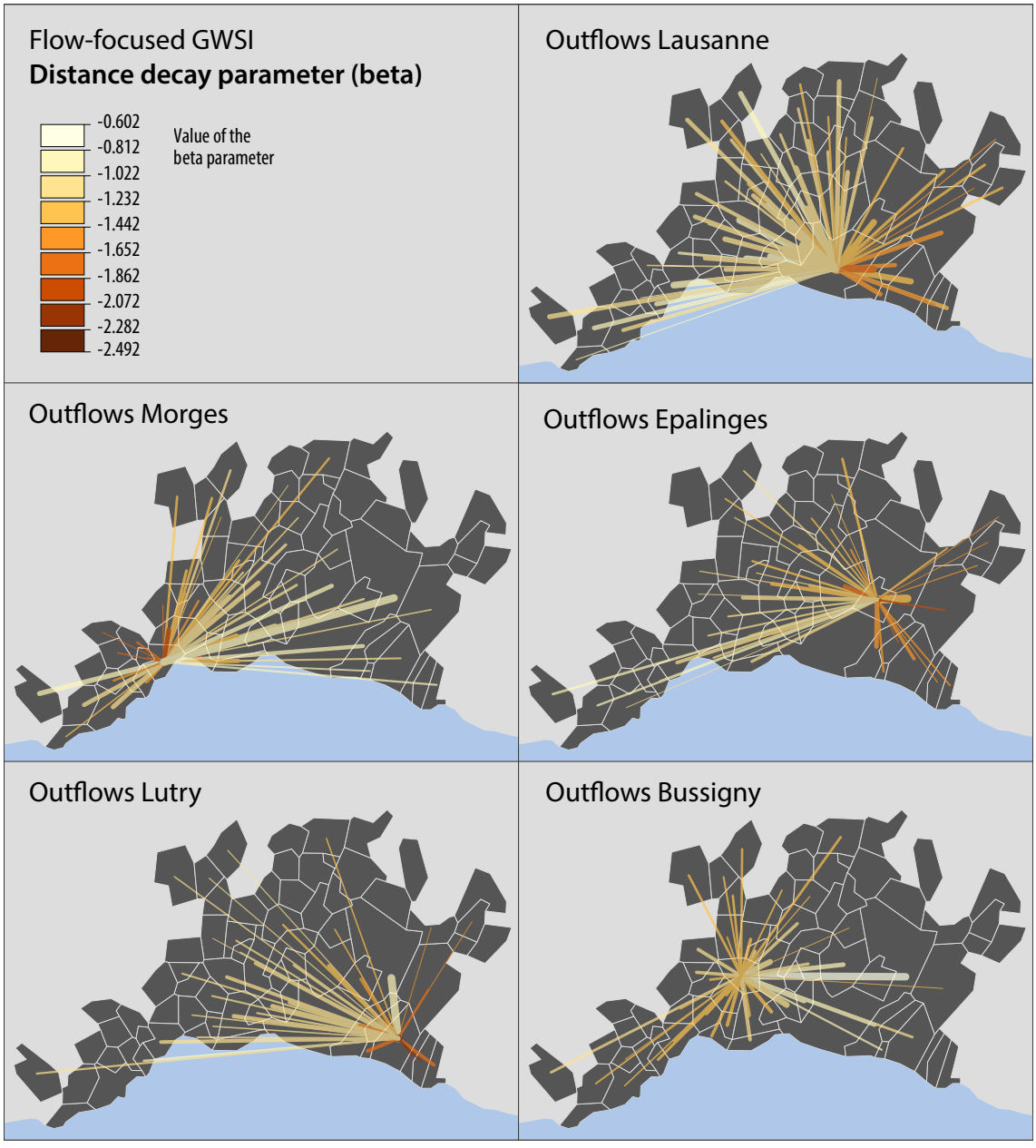


Figure 7.12: Distance decay parameters for outflows from 5 communes in the agglomeration of Lausanne. The value of the parameter estimates are represented by different colours and the width of the lines shows original flow data values (flow size).

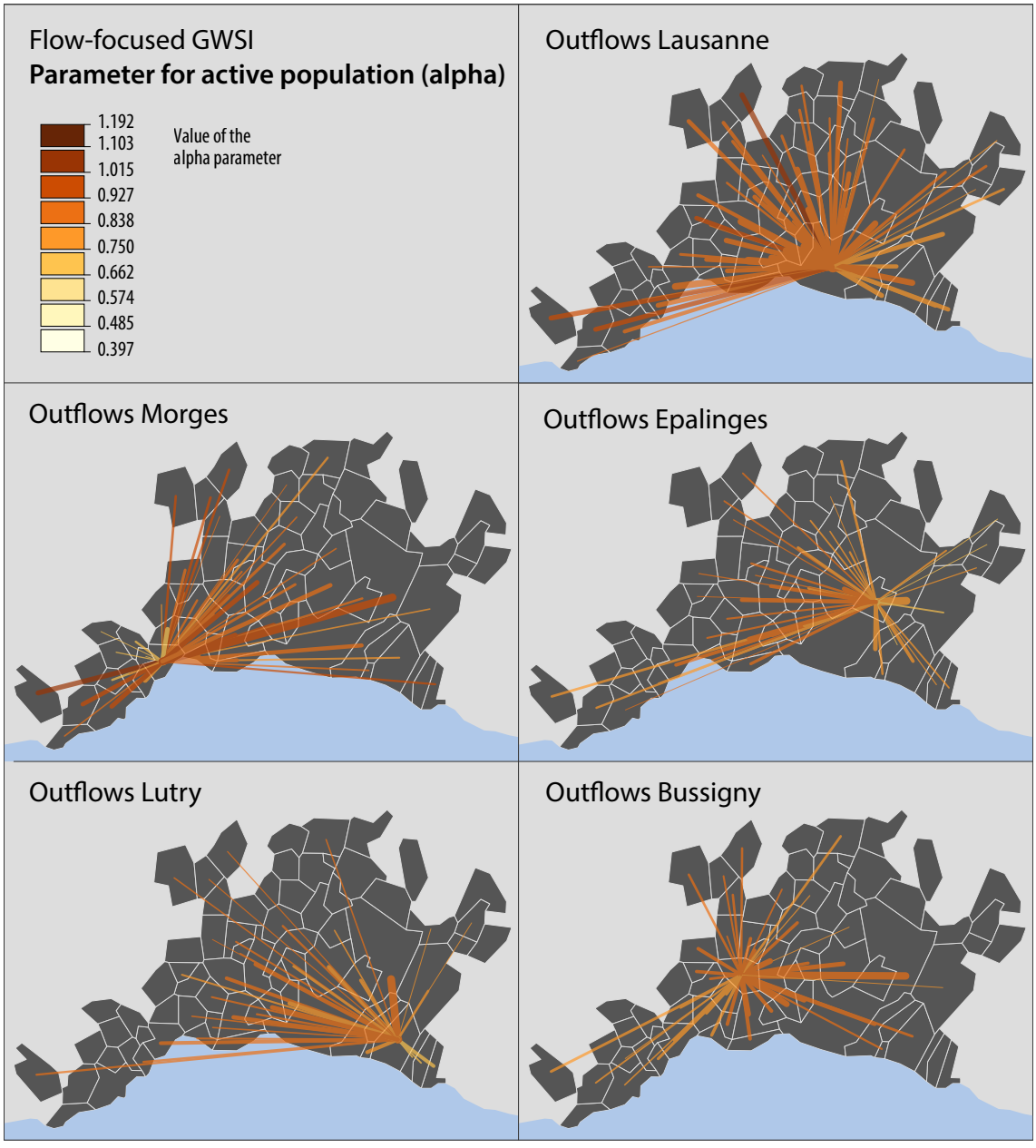


Figure 7.13: Active population parameters (alpha) for outflows from 5 communes in the agglomeration of Lausanne. The value of the parameter estimates are represented by different colours and the width of the lines shows original flow data values (flow size).

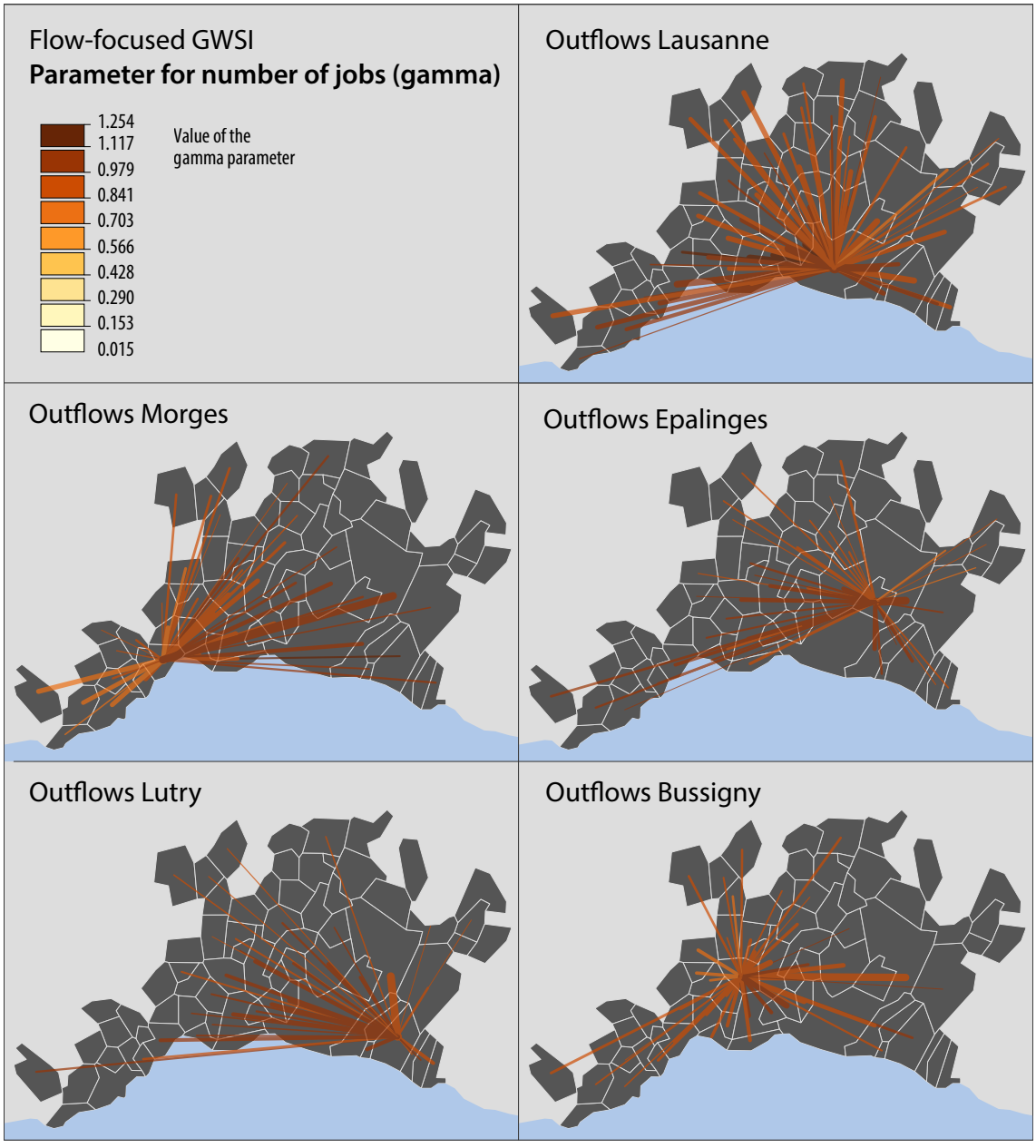


Figure 7.14: Number of jobs parameters (gamma) for outflows from 5 communes in the agglomeration of Lausanne. The value of the parameter estimates are represented by different colours and the width of the lines shows original flow data values (flow size).

Chapter 8

Discussion and examples

The underlying principle of the family of GWSI models and their formulations has been introduced in the previous chapters of this thesis. Furthermore, the local parameter estimates of the GWSI models have been compared to those from global spatial interaction models through a real-world commuting application in Lausanne. In the following chapter we discuss some issues that arise from the geographically weighting spatial interaction models along with some suggestions and discussion on possible approaches to deal with these issues.

8.1 Network distance and travel time approaches

Spatial interaction models incorporate the distance between origins and destinations as one of the explanatory variables. Euclidean distance is frequently used for this purpose; however, it is also possible to replace this by the transportation network distance or travel time. Today, calculating the network distance and travel time is fairly simple as Web services become available for estimating both the fastest and shortest path from one location to another[1]. The travel time estimate often includes the road types and speed limits and sometimes even the traffic conditions. This modification of the distance parameter applies for both global and local interaction models. A feasible way to compute the network and time distances is to query one of the Web services repeatedly using the centroids of all the origin and destination zones as departure and arrival points. The Web service selects automatically the closest point of the road network to the centroid of the zone and then calculate the fastest/shortest path between the two points. The resulting distances can be used directly in the spatial interaction models, in the same way as other distance measures.

We have used the CloudMade Routing Web service for obtaining the fastest route along with network and time distances for the agglomeration of Lausanne. We have then calibrated the global Poisson spatial interaction model shown in equation 3.41 and

[1] e.g. *Google Directions:* `https://developers.google.com/maps/documentation/directions/` or *CloudMade Routing:* `http://developers.cloudmade.com/projects/show/routing-http-api`

Table 8.1: Global Poisson spatial interaction model with different distance measures

Distance measure	α	γ	β	*Intercept*	*Deviance*	*Pseudo* R^2
Density-based scattering	0.820	0.998	- 0.878	-3.982	10254	0.9490
Centroid-to-centroid	0.846	1.011	-0.668	-5.700	11362	0.9453
Road network	0.851	0.984	-0.745	-4.642	12314	0.9443
Time distance	0.810	0.949	-0.667	-6.484	16029	0.9351

Table 8.2: Poisson flow-focused model with different distance measures

Distance measure	*Median* α	*Median* γ	*Median* β	*Median intercept*	*Median deviance*	*Median pseudo* R^2
Density-based scattering	0.876	1.032	-0.867	-4.516	15361	0.9515
Centroid-to-centroid	0.897	1.047	-0.720	-6.277	15457	0.9480
Road network	0.903	1.021	-0.804	-5.072	15967	0.9467
Time distance	0.869	0.991	-0.713	-7.049	20065	0.9368

the local flow-focused GWSI model in equation 7.3 for the Lausanne commuting dataset considering different distance parameters of the density-based scattered distance (see section 4.3.2), the centroid-to-centroid distance, the network and the time distances. The variables of the models are defined in the same way as in equations 3.41 and 7.3 respectfully and intra-zonal flows are excluded from the analysis. In the flow-focused model a squared Cauchy kernel is used for weighting the flows and the bandwidth is set to 1320 metres obtained in section 7.3.1. The results of the global and the flow-focused Poisson spatial interaction models are shown in table 8.1 and table 8.2 respectively.

For all models, the resultant parameter estimates show the expected effect on the total interaction (i.e. distance-decay negative, population and number of jobs positive) and the goodness-of-fit assessed by deviance and pseudo R^2 represent that models are well fitted. However, both deviance and pseudo R^2 suggest that models using the population density-based distance are the best and the goodness-of-fit decreases slightly by using the centroid-to-centroid, the road network and the time distances. This is a rather unexpected result although the differences between the models are small. However, the results only show the behaviour of the models in Lausanne and other example applications should be studied in order to get better insight into this issue.

There are some possible explanations for this behaviour, for instance, the chosen approach of calculating the network distances based on the origin and destination centroids might not give a satisfactory approximation of the reality. Furthermore, the chosen Web service (the one provided by CloudMade based on OpenStreetMap data) might not give time distance estimates with a sufficient accuracy. The other possible source of error is the assumption that all commuters are using a car, while in reality many people use alternative means of transportations. Ignoring the internal flows might also give rise to a biased result. For the GWSI models, it also has to be noted that the flow-focused model uses Euclidean distances for weighting the flows (four-dimensional Euclidean distance in

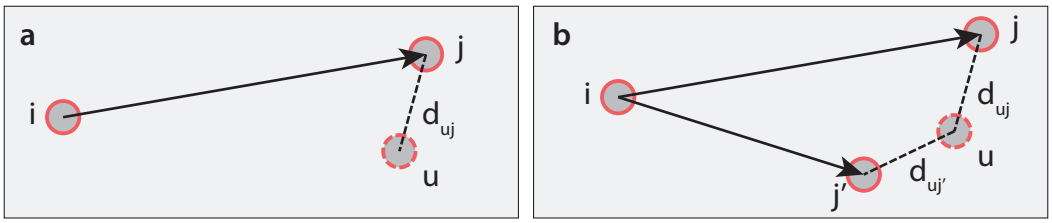


Figure 8.1: Weighting of flows in a destination-focused approach is done using the distance to the destination only.

the case of the flow-focused model). It is unclear if using network/time distances for weighting the flows would be a feasible approach in the case of GWSI models.

8.2 Mixed kernel approach

During calibration of a GWSI model, we give a weight to each of the flows. This weight is calculated based on some distance measures (e.g. distance from the calibration point to each flow's origin on the origin-focused model), and using a kernel function, mostly a Gaussian kernel in GWR or a squared Cauchy kernel in GWSI. In the case of a destination-focused model (see section 6.3), the squared Cauchy kernel function can be written as:

$$W_{u_{ij}} = \left[1 + (\frac{d_{uj}}{b})^2\right]^{-2} \tag{8.1}$$

where d_{uj} is the geographical distance between the calibration point u and destination j of the observed flow ij (see figure 8.1a for an illustration). Using this type of spatial kernel, in geographically weighting spatial flows we consider only distance as the measure of spatial similarity between flows. This means that a flow ij' where $d_{uj'} = d_{uj}$ is weighted the same as flow ij since their destinations are the same distance to the calibration point u (illustrated by figure 8.1b). However, some situations call into question the equal weighting of flows only based on the physical distance to their respective destinations. This might for example happen if destinations j and u are located in the same spatial cluster with many exchanges between j and u, while j' is located outside the cluster with only few exchanges between j' and u. In this case, destination u is probably more similar to destination j than to j', and consequently flow ij should be weighted higher than flow ij'. The question therefore is: how to improve the weighting scheme in GWSI? A possible solution is to include the similarity between destinations into the destination-focused model (in the case of origin-focused model, similarity between origins). We discuss here a possible approach for the case of a destination-focused model, using the *"strength of connection"* between destinations.

8.2.1 Strength of connection as a similarity measure between destinations

As already mentioned in chapter 5, Nissi and Sarra (2011) integrate a parameter called the strength of connection between destinations into the weighting function by multiplying the strength of connection with the spatial distance (see equation 5.23). They defined the strength of connection between destinations u and j as $T_{uj}^2/(T_{\bullet u}T_{\bullet j})$, where T_{uj} represents the flow from u to j, $T_{\bullet u}$ the total of all flows to destination u, and $T_{\bullet j}$ the total of all flows to destination j (see also equation 5.24). They considered the strength of connection between destinations as a measure of similarity. In their equation, flows T_{uj} are powered by 2 and also flows $T_{\bullet j}$ and $T_{\bullet j}$ are multiplied; however, from a mathematical point of view, it is not clear whether powering or multiplying the flows is feasible or what is the meaning of powered or multiplied flows. On the other hand, the strength of connection parameter is then integrated into the weighting function with simply a direct multiplication to the power term which again does not have a mathematical reason (equation 5.23). Furthermore, as already discussed in chapter 5, this approach fails if $T_{uj} = 0$. Also, only a directional flow from u to j is considered in the straight of connection equation, and the flow from j to u is ignored.

We suggest two alternative formulations for the strength of connection:

$$c_{uj} = \frac{T_{uj} + T_{ju}}{T_{\bullet u} + T_{u\bullet}} \quad (a)$$

$$or \tag{8.2}$$

$$c_{uj} = \frac{1}{2}\left(\frac{T_{uj}}{T_{\bullet j}} + \frac{T_{ju}}{T_{\bullet u}}\right) \quad (b)$$

where c_{uj} is the strength of connection of destination u to destination j, T_{uj} is the flow from u to j, T_{ju} the flow from j to u, $T_{\bullet u}$ and $T_{\bullet j}$ are the total of flows to destination u and j respectively, and $T_{u\bullet}$ the total of outgoing flows from u. In formula 8.2(a), as T_{uj} is part of $T_{u\bullet}$ and T_{ju} part of $T_{\bullet u}$, c_{uj} is always smaller or equal to 1. Also since all flows are always 0 or positive, then $0 \leq c_{uj} \leq 1$. We can see c_{uj} as the proportion of flows between u and j compared to the total flows with origin or destination u. In formula 8.2(b), we are summing up the proportion of flows from u to j over all flows going to j and the proportion of flows from j to u over all flows going to u. Multiplying this by $\frac{1}{2}$, the strength of connection is then $0 \leq c_{uj} \leq 1$. In cases where no flows occur between two destinations, c_{uj} will be 0. If the flows between u and j are small compared to the overall flows from and to u, c_{uj} will take a relatively small value. If all flows from and to u (j in formula 8.2(b)) are between u and j, c_{uj} is 1.

8.2.2 Integrating destination similarity into the weighting function

An important question is how to integrate the similarity measure into the weighting function, for example into the Gaussian or squared Cauchy kernel function. Nissi and

Sarra (2011) multiply the strength of connection with the squared geographical distance inside a Gaussian kernel function (see also equation 5.23):

$$W_{u_{ij}} = \exp(-\frac{1}{2}(\frac{d_{uj}}{b})^2 \ \times \ \text{strength of connection}). \tag{8.3}$$

Instead of altering the distance measure directly, we can make a separate kernel for the spatial distance and the similarity measure between flows. If we take the strength of connection c_{uj} as the similarity measure, we can mix the two kernels by computing a weighted average. In the case of a Gaussian kernel function, this gives:

$$W_{u_{ij}} = w_{geo} \cdot e^{-\frac{1}{2}(\frac{d_{uj}}{b})^2} + (1 - w_{geo}) \cdot e^{-\frac{1}{2}(\frac{1-c_{uj}}{b_c})^2} \tag{8.4}$$

where w_{geo} is the weight given to the geographical weighting kernel, d_{uj} is the Euclidean distance between the calibration point and destination j, c_{uj} is the strength of connection, b the bandwidth for the spatial kernel and b_c the bandwidth for the destination similarity kernel (strength of connection in our case). In the case of a squared Cauchy kernel, the mixed kernel can be written as:

$$W_{u_{ij}} = w_{geo} \cdot \left[1 + (\frac{d_{uj}}{b})^2\right]^{-2} + (1 - w_{geo}) \cdot \left[1 + (\frac{1 - c_{uj}}{b_c})^2\right]^{-2}. \tag{8.5}$$

It is also possible to mix two different types of kernels together, for example a squared Cauchy kernel for the geographical distance, and a Gaussian kernel for the destination similarity kernel:

$$W_{u_{ij}} = w_{geo} \cdot \left[1 + (\frac{d_{uj}}{b})^2\right]^{2} + (1 - w_{geo}) \cdot e^{-\frac{1}{2}(\frac{1-c_{uj}}{b_c})^2}. \tag{8.6}$$

All variants of this mixed kernel have two bandwidths, b and b_c; one for the spatial kernel and one for the distance similarity kernel. Both bandwidths need to be calibrated using AICc, BIC or a similar measure. Additionally, w_{geo} can be included into the bandwidth calibration process as third parameter, with possible values between 0 and 1. This last parameter allows for different weighting of the two kernels. By including this parameter into the bandwidth calibration process, the optimal weight for each of the two kernels can be found, giving an interesting insight into the importance of geographical distance in the model compared to the destination similarity. If the dataset is not too large, the calibration of the three parameters can be done by a simple grid search, where all parameters are varied in relatively small regular steps and the AICc or BIC is calculated for each possible parameter combination. Different variants of the mixed kernel (e.g. equation 8.4, 8.5 or 8.6) can also be considered.

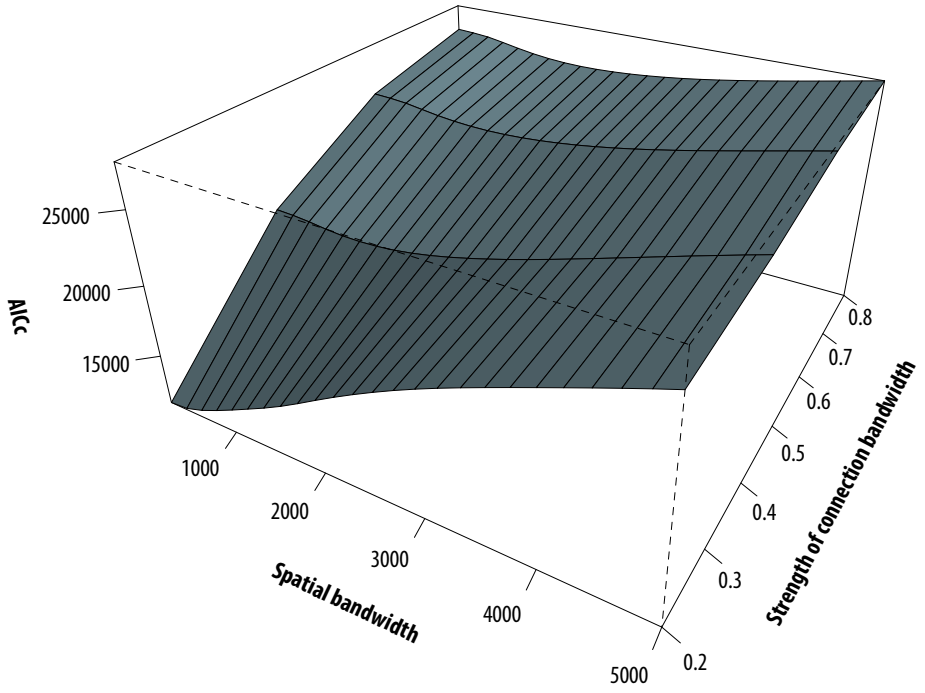


Figure 8.2: Bandwidth optimisation plot using AICc for both spatial and strength of connection bandwidths, with a weight of 0.8 for the spatial kernel and 0.2 for the strength of connection kernel.

8.2.3 Example application of a mixed kernel

In order to illustrate the approach of including the strength of connection into a GWSI destination-focused model, we show an example application to the journey-to-work dataset for the agglomeration of Lausanne. The strength of connection between destinations has been computed according to equation 8.2(a), and a mixed kernel approach with a Gaussian kernel for both spatial and strength of connection bandwidths has been chosen (equation 8.4). The three parameters have been calibrated using AICc. Figure 8.2 shows the AICc against the two bandwidths, with a constant value for w_{geo} of 0.8. The bandwidth calibration yields an optimal bandwidth of 200 metres for the spatial bandwidth, 0.2 for the strength of connection bandwidth, and a weight w_{geo} of 0.8. The optimal bandwidths are both very small: with smaller bandwidths, the model is getting better, which represents a tendency towards a destination-specific model. The calibration of the third parameter, the weight for the spatial kernel w_{geo}, yields the importance of the strength of connection in the model. Figure 8.3 shows w_{geo} against the AICc for a constant spatial bandwidth of 200 metres and constant strength of connection bandwidth of 0.2. The optimal weight is in this case 0.8, indicating that the model is mainly influenced by the spatial kernel.

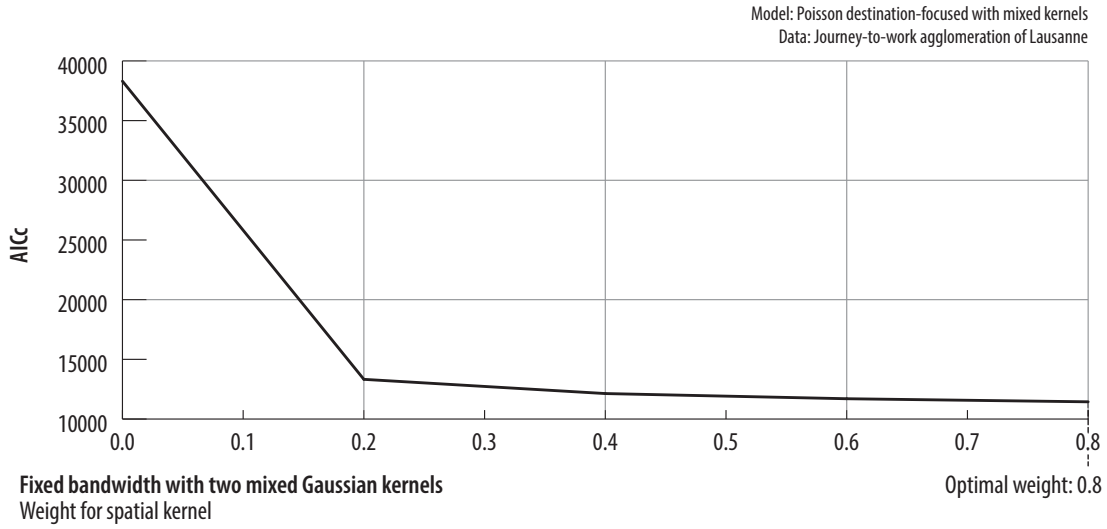


Figure 8.3: Plot showing the AICc for different weights of the spatial kernel, for spatial bandwidth of 200 metres and strength of connection bandwidth of 0.2.

8.2.4 Discussion of the mixed kernel approach

In the case of the presented example application of a destination-focused mixed kernel model including the strength of connection between destinations, this approach did show that the GWSI models can be extended to include some additional non-geographic variables in the weighting function in the model. Through the bandwidth optimisation process, not only the optimal bandwidths can be found, but the strength of connection kernel or the spatial kernel can potentially get eliminated when its weight is zero. In this example, the strength of connection between destinations does not improve the model considerably and the spatial bandwidth indicates that our model is close to a standard destination-focused model. The mixed kernel approach might improve the GWSI model in the case of other datasets, or by using another measure for the strength of connection between destinations. In this section, we have presented the case of a destination-focused GWSI model.

The formula of the mixed kernel approach has one important limitation in the presented form. The calibration point in the GWSI model needs to be an existing destination, because the flow from and to the calibration point to all other origins and destinations must be known. This will somewhat limit the utility of this approach, especially for forecasting flows (i.e. for calibration the local model at a point with no flow). It should be noted here that the presented mixed kernel approach could also be applied to other types of GWSI models such as an origin-focused GWSI model, and potentially the flow-focused models.

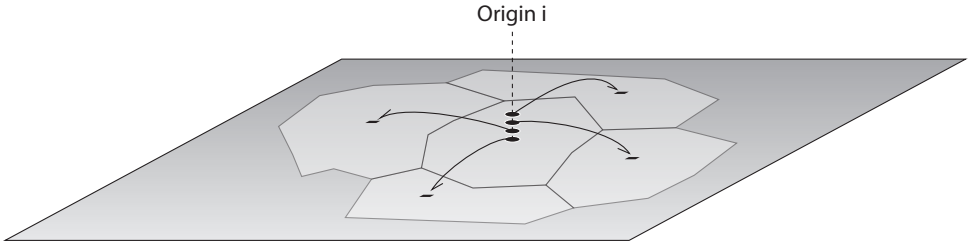


Figure 8.4: Flows from region i stacked at the centroid of the region.

8.3 GWSI: Scattered approach

In the origin- and destination-focused GWSI approaches for local calibration of spatial interaction models, centroids of origin and destination regions have been so far considered as the reference points for geographically weighting the observed flows. In the case of an origin-focused model, if origin i is connected to n destinations, all n outgoing flows from i will obtain the same weight as the distance between their origin (i.e. the centroid of origin region i) to the calibration point u is the same for all flows. Figure 8.4 illustrates this situation where the origins of n flows going to n different destinations are stacked up at the centroid of the region i. This behaviour implies that in origin- and destination-focused approaches, flows are always weighted in groups. For example in a GWSI origin-focused approach, if the model is calibrated at the centroid of existing origin i connected to n destinations then the weight for all these n flows is 1. The same situation is valid for the destination-focused approach when a group of n flows from different origins is piled up at the centroid of the destination j, leading to equal weights for all the n incoming flows to j.

To study the effect of stacking the origins and destinations at the region centroids in both origin- and destination-focused approaches, we consider a scenario where origins and destinations of the flows are scattered randomly within their region polygons. In this situation, the observed flows will have their origins and destinations in different locations rather than to be piled up at the region centroids. When geographically weighting the flows, they will have slightly different weights. The weighting schema is still based on the proximity of origins or destinations of the observed flows to the calibration point. In order to evaluate the differences between the two approaches, we apply the geographically weighted concept on a Poisson origin-focused model using the scattered-based method and compare it with the result of the GWSI origin-focused model presented in section 6.2.1.

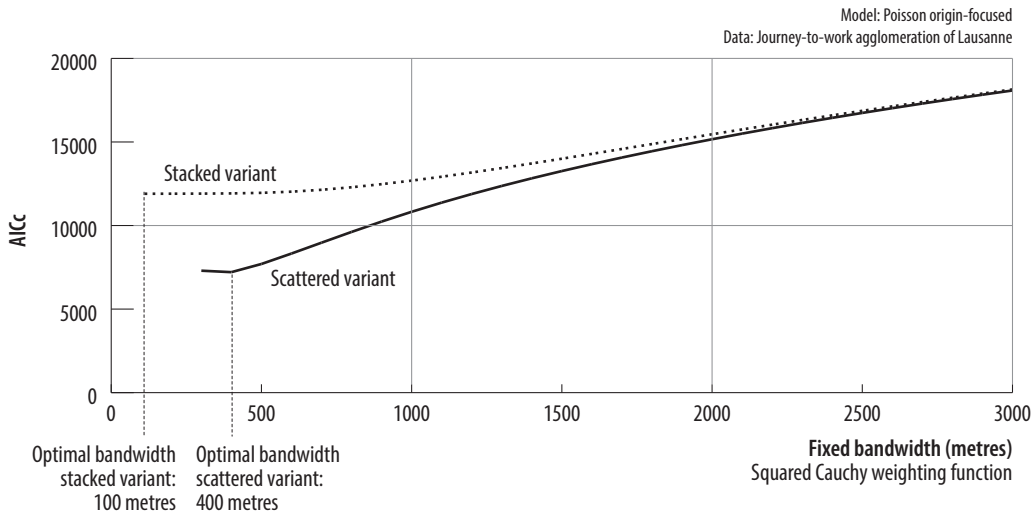


Figure 8.5: Bandwidth value (fixed) against AICc score for the origin-focused Poisson model using stacked and the scattered-based method.

8.3.1 Origin-focused approach for Poisson spatial interaction using scattered-based method

The same model formula as in equation 6.7 will be applied here for the GWSI origin-focused approach for a Poisson spatial interaction model on journey-to-work data in Lausanne. The model is calibrated in each origin commune where the origins and destinations of the observed flows are located randomly within their corresponding regions. The squared Cauchy kernel is used as the spatial weighting function for the geographically weighting of the flows and the selection of the optimal bandwidth has been performed using AICc. Figure 8.5 shows the bandwidth values against the AICc score along with the AICc scores for the stacked model. In the case of the scattered-based model, the optimal bandwidth is equal to 400 metres and for the stacked model, the optimal bandwidth is 100 metres. This result was expected since in the scattered model, since the origins are not located at a same location as in a stacked model but they are scattered within the zone, so the spatial kernel does not wrap around one single location and obtains a bigger bandwidth.

The results of the calibrated model are listed in table 8.3. Compared to the results of the stacked variant of the origin-focused model (table 6.1) in the agglomeration of Lausanne, the parameter values vary slightly. The biggest difference is in the parameter estimates for the distance variable which has a median value of -1.58 for the stacked approach and -1.83 for the scattered variant. Interestingly, the p-values indicate that for all parameters in the scattered approach the majority of models have significant values, contrary to the stacked approach where the active population and intercept estimates are mostly not significantly different from 0. Most of the models resulting from the scattered approach are significant and show very good Pseudo R^2 values of 0.92 or more for

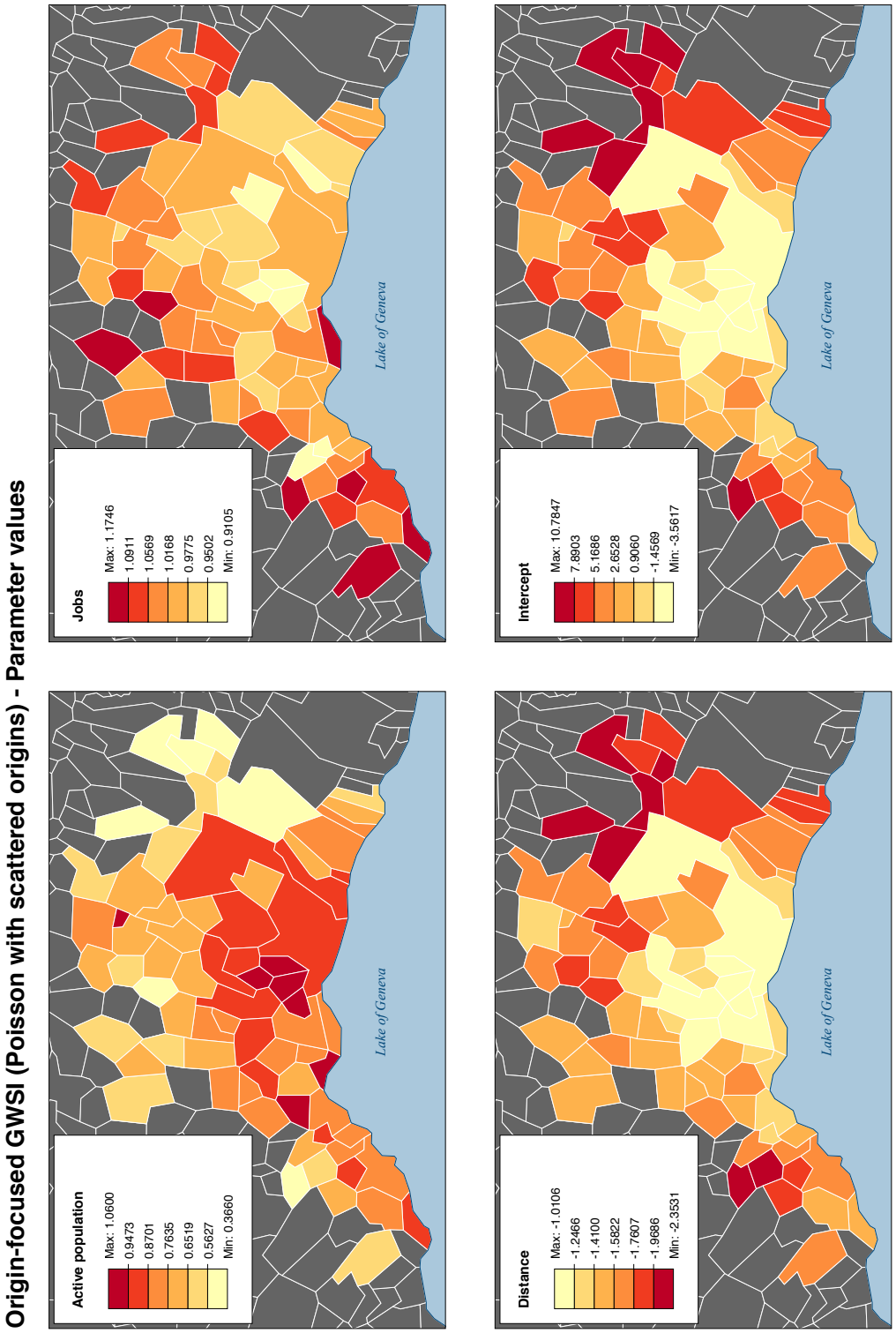


Figure 8.6: The median parameter estimates for origin-focused Poisson model using scattered-based method.

Table 8.3: Origin-focused GWSI (Poisson with scattered origins)

Bandwidth	400 metres (fixed)						
Parameter	*Mean*	*Min*	*Max*	*Std Dev*	*Quartiles*		
					25%	*50%*	*75%*
Active pop.	0.7539	-0.4837	1.3198	0.1878	0.6252	0.7589	0.8910
Jobs	1.0158	0.4957	1.6471	0.1010	0.9504	1.0181	1.0800
Distance	-1.5997	-2.7076	-0.0277	0.3393	-1.8437	-1.5784	-1.3812
Intercept	3.1310	-13.1462	14.7351	3.9885	0.4505	2.9231	6.0377
t-values active pop.	6.3	-1.1	30.3	4.9	2.6	5.0	8.8
t-values jobs	13.3	1.6	78.5	8.4	7.7	11.2	16.2
t-values distance	-10.3	-47.7	-0.1	5.7	-12.9	-9.0	-6.2
t-values intercept	1.4	-11.4	16.2	2.5	0.2	1.3	2.7
p-values active pop.	0.06	0.00	0.87	0.15	0.00	0.00	0.01
p-values jobs	0.00	0.00	0.12	0.00	0.00	0.00	0.00
p-values distance	0.00	0.00	0.91	0.02	0.00	0.00	0.00
p-values intercept	0.24	0.00	1.00	0.29	0.00	0.09	0.40
Deviance	-28978.5	-399250.5	956985.1	114814.6	-99331.9	-17485.8	51242.0
Pseudo R^2	0.8922	-0.1360	0.9988	0.1073	0.8825	0.9271	0.9466

the majority of the models, which is higher than for the stacked approach model with a median Pseudo R^2 of 0.90. The scattered approach shows smaller variation (standard deviation) for the parameters and also Pseudo R^2. Figure 8.6 shows the median parameter values for each origin commune for all variables and figure 8.7 illustrates the respective t-values of the parameters. Most parameters show high t-values, especially in the centre of the agglomeration. In short, scattering the points for geographically weighting has an impact on the optimal bandwidth selection. In the case of this dataset, a bigger optimal bandwidth resulted for the scattered approach, leading to a more stable regression result (less variation in parameters, most t-values and Pseudo R^2). Overall, a better model fit seems to be the result of the scattering approach.

8.4 Adaptive bandwidth

As discussed earlier in chapter 5, the spatial kernel in geographically weighting can be fixed in terms of shape and magnitude over space or it can vary spatially (Fotheringham et al., 2002). This spatial kernel variation in size can be based on the density of the data in space so that the bandwidth parameter in the kernel will be larger where the density of data is low and smaller in regions where the density of data points is high. To investigate the adaptive bandwidth in GWSI, we have calculated the adaptive bandwidth for an origin-focused model using the AICc method for the journey-to-work dataset of Lausanne. Figure 8.8 illustrates the adaptive bandwidth for Poisson GWSI origin-focused approach (see equation 6.7), where the number of data points (in GWSI the number of flow observations) are plotted against the AICc score. The geographical weighting function used for this example is a bi-square kernel (see equation 5.6).

The minimum AICc value occurs at the optimum adaptive bandwidth of 71 to 140 flows, the lowest band. In the origin-focused GWSI model, each calibration point (each

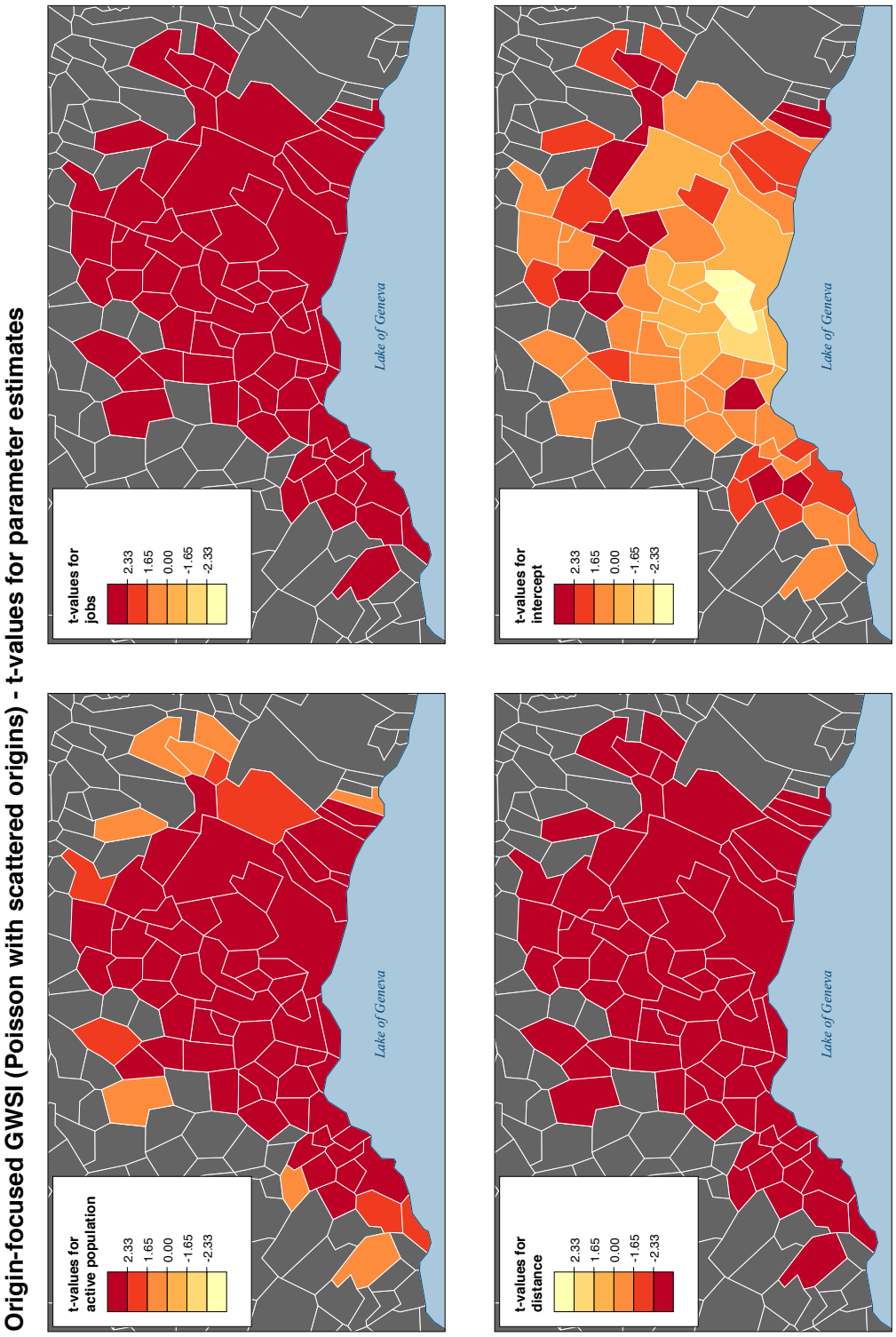


Figure 8.7: The median t-values of parameters for origin-focused Poisson model using scattered-based method. These maps display absolute t-values, where values greater than 2.33 are significant at a level of 99%, and values greater than 1.65 are significant at a level of 95%. For negative parameter values (for the distance-decay parameter), the negative t-values of -2.33 and -1.65 correspond to the significance levels of 99% and 95% respectively.

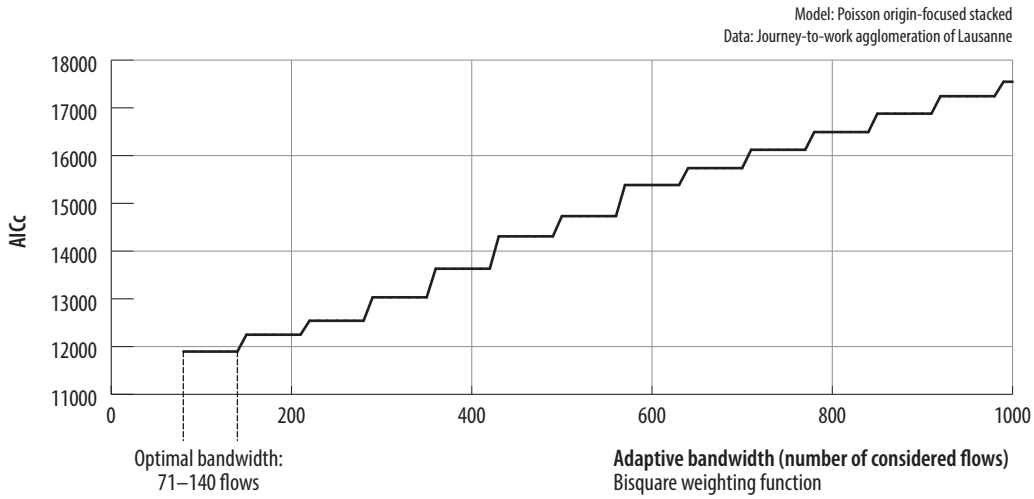


Figure 8.8: Bandwidth value (adaptive; number of flows) against AICc for origin-focused Poisson model.

origin i) is connected to n destinations (in the Lausanne dataset, 70 destinations), so if we divide the adaptive bandwidth by the number of *destinations* (i.e. number of flows with origins stacked up at the calibration point), it shows the number of origins that are considered in the calibration process of origin i (here maximum 2 neighbouring origins). Figure 8.8 shows well the fact that in the origin-focused model, neighbours can only be included by sets of 70 points due to the stacked nature of the origins. We have calculated the adaptive bandwidth for the GWSI origin-focused model with scattered origins to investigate how the number of observations included in the model calibration changes. Figure 8.9 shows the bandwidth values plotted against the AICc scores and the optimum adaptive bandwidth is equal to 140 flows corresponding again to 2 origins. In GWSI, both fixed and adaptive bandwidth approaches are possible. In the case of the journey-to-work dataset for the Lausanne agglomeration, no big difference seems to occur between the two approaches.

8.5 Application examples of the GWSI models

In order to illustrate possible use cases of the GWSI models, we show two different examples in this section.

8.5.1 Evaluate impact of new business centre

In this first example, we study the impact of a new (fictional) business centre that could be located near the city centre of Lausanne. We assume that this new business centre creates 1000 new jobs and we are interested in forecasting where the workers would come from according to a spatial interaction model. Figure 8.10 shows the location of the new

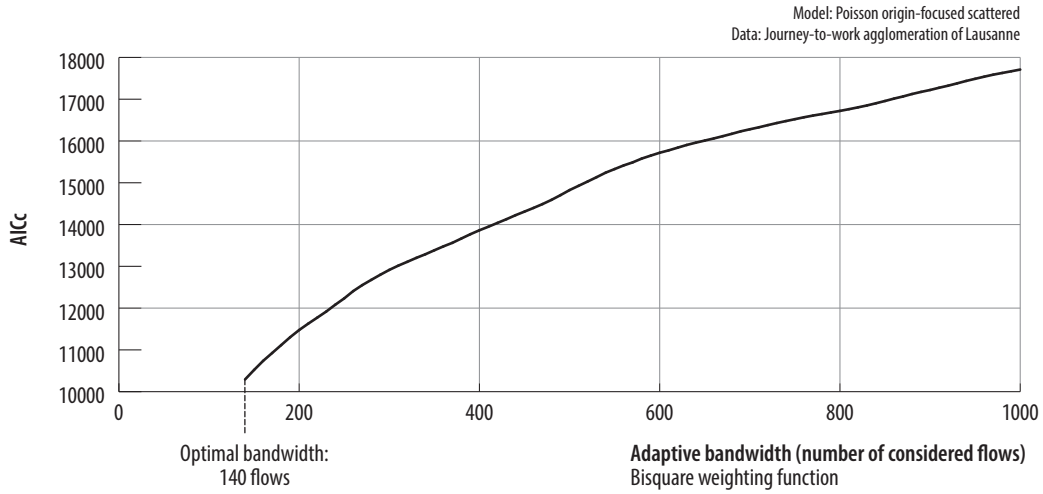


Figure 8.9: Bandwidth value (adaptive; number of flows) against AICc for origin-focused Poisson model using scattered approach.

business centre. Its location is strategic as it is quite close to the motorway and within good proximity to the city centre to guarantee good accessibility for potential workers.

We use a destination-focused model with a squared Cauchy kernel as a weighting function for this example. The bandwidth of the destination-focused model is calibrated in a first step using BIC and AICc scores, using all journey-to-work flows for the agglomeration of Lausanne. We use a fixed bandwidth approach. In our case, the bandwidth tends towards a specific model, as both BIC and AICc scores decrease with decreasing bandwidth (figure 8.11), so then we use the smallest calculated bandwidth of 100 metres. In a second step, we calibrate the destination-focused model for the business centre as a destination, yielding local estimates for the model parameters α, γ and β. We get a value of $\alpha = 0.779$, $\gamma = 1.008$ and $\beta = -1.044$ with the intercept value being -2.593. The value for the pseudo R^2 is 0.957. A global unconstrained interaction model yields values of $\alpha = 0.791$, $\gamma = 0.949$, and $\beta = -1.297$, with an intercept of 0.818. The values for the local destination-focused model for the distance-decay parameter β is less negative compared to the global model. This means that for this specific location, the workers accept to travel longer distances. The local parameter values can now be used to predict the individual journey-to-work flows from each of the 70 communes towards the business centre; figure 8.12 shows these predictions graphically. As expected, the biggest commuting flow would be from the city of Lausanne, and from the neighbouring communes around the business centre.

Instead of using a destination-focused model, we can also predict the flows to the business centre by using an origin-specific destination-focused model. In this case, a separate bandwidth is calculated for each origin, resulting in an even more localised model. Instead of calibrating the parameters on the whole dataset of 4900 flows, we now

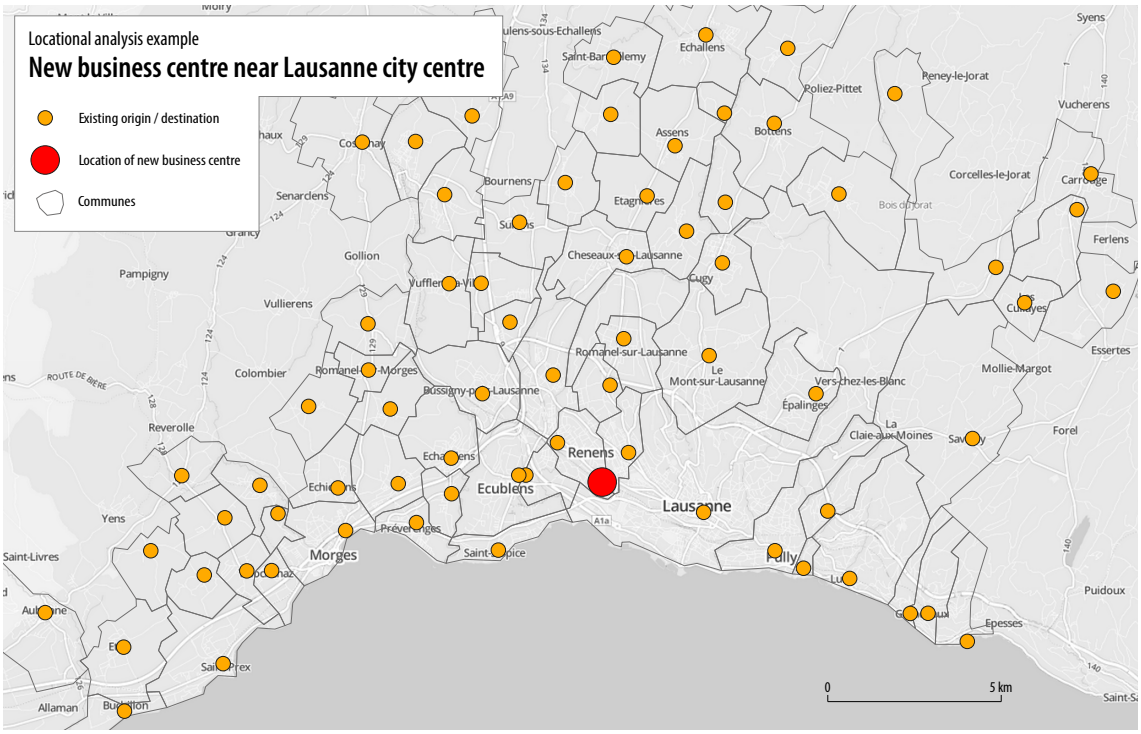


Figure 8.10: Location of the fictional new business centre in the agglomeration of Lausanne

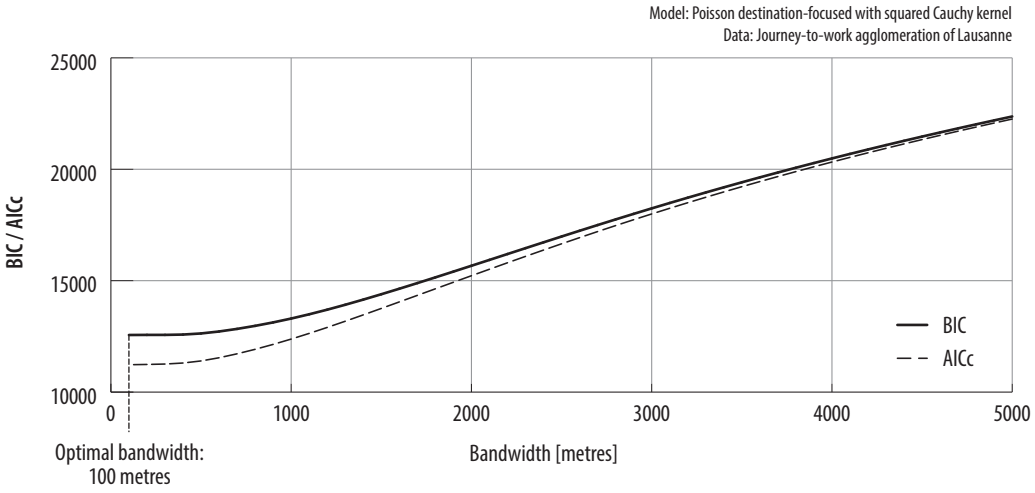


Figure 8.11: Bandwidth selection plot for the destination-focused model using a squared Cauchy kernel.

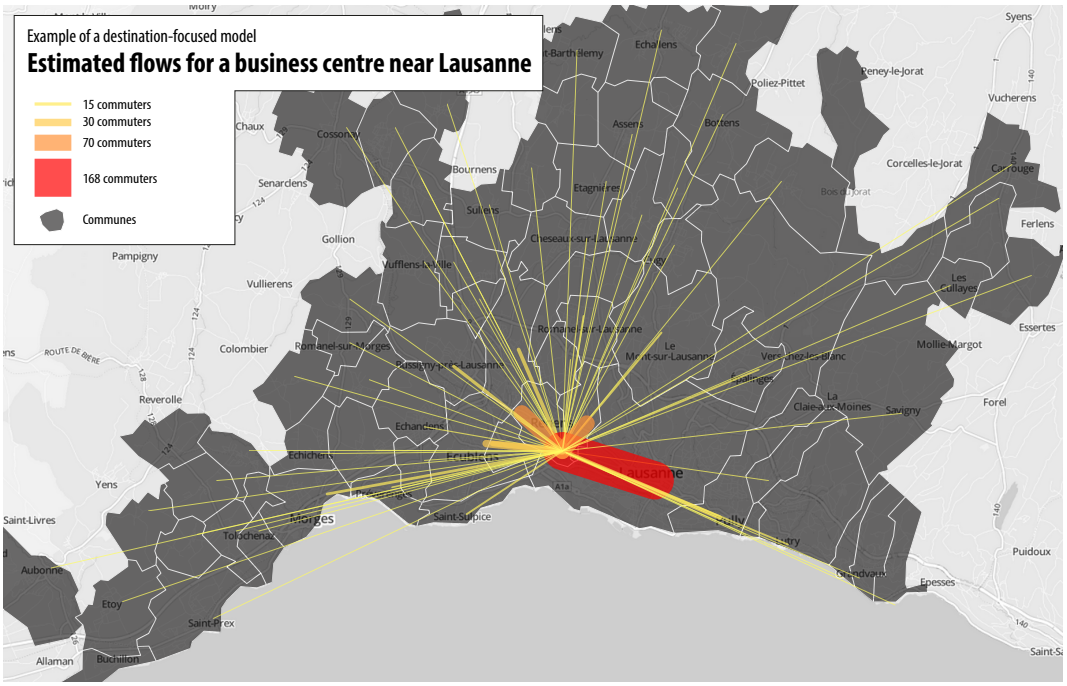


Figure 8.12: Estimated flows based on the destination-focused model calibrated at the location of the business centre.

calibrate 70 local models with each time 70 flows only, resulting in a set of parameter estimates for each flow. Table 8.4 summarises the results of the 70 local models. The γ parameter has a similar median value compared to the destination-focused model (1.069 against 1.008). The median of the distance-decay parameter has a value of -1.823, which is considerably more negative than for the destination-focused model (-1.297).

Table 8.4: Poisson origin-specific destination-focused GWSI model

Parameter	*Mean*	*Min*	*Max*	*Std Dev*	*Quartiles*		
					25%	*50%*	*75%*
Jobs	1.094	0.616	1.381	0.117	1.025	1.069	1.169
Distance	-1.816	-2.689	-0.851	0.396	-2.039	-1.823	-1.501
Intercept	9.330	3.425	18.145	3.032	7.140	9.139	11.274
Pseudo R^2	0.892	0.626	0.996	0.078	0.858	0.906	0.951

Both destination-focused and origin-specific destination-focused models yield unconstrained estimates of flows towards the business centre, so that the total number of estimated flows does not necessarily sum up to the number of jobs in the buisness centre (1000 in our assumption). A simple correction can be the application of a multiplication factor constraining the total number of flows to correspond to the number of jobs. This multiplication factor c can be calculated as $c = \sum N/\hat{T}$, where $\hat{T}$ is the number of estimated flows, and N the number of jobs. Figure 8.13 shows the estimated flows predicted using the origin-specific destination-focused model with a total flow constraint.

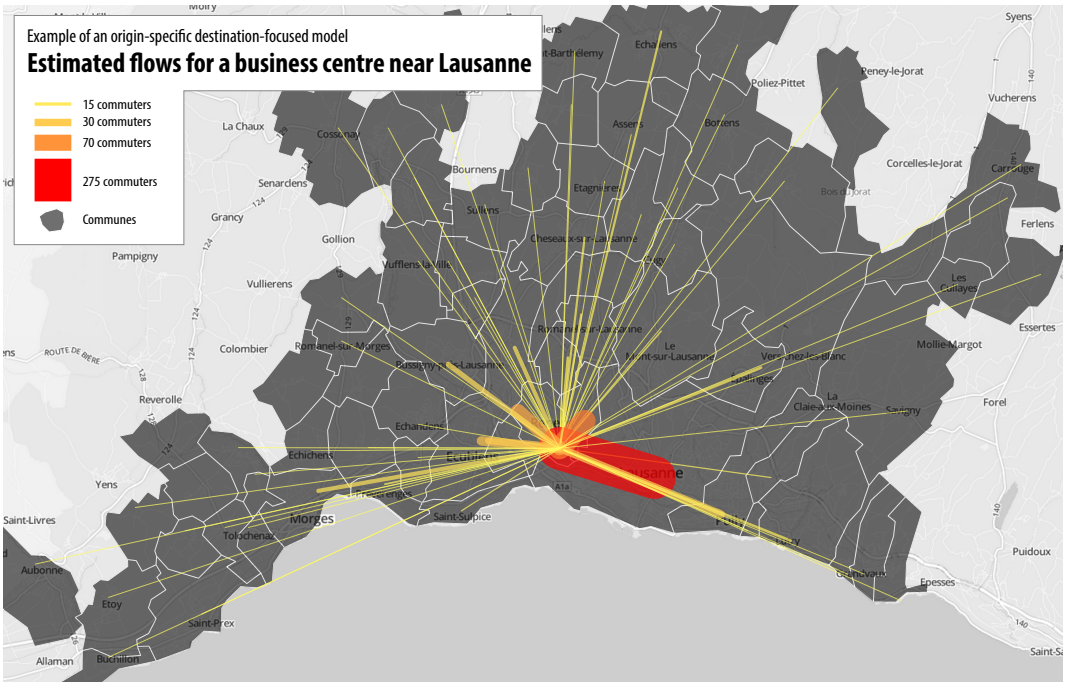


Figure 8.13: Estimated flows based on the origin-specific destination-focused model calibrated at the location of the business centre, corrected using a total-flow constraint.

8.5.2 Estimating commuting flows based on historic data

In Switzerland, the commuting flows have been measured in 1990 and 2000. More recent data are currently not available, while data both on active population and number of jobs are available for 2010. We can use the flow-focused model to estimate the flows for 2010 based on the commuting data for 2000. A squared Cauchy kernel is used as weighting function. The optimal bandwidth is estimated using AICc for the flow-focused model using the commuting data of 2000 for calibration; the optimal bandwidth is 1320 metres (figure 8.14).

Once the optimal bandwidth has been found, we can calibrate the model for each individual flow, and we get a set of parameters local to each flow, as already seen in chapter 7. The set of parameters can then be used to estimate the flow for 2000, and assess the estimation error as the flows for 2000 are known. Figure 8.15 shows the estimated flows for 2000 on top left, and the the estimation error on top right as a origin-destination matrix. The overall model quality is excellent, with only a few errors. For each calibration flow, pseudo R^2 values can be computed. Average pseudo R^2 is 0.954 with a standard deviation of 0.016; half of the models have values between 0.950 and 0.963.

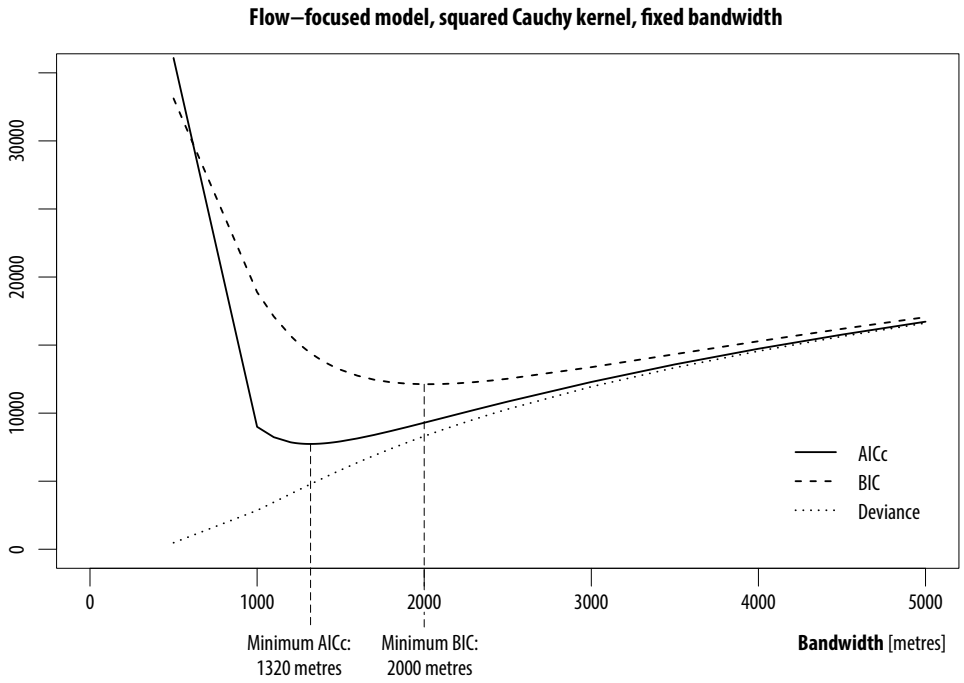


Figure 8.14: Bandwidth selection plot for the flow-focused model with squared Cauchy kernel.

Using the calibrated flow-focused model for each flow with active population and number of jobs for 2010 instead of 2000, we can make an estimation of the expected flow between any origin and destination for 2010 (figure 8.15 bottom left). We can also estimate the increase or decrease of commuters between each origin and destination (figure 8.15 bottom right). In figure 8.15, the origins and destinations are ordered by the total number of flows, communes having bigger flows being on the top left of the matrix. The bigger communes all show an increase in the number of commuters, for both incoming and outgoing flows. This is mainly due to an increase of active population between 2000 and 2010, as shown in figure 8.16 (left), and also to the increase in the number of jobs (figure 8.16, right). All communes in the agglomeration show an increase in active population with the city of Lausanne increasing by more than 20,000 people. Also in most communes, the number of jobs increased; however, a small decrease could be observed in some places. Overall, at the level of the agglomeration, both the total number of active population and jobs increased resulting in an increase in the total number of commuting flows.

8.5.3 Discussion of locational analysis examples

Two different examples of the use of a local spatial interaction model have been shown in sections 8.5.1 and 8.5.2. Many other applications are possible. For example, if we

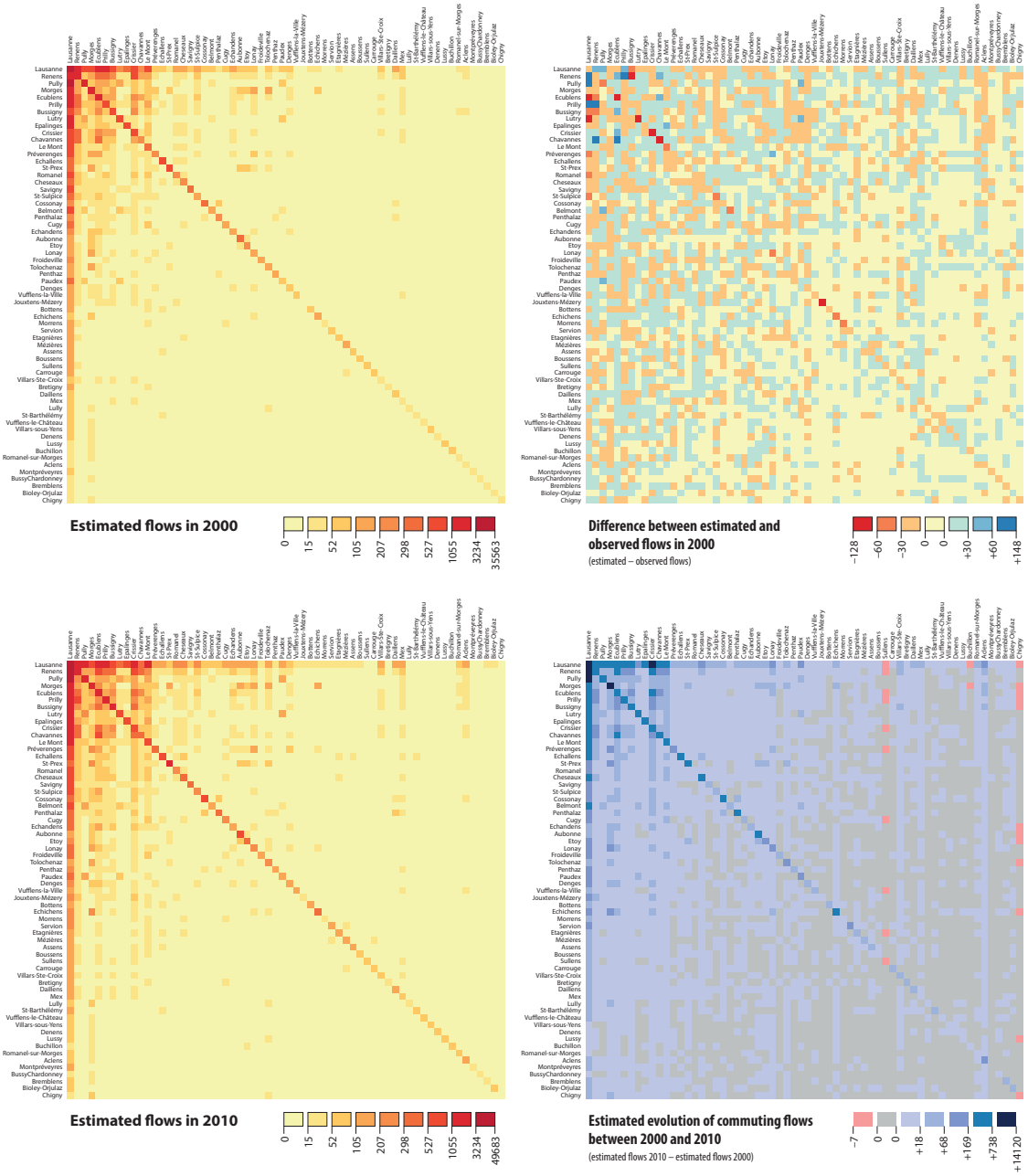


Figure 8.15: Estimated flows in 2000 based on the flow-focused model (top left), difference of estimated and observed flows in 2000 (top right), estimated flows in 2010 (bottom left) and estimated increase in commuting flows from 2000 to 2010 (bottom right).

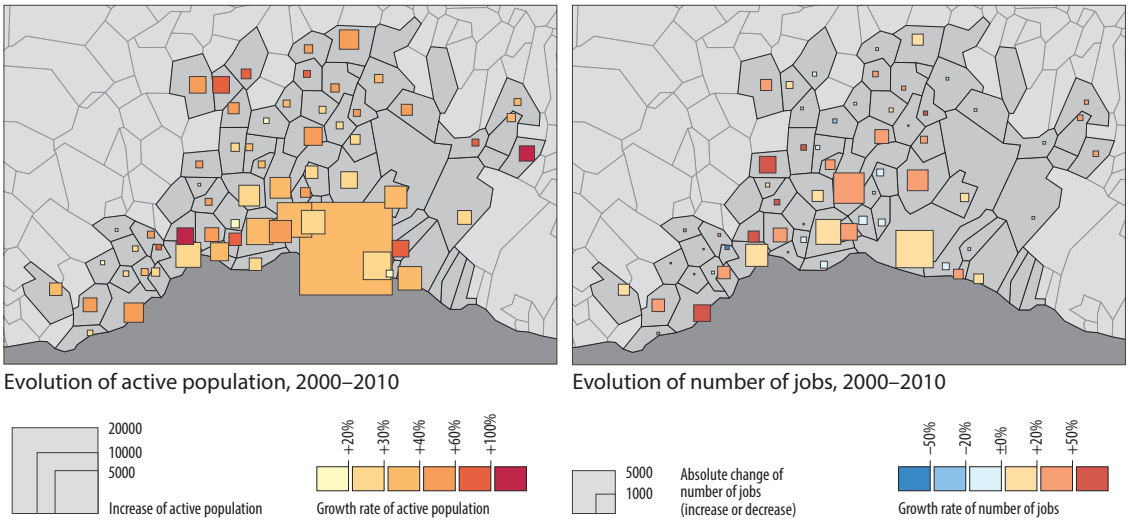


Figure 8.16: Evolution of active population 2000-2010 (both absolute values and growth rate), and of the number of jobs 2000-2010 (both absolute values and growth rate).

use network distances or travel times inside the spatial interaction model, the impact of improved transportation infrastructure on the journey-to-work flows could be studied. Other examples of GWSI models in the context of commuting interaction can be the analysis of the decision process during the search for a new home, given a known location for the workplace and during the search for a job when the residential place is fixed. The connection between residence and workplace is the core consideration in commuting models (see Clark et al., 2003). Although several factors can affect commuting behaviour, the cost of travel, job viability and housing price are three important factors to be considered in the commuting analysis.

The journey-to-work problem can be seen through two different frames. First, when the workforce are resident in a area, for instance owning their house, and are looking for suitable jobs nearby. This commuting scenario can be analysed using an origin-focused GWSI model. Here, we assume that people from the same origin are more likely to have the same characteristics and behave similarly (see Haynes and Fotheringham, 1984), so then flows related to them will have more weight and influence on the calibration using an origin-focused model. Further dissaggregated information can be obtained if a destination-specific origin-focused model is used.

A second scenario on commuting which perhaps is more likely is when workers find a job and then look for a residence close to their work. For instance, Clark and Burt (1980) have studied that when households move, there is a considerable tendency to relocate closer to their working place (see Clark et al., 2003). These types of behaviour can be analysed using a destination-focused model or an origin-specific destination-focused model. Here we assume it is more likely that people who are working in the same locations or nearby destinations have more similar characteristics than people coming from the same locations. Therefore, we would then weight the flows with destination

information rather than origin information.

A frequent problem in spatial interaction modelling is the problem of non-geographic disaggregation of the model. This arises for example in commuting where workers have different qualifications and the companies are active in different economic sectors. In this situation, a different model might be required for each group of commuters. In the case of the business centre discussed above, it might happen that the new centre is located in an area where the active population does not have the required qualification for working in a service-oriented business. In order to model this situation accurately, we need to make sure we account for the qualifications both of the population and the jobs. This problem is similar to the problem of market segmentation, discussed in section 3.4, with one possible model formulation in equation 3.18. We could apply the same principle to the geographically weighted spatial interaction model, for example in the case of the business centre with a destination-focused model:

$$T_{ij}(q) = \kappa_{\{j\}}(q)\ P_i^{\alpha_{\{j\}}}(q)\ N_j^{\gamma_{\{j\}}}(q)\ d_{ij}^{\ \beta_{\{i\}}}(q) \tag{8.7}$$

where q is a specific qualification needed for the business centre and for which we want to estimate the commuting flows. All other variables are the same as in equation 6.7. Of course, we would need to know the commuting flow for the group having the required qualifications towards other existing destinations in order to successfully calibrate the model.

Another possibility is the use of the mixed kernel approach presented in section 8.2. Instead of using the strength of connection for the non-geographical kernel, information about the belonging to a specific group (e.g. having the required qualifications for working in the new business centre) could be included. This approach might be especially useful if the membership to a group is not very well defined. For example, the population with similar qualifications could also be included in the model and given a weight according to the probability of having an employee with those qualifications in the new business centre. In other applications, this approach could also be useful in cases where segmentation is in arbitrary groups, such as age groups. For these continuous variables, a finer weighting scheme might lead to better results. One problem with further disaggregation of the model is also the risk of overfitting. In geographically weighted models, overfitting is addressed by using a sufficiently big kernel bandwidth, allowing for drawing in a sufficient amount of information from neighbouring locations. In the case of non-geographical disaggregation, borrowing information from similar groups might be necessary if data availability is sparse.

In the examples above, the dataset is limited to 70 origins and 1 (section 8.5.1) or 70 destinations (section 8.5.2), resulting in 4900 flows at maximum. This size of a dataset for a spatial interaction system is rather small, but even so it required considerable computational resources, especially for bandwidth selection where many different models are computed and compared to each other. Strategies to deal with bigger datasets have

to be developed. One attempt to deal with big datasets in local regression models is the Scalable Local Regression (SLR) approach which is an incremental variant of GWR developed by Kaiser et al. (2011). In this study, (i.e. Kaiser et al. (2011)) SLR in a GWSI example is applied to roughly 150,000 journey-to-work flows across the whole of Switzerland. However, only Gaussian regression has been considered in SLR and Poisson regression has not been implemented in this model, limiting its usability for GWSI models.

In the following chapter we summarise the major findings of the thesis with regard to the research questions posed initially and will provide a conclusion of the study along with some suggestions for future research.

Chapter 9

Conclusion

The main research objectives in this thesis were how spatial heterogeneity can be detected in spatial interaction processes and how spatial interaction models can be localised in order to take spatial heterogeneity into account. These objectives were addressed by creating a *family of local spatial interaction models* taking into account spatial heterogeneity through geographical weighting of the flows.
This family is composed by the following models:

- **Origin-focused spatial interaction model.** In this model the flows are weighted according to the geographical distance between the calibration point and the origin of the flows. The origin-specific model can be seen as a special case of the origin-focused model with a bi-square kernel and bandwidth 0, except for the choice of the variables.

- **Destination-focused spatial interaction model.** This model is similar to the origin-focused model, except that the flows are weighted according to the geographical distance between the calibration point and the *destination* of the flows. Again, the destination-specific model can seen as a special case of the destination-focused model, except for the choice of the model variables.

- **Origin-focused destination-specific spatial interaction model.** In this case, the origin-focused geographically weighted approach is applied to a destination-specific model. Individual flows towards one single location are weighted according to the distance between their origin to the calibration point.

- **Destination-focused origin-specific spatial interaction model.** In this case, the destination-focused approach is applied to a origin-specific model.

- **Flow-focused spatial interaction model.** Individual flows are weighted based on the distance between the flows. Several measures can be used for the distance between the flows, such as the 4-dimensional Euclidean distance or a spatial trajectories distance.

These different models are built on top of the GWR principle in which the geographical weighting approach allows finding local parameters at a given calibration point and takes into account possible spatial heterogeneity through localised parameter estimates. The spatial variation can refer to origins or destinations only, or to both simultaneously; for each spatial heterogeneity case a different model can be chosen from the GWSI family. All models are also able to adapt to situations with or without spatial variation through the bandwidth calibration process. If no spatial heterogeneity is found in the dataset, the kernel bandwidth will be large. Indeed, a global unconstrained spatial interaction model can be seen as a special case of a local model where the kernel bandwidth is infinity and all flows are given weight 1.

This family of local spatial interaction models has been gradually introduced and studied in successive chapters of this thesis and related aspects and issues have been highlighted, discussed, and in some cases solutions suggested. More specifically, the contributions of the individual chapters are as follows:

- Chapter 2 describes the journey-to-work dataset of the agglomeration of Lausanne which has been used throughout the thesis for illustrating and testing the individual models. Additional data required for spatial interaction modelling such as active population and number of jobs is also described.

- Chapter 3 briefly reviews existing spatial interaction models and their underlying theoretical frameworks. The problem of the intra-zonal flows which are commonly ignored is highlighted.

- Chapter 4 introduces some novel considerations on how to deal with the intra-zonal flows issue in spatial interaction models by estimating the average trip length within a zone. The best results are given by the population density-based scattering approach which estimates the average trip length based on the population distribution within the zone. As a result, intra-zonal flows can also be considered in the spatial interaction models. In the case of the Lausanne journey-to-work dataset, integrating the intra-zonal flows is important since more than 45% of all the flows are internal flows (see Kordi et al., 2012).

- Chapter 5 introduces the principle of local models and more specifically the principle of GWR. It also discusses existing methods for localising spatial interaction models, such as origin- or destination-specific models. Finally, chapter 5 also presents some attempts to combine the GWR principle with spatial interaction models.

- The geographical weighting principle of spatial interaction models is introduced in chapter 6 and the origin- and destination-focused models are developed and discussed. Chapter 6 also introduces the combination of the geographical weighting of flows with the origin- and destination-specific models, resulting in the origin-focused

destination-specific and destination-focused origin-specific models (see Kordi and Fotheringham, 2011).

- Chapter 7 introduces the last member of the family of local GWSI models, the flow-focused model. First, different approaches for measuring the distance between two flows are discussed. One approach introduced in this chapter is the four-dimensional kernel approach, where the flow is considered as a four-dimensional object and the distance between the flows is calculated by the Euclidean distance in four-dimensional space. Different variants to the Euclidean distance, such as Manhattan or Minkowski distance are also discussed. A second approach, the spatial trajectories distance measure, is also introduced, based on the sum-of-pairs distance by Agrawal et al. (1993). The model has been applied to the Lausanne journey-to-work dataset, giving the opportunity to discuss bandwidth calibration issues for the local spatial interaction models. Visualisations for the parameter estimates for each flow have also been developed using matrix-style visualisations and a series of flow maps.

- Chapter 8 discusses some issues arising from the geographical weighting of spatial interaction models that have not been treated in depth in previous chapters. This includes the replacement of Euclidean distance measures by network distance and travel time distance, which in the case of the Lausanne journey-to-work dataset does not improve the model estimates. Chapter 8 also introduces a mixed-kernel approach, where a second kernel is added to the geographical weighting scheme in order to take into account the strength of connection between destinations in a destination-focused model. The two kernels are combined together using a weighted average. The two bandwidths and the weight for the geographical kernel can be optimised using the AICc score; in the case of the Lausanne dataset, the geographical kernel is weighted much more than the kernel for the strength of connection. The mixed kernel approach is flexible in the sense that it allows for excluding one of the kernels.

  Chapter 8 also addresses the issue of having all outgoing flows of one region located at the same origin coordinates ("stacked" origins), and similarly for destinations, by scattering the origins and destinations. This scattering approach leads to a more robust spatial interaction model, at least for the studied dataset. In previous chapters, only a fixed bandwidth has been considered. The principle of geographical weighting of spatial flows does not exclude an adaptive bandwidth commonly used in GWR. In chapter 8, an example of an adaptive bandwidth is shown, with similar results to the fixed bandwidth approach.

  Chapter 8 also contains two locational analysis examples where variants of local spatial interaction models have been tested. The first is to study the flows towards a planned business centre where 1000 jobs are to be created, by using a

destination-focused model. The second example predicts commuting flows for the agglomeration of Lausanne for the year 2010, based on flow data for 2000. The resulting estimates shows an increase in the journey-to-work flows which seems to be consistent with the overall increase of active population and the number of jobs. This flow prediction has been done using a flow-focused model.

The empirical experiments in this thesis indicate that the local GWSI models might be an improvement over conventional global spatial interaction models since they provide more useful spatial information on interaction behaviour. These models also can act as a generalisation to global and previous local spatial interaction models such as origin- and destination-specific models. The family of local spatial interaction models represents a feasible solution to detect and take into account spatial heterogeneity in spatial interaction processes.

Beyond these two main research questions, other minor questions have also been addressed. Two different approaches have been developed on how to define distances between flows. Considering flows as four-dimensional directed objects allows the opportunity to consider them using similar techniques as for two-dimensional spatial analysis. Also, the problem of visualising local parameter estimates for spatial interaction models could be addressed using traditional maps for origin- and destination-focused models and matrix-style visualisations for the flow-focused models; extracts of the matrix visualisations also have been used to develop a series of flow maps.

One of the minor questions that needed to be answered was if existing GWR software can be used also for GWSI models. For all models in the family of local spatial interaction models except for the flow-focused model, existing GWR can potentially be used, as long as Poisson GWR is implemented. The flow-focused model needs a special weighting scheme, which is either a four-dimensional distance or a spatial trajectories distance. These distance calculations require a specific implementation of GWR. The mixed kernel approach also requires implementing a specific weighting scheme. In practice, existing GWR software tested with GWSI models presented some problems. In most cases, the computation was very slow, taking several hours for the Lausanne journey-to-work dataset with 4900 flows. In many cases, errors occurred during the calculation for unknown reasons. Consequently, a Python implementation of GWR has been built on top of the statsmodels Python package[1], which uses the computationally efficient Numpy library[2] (especially for matrix calculations). Statsmodels provides the required regression models (especially weighted Poisson regression), and only the geographical weighting scheme had to be added. This custom GWR implementation has been validated against existing GWR software (GWR3[3] and GWR4[4]).

[1] `http://statsmodels.sourceforge.net`
[2] `http://www.numpy.org`
[3] `http://ncg.nuim.ie/ncg/gwr/software.htm`
[4] `http://www.st-andrews.ac.uk/geoinformatics/gwr/gwr-software/`

Another minor research question that has been raised in the introduction was how intra-zonal flows can be taken into account in spatial interaction modelling. A novel approach has been developed in chapter 4 for estimating average trip length in which origins and destinations of flows are distributed randomly or based on an available density surface within their respective zones. The random and population-based scattering methods have shown good results in the case of the Lausanne journey-to-work dataset.

9.1 Future research

The work presented in this thesis opens several new interesting avenues for future research. First, although the GWSI models in this thesis are applied and analysed in the context of commuting interaction, they could be applied equally to other types of spatial interaction data. So, further work on the application of different variants of GWSI models to other interaction datasets, for instance on bigger datasets from different regions on migration data, flow of information (phone calls, e-mail traffic etc.) or shopping behaviour, would be interesting. Also, testing the GWSI approach with other types of regression, such as Poisson-Gamma or negative binomial regression would be another pertinent issue to be addressed.

Furthermore, so far all geographical weighted spatial interaction models are extensions of an unconstrained gravity model. In traditional spatial interaction modelling, constrained models play an important role; for instance both of the example applications presented in section 8.5 would traditionally be studied with constrained interaction models. Geographically weighting of a constrained model raises interesting issues. Future research could address the relevance of such an approach.

In section 8.2, the mixed kernel approach has been introduced and applied to the Lausanne dataset. This approach integrates a simple measure for the strength of connection between destinations. The method can be tested on other datasets and also other measures for the strength of connection could be examined. For example, using a clustering approach that tries to group together similar destinations based on some features that need to be defined might be an alternative. Applications of the mixed kernel in other models could also be tested.

In order to ease the application of the GWSI models and establish the method, a reliable software package should be built with all the different algorithms. A well-documented Python package with some example applications would enable many researchers to use the models with their own data and integration into other user-friendly software packages such as Quantum GIS[5] would be possible. Also, computationally efficient variants such as the Scalable Local Regression presented in Kaiser et al. (2011) should be further extended to better take into account the specificities of spatial interaction models, such as using Poisson or negative binomial regression instead of simple

[5] http://qgis.org

Gaussian regression. Such an algorithm would open the application of the GWSI models to huge datasets.

Bibliography

Agnew, J. (1996). Mapping politics: how context counts in electoral geography. *Political geography*, 15:129–146.

Agrawal, R., Faloutsos, C., and Swami, A. (1993). Efficient similarity search in sequence databases. In Lomet, D., editor, *Foundations of Data Organization and Algorithms*, volume 730 of *Lecture Notes in Computer Science*, pages 69–84. Springer, Berlin.

Agresti, A. (1990). *Categorical data analysis*. NY: Wiley, New York.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. and Csáki, F., editors, *2nd International Symposium on Information Theory*, pages 267—281. Akadémiai Kiadó.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Alonso, W. (1978). A theory of movement. In Hansen, N. M., editor, *Human Settlement Systems*, pages 197–211. Ballinger, Cambridge.

Anselin, L. (1988). *Spatial Econometrics, Methods and Models*. Dordrecht: Kluwer Academic.

Anselin, L. (1995). Local Indicators of Spatial Association–LISA. *Geographical Analysis*, 27(2):93–115.

Anselin, L. (1999). The future of spatial analysis in the social sciences. *Geographical Information Sciences*, 5:67–76.

Anselin, L. (2001a). Rao's score test in spatial econometrics. *Journal of Statistical Planning and Inference*, 97:113–139.

Anselin, L. (2001b). Spatial econometrics. In Baltagi, B. H., editor, *A Companion to Theoretical Econometrics*. Wiley-Blackwell, Oxford.

Anselin, L. (2009). Spatial regression. In Fotheringham, A. S. and Rogerson, P. A., editors, *The SAGE Handbook of Spatial Analysis*, pages 254–275. SAGE Publications Ltd, London.

Anselin, L. (2010). Thirty years of spatial econometrics. *Papers in Regional Science*, 89(1):3—25.

Anselin, L. and Getis, A. (1992). Spatial statistical analysis and geographic information systems. *Annals of Regional Science*, 26(1):19–33.

Atkinson, P. M. (2001). Geographical information science: geocomputation and nonstationarity. *Progress in Physical Geography*, 25:111–122.

Bailey, T. C. and Gatrell, A. C. (1995). *Interactive Spatial Data Analyses*. Longman, Essex.

Balcan, D., Colizza, V., Gonçalves, B., Hu, H., and JJ, R. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences of the United States of America*, 106(51):21484–21489.

Ballas, D. and Clarke, G. (2001). Towards local implications of major job transformations in the city: a spatial microsimulation approach. *Geographical Analysis*, 31:291–311.

Banerjee, S., Gelfand, A. E., and Polasek, W. (2000). Geostatistical modelling for spatial interaction data with application to postal service performance. *Journal of Statistical Planning and Inference*, 90:87–105.

Batten, D. F. and Boyce, D. E. (1986). Spatial interaction, transportation, and interregional commodity flow models. In Nijkamp, P., editor, *Handbook of Regional and Urban Economics*, volume 1 of *Regional Economics*, pages 357–406. North-Holland, Amsterdam.

Batty, M. (1974). Spatial entropy. *Geographical Analysis*, 6(1):1–31.

Batty, M. (1976). *Urban Modeling: Algorithms, Calibrations, Predictions*. Cambridge University Press, London.

Bavaud, F. (2008). Local concentrations. *Papers in Regional Science*, 87(3):357–370.

Bavaud, F. (2010a). Euclidean distances, soft and spectral clustering on weighted graphs euclidean distances, soft and spectral clustering on weighted graphs euclidean distances, soft and spectral clustering on weighted graphs. In *Proceedings of the ECML PKDD'10. Lecture Notes in Computer Science 6321*, pages 103–118. Springer.

Bavaud, F. (2010b). Models for Spatial Weights: A Systematic Look. *Geographical Analysis*, 30(2):153–171.

Behrens, K., Ertur, C., and Koch, W. (2010). Dual ' gravity : Using spatial econometrics to control for multilateral resistance. *Journal of Applied Econometrics*, 26.

Bentley, G., Cromley, R., and Atkinson-Palombo, C. (2013). The network interpolation of population for flow modeling using dasymetric mapping. *Geographical Analysis*, 45(3):307–323.

Berglund, K. and Karlström, A. (1999). Identifying local spatial association in flow data. *Journal of Geographical Systems*, 1:219–236.

Berry, B. J. L., Goheen, P., and Goldstein, H. (1969). Metropolitan area definition: a reevaluation of concept and statistical practice. Working paper, United States Bureau of Census, Washington DC.

Bertin, J. (1983). *Semiology of graphics: diagrams, networks, maps*. The University of Wisconsin Press, Madison, Wisconsin. Translated from French by Berg, W.

Bettman, J. (1979). Memory factors in consumer choice: a review. *Journal of Marketing*, 43(2):37–53.

Bharat, P. and Larsen, O. (2011). Are intrazonal trips ignorable? *Transport Policy*, 18:13–22.

Birkin, M., Clarke, G., and Clarke, M. (2002). *Retail geography and intelligent network planning*. Wiley, Chichester.

Birkin, M., Clarke, G., and Clarke, M. (2010). Refining and operationalizing entropy-maximizing models for business applications. *Geographical Analysis*, 42(4):422–445.

Birkin, M. and Clarke, M. (1985). Comprehensive dynamic urban models: integrating macro- and microapproaches. In Griffith, D. and Haining, R., editors, *Transformations through space and time: an analysis of nonlinear structures, bifurcation points and autoregressive dependencies*, pages 165–292. Martinus Nijhoff, Dordrecht.

Birkin, M. and Clarke, M. (1988). SYNTHESIS - a synthetic spatial information system for urban and regional analysis: methods and examples. *Environment and Planning A*, 20:1645–1671.

Bitter, G., Mulligan, G. F., and Dall'erba, S. (2007). Incorporating spatial variation in housing attribute prices: a comparison of geographically weighted regression and the spatial expansion method. *Journal of Geographical Systems*, 9(1):7–27.

Black, W. R. (1992). Network autocorrelation in transport network and flow systems. *Geographical Analysis*, 24(3):207–222.

Blondel, V., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 10.

Boots, B. and Okabe, A. (2007). Local statistical spatial analysis: Inventory and prospect. *Journal of Geographical Information Science*, 21(4):355—375.

Bowman, A. W. (1984). An Alternative Method of Cross-Validation for the Smoothing of Density Estimates. *Biometrika*, 71:353—360.

Boyandin, I., Bertini, E., Bak, P., and Lalanne, D. (2011). Flowstrates: An approach for visual exploration of temporal origin-destination data. *Computer Graphics Forum*, 30(3):971–980.

Brown, L. A. and Kodras, J. E. (1987). Migration human resources transfer and development contexts: a logit analysis of Venezuelan data. *Geographical Analysis*, 19:243–263.

Brown, W. M. and Andreson, W. P. (2002). Spatial markets and the potential for economic integration between Canadian and US regions. *Papers in Regional Science*, 81(1):99–120.

Brunsdon, C., Aitkin, M., Fotheringham, A. S., and Charlton, M. (1999a). A Comparison of Random Coefficient Modelling and Geographically Weighted Regression for Spatially Non-Stationary Regression Problems. *Geographical and Environmental Modelling*, 3:47–62.

Brunsdon, C., Fotheringham, A. S., and Charlton, M. (1996). Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geographical Analysis*, 28:281–289.

Brunsdon, C., Fotheringham, A. S., and Charlton, M. (1997). Geographical instability in linear regression modelling – a preliminary investigation. In *New techniques and technologies for statistics II*, pages 149–158. IOS Press, Amsterdam.

Brunsdon, C., Fotheringham, A. S., and Charlton, M. (1998a). Geographically weighted regression - modelling spatial non-stationarity. *The Statistician*, 47:431–443.

Brunsdon, C., Fotheringham, A. S., and Charlton, M. (1998b). Spatial nonstationarity and autoregressive models. *Environment and Planning A*, 30:957–973.

Brunsdon, C., Fotheringham, A. S., and Charlton, M. (1999b). Some notes on parametric significance tests for geographically weighted regression. *Journal of Regional Science*, 39(3):497–524.

Brunsdon, C., McClatchey, J., and Unwin, D. J. (2001). Spatial variations in the average rainfall – altitude relationship in Great Britain: an approach using geographically weighted regression. *International Journal of Climatology*, 21:455–466.

Burger, M., van Oort, F., and Linders, G. (2009). On the Specification of the Gravity Model of Trade: Zeros, Excess Zeros and Zero-inflated Estimation. *Spatial Economic Analysis*, 4(2):167–190.

Burnham, K. and Anderson, D. R. (2002). *Model Selection and Multimodal Inference: A Practical Information-Theoretic Approach*. Springer, New York, 2nd edition.

Cahill, M. and Mulligan, G. (2007). Using geographically weighted regression to explore local crime patterns. *Social Science Computer Review*, 25(2):147–193.

Cameron, A. C. and Trivedi, P. K. (1998). *Regression analysis of count data*. Cambridge, New York.

Carey, H. C. (1858). *Principles of social science*, volume 1. Lippincott, Philadelphia.

Casetti, E. (1972). Generating Models by the Expansion Method: Applications to Geographical Research. *Geographical Analysis*, 4(1):81–91.

Casetti, E. (1997). The Expansion Method, Mathematical Modeling, and Spatial Econometrics. *International Regional Science Review*, 20:9–32.

Charlton, M., S, F. A., and Brunsdon, C. (1997). The Geography of Relationships: An Investigation of Spatial Non-Stationarity. In Bocquet-Appel, J., Courgeau, D., and Pumain, D., editors, *Spatial Analysis of Biodemographic Data*, pages 23–47. John Libbey Eurotext.

Chen, Y. (2009). Spatial interaction creates period-doubling bifurcation and chaos of urbanization. *Chaos, Solitons and Fractals*, 42:1316–1325.

Cheng, E. M., Atkinson, P. M., and Shahani, A. K. (2011). Elucidating the spatially varying relation between cervical cancer and socio-economic conditions in england. *International Journal of Health Geographics*, 10(51).

Chun, Y. (2008). Modeling network autocorrelation within migration flows by eigenvector spatial filtering. *Journal of Geographical Systems*, 10:317–344.

Chun, Y. and Griffith, D. A. (2011). Modeling network autocorrelation in space–time migration flow data: An eigenvector spatial filtering approach. *Annals of the Association of American Geographers*, 101(3):523–536.

Chun, Y., Kim, H., and Kim, C. (2012). Modeling interregional commodity flows with incorporating network autocorrelation in spatial interaction models: An application of the us interstate commodity flows. *Computers, Environment and Urban Systems*.

Clark, W. and Burt, J. (1980). The impact of workplace on residential relocation. *Annals of the Association of American Geographers*, 79:59–67.

Clark, W., Huang, Y., and Withers, S. (2003). Does commuting distance matter? commuting tolerance and residential change. *Regional Science and Urban Economics*, 33:199–221.

Clarke, G. and Clarke, M. (2001). Applied spatial interaction modelling. In Clarke, G. and Madden, M., editors, *Regional Science in Business*, pages 137–157. Springer.

Clarke, G., Eyre, H., and Guy, C. (2002). Deriving indicators of access to food retail provision in British cities: studies of Cardiff, Leeds and Bradford. *Urban Studies*, 39(11):2041–2060.

Clarke, G., Langley, R., and Cardwell, W. (1998). Empirical applications of dynamic spatial interaction models. *Computers, Environment and Urban Systems*, 22:157–184.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836.

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

Cox, K. R. (1969). The voting decision in a spatial context. *Progress in geography*, 1:81–117.

Csáji, B., Browet, A., Traag, V., Delvenne, J.-C., Huens, E., Van Dooren, P., Smoreda, Z., and Blondel, V. (2013). Exploring the mobility of mobile phone users. *Physica A: Statistical Mechanics and its Applications*, 392(6):1459–1473.

Curry, L. (1972). A Spatial Analysis of Gravity Flows. *Regional Studies*, 9:285–288.

Curry, L., Griffith, D. A., and Sheppard, E. S. (1975). Those Gravity Parameters Again. *Regional Studies*, 9:289–296.

Dahmann, D. C. and Fitzsimmons, J. D. (1995). Metropolitan and nonmetropolitan areas: new approaches to geographical definition. Working Paper 12, United States Bureau of Census, Washington DC.

Dalziel, B., Pourbohloul, B., and Ellner, S. (2013). Human mobility patterns predict divergent epidemic dynamics among cities. *Proceedings of the Royal Society B: Biological Sciences*, 280(1766).

de Vries, J. J., Nijkamp, P., and Rietveld, P. (2009). Exponential or power distance-decay for commuting? an alternative specification. *Environment and Planning A*, 41(2):461–480.

Dendrinos, D. S. and Sonis, M. (1990). *Chaos and socio-spatial dynamics*. Springer, New York.

Dessemontet, P. (2011). *Changes in Employment Localization and Accessibility: the Case of Switzerland between 1939 and 2008*. EPFL, Lausanne.

Dessemontet, P., Kaufmann, V., and Jemelin, C. (2010). Switzerland as a single metropolitan area? a study of its commuting network. *Urban Studies*, 47(13):2785—2802.

Ding, Y. and Fotheringham, A. S. (1992). The Integration of Spatial Analysis and GIS. *Computers, Environment and Urban Systems*, 16:3–19.

Dodd, S. C. (1950). The interactance hypothesis: A model fitting physical masses and human groups. *American Sociological Review*, 15:245–257.

Dorling, D. (1991). *The Visualization of Spatial Structure*. Phd thesis, Department of Geography, University of Newcastle upon Tyne, Newcastle upon Tyne.

Dueñas, M. and Fagiolo, G. (2013). Modeling the international trade network: a gravity approach. *Journal of Economic Interaction and Coordination*, 8(1):155–178.

Eichengreen, B. and Irwin, D. A. (1998). The role of history in bilateral trade flows. In Frankel, J. A., editor, *The Regionalization of the World Economy*, pages 33–57. University of Chicago Press.

Eilon, S., Watson-Gandy, C., and Christofides, N. (1971). *Distribution management: mathematical modelling and practical analysis*. Griffin, London.

England, J. (2000). *Retail Impact Assessment: A Guide to Best Practice*. Routledge, London.

Expert, P., Evans, T., Blondel, V., and Lambiotte, R. (2011). Uncovering space-independent communities in spatial networks. *Proceedings of the National Academy of Sciences of the United States of America PNAS*, 108(19):7663–7668.

Farber, S. and Páez, A. (2007). A systematic investigation of cross-validation in GWR model estimation: empirical analysis and Monte Carlo simulations. *Journal of Geographical Systems*, 9(4):371–396.

Farmer, C. J. Q. and Fotheringham, A. S. (2011). Network-based functional regions. *Environment and Planning A*, 43(11):2723–2741.

Ferguson, N., Cummings, D., Fraser, C., Cajka, J., Cooley, P., and Burke, D. (2006). Strategies for mitigating an influenza pandemic. *Nature*, 442:448–452.

Fischer, M. M. (2000). Spatial Interaction Models and the Role of Geographic Information Systems. In Fotheringham, A. S. and Wegener, M., editors, *Spatial Models and GIS: New Potential and New Models*, pages 33–43. Taylor and Francis, Philadelphia.

Fischer, M. M. (2002). Learning in neural spatial interaction models: a statistical perspective. *Journal of Geographical Systems*, 4(3):287–299.

Fischer, M. M. and Getis, A. (1999). New advances in spatial interaction theory. *Papers in Regional Science*, 78:117–118.

Fischer, M. M. and Griffith, D. A. (2008). Modeling spatial autocorrelation in spatial interaction data: An application to patent citation data in the European union. *Journal of Regional Science*, 48:969–989.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest number of a sample. *Proceedings of the Cambridge Philosophical Society*, 24:180–190.

Flowerdew, R. (1982). Fitting the lognormal gravity model to heteroscedastic data. *Geographical Analysis*, 14:263–267.

Flowerdew, R. and Aitkin, M. (1982). A method of fitting the gravity model based on the Poisson distribution. *Journal of Regional Science*, 22:191–202.

Flowerdew, R. and Lovett, A. (1988). Fitting Constrained Poisson Regression Models to Interurban Migration Flows. *Geographical Analysis*, 20(4):297–307.

Foody, G. M. (2004). Gis: stressing the geographical. *Progress in Physical Geography*, 28(1):152–158.

Fotheringham, A. S. (1981). Spatial structure and distance-decay parameters. *Annals of the Association of American Geographers*, 71:425–436.

Fotheringham, A. S. (1982a). Distance-Decay Parameters: A Reply. *Annals of the Association of American Geographers*, 72(4):551–553.

Fotheringham, A. S. (1982b). Multicollinearity and parameter estimates in a linear model. *Geographical Analysis*, 14:64–71.

Fotheringham, A. S. (1983). A new set of spatial interaction models: the theory of competing destinations. *Environment and Planning A*, 15(1):15–36.

Fotheringham, A. S. (1984a). Spatial flows and spatial patterns. *Environment and Planning A*, 16(4):529–543.

Fotheringham, A. S. (1984b). Spatial Flows and Spatial Patterns: A Generalization. *Environment and Planning A*, 16(11):1521–1984.

Fotheringham, A. S. (1986). Modelling hierarchical destination choice. *Environment and Planning A*, 18:401–418.

Fotheringham, A. S. (1987). Hierarchical destination choice: Discussion with evidence from migration in The Netherlands. Working Paper 9, Interuniversity Demographic Institute Netherlands, The Hague, Netherlands.

Fotheringham, A. S. (1988a). Consumer store choice and choice-set definition. *Marketing Science*, 7:299–310.

Fotheringham, A. S. (1988b). Market share analysis techniques: a review and illustration of current U.S. practice. In Wrigley, N., editor, *Store Choice, Store Location and Market Analysis*, pages 120–159. Routledge, London.

Fotheringham, A. S. (1991). Migration and spatial structure: the development of the competing destinations model. In Stillwell, J. and Congdon, P., editors, *Migration Models: Macro and Micro Approaches*, pages 57–72. Belhaven, London.

Fotheringham, A. S. (1992). The Integration of Spatial Analysis and GIS. *Computers, Environment and Urban Systems*, 16(1):3–19.

Fotheringham, A. S. (1993). On the Future of Spatial Analysis: The Role of GIS. *Environment and Planning A*, Anniversary lssue:30–34.

Fotheringham, A. S. (1997). Trends in quantitative methods I: stressing the local. *Progress in Human Geography*, 21, 1:88–96.

Fotheringham, A. S. (1999a). GIS-Based Spatial Modelling: A Step Forwards or a Step Backwards? In Fotheringham, A. S. and Wegener, M., editors, *Spatial Models and GIS: New and Potential Models*. Taylor and Francis, London.

Fotheringham, A. S. (1999b). Guest Editorial: Local Modelling. *Geographical and Environmental Modelling*, 3:5–7.

Fotheringham, A. S. (2000). Context-dependent spatial analysis: A role for GIS ? *Journal of Geographical Systems*, 2:71–76.

Fotheringham, A. S. (2001). Spatial interaction models. In Smelser, N. J. and Baltes, P. B., editors, *International Encyclopedia of the Social & Behavioral Sciences*, pages 14794–14800. Elsevier Ltd.

Fotheringham, A. S. (2009). Geographically Weighted Regression. In Fotheringham, A. S. and Rogerson, P. A., editors, *The SAGE Handbook of Spatial Analysis*, page 243. SAGE Publications Ltd, London.

Fotheringham, A. S., Brundson, C., and Charlton, M. (2000). *Quantitative Geography: Perspectives on spatial data analysis*. SAGE, Los Angeles.

Fotheringham, A. S. and Brunsdon, C. (1999). Local Forms of Spatial Analysis. *Geographical Analysis*, 31(4):338–358.

Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2002). *Geographically Weighted Regression: the analysis of spatially varying relationships*. John Wiley, Chichester, Sussex.

Fotheringham, A. S. and Charlton, M. (1994). GIS and Exploratory Spatial Data AnalysIs: An Overview of Some Research Issues. *Geographical Systems*, 1:315–327.

Fotheringham, A. S., Charlton, M., and Brunsdon, C. (1996). The geography of parameter space: an investigation into spatial nonstationarity. *International Journal of Geographical Information Systems*, 10(5):605–627.

Fotheringham, A. S., Charlton, M., and Brunsdon, C. (1997a). Measuring Spatial Variations in Relationships with Geographically Weighted Regression. In Fischer, M. M. and Getis, A., editors, *Recent Developments in Spatial Analysis: Spatial Statistics, Behavioural Modelling and Computational Intelligence*, pages 60–82. Springer-Verlag, Berlin.

Fotheringham, A. S., Charlton, M., and Brunsdon, C. (1997b). Two techniques for exploring non-stationarity in geographical data. *Geographical Systems*, 4:59–82.

Fotheringham, A. S., Charlton, M., and Brunsdon, C. (1998). Geographically weighted regression: a natural evolution of the expansion method for spacial data analysis. *Environment and Planning A*, 30:1905–1927.

Fotheringham, A. S., Charlton, M., and Brunsdon, C. (2001a). Spatial Variations in School Performance: A Local Analysis using Geographically Weighted Regression. *Geographical and Environmental Modelling*, 5(1):43–66.

Fotheringham, A. S. and Dignan, T. (1984). Further contributions to a general theory of movement. *Annals of the Association of American Geographers*, 74(4):620–633.

Fotheringham, A. S., Nakaya, T., Yano, K., Openshaw, S., and Ishikawa, Y. (2001b). Hierarchical destination choice and spatial interaction modelling: A simulation experiment. *Environment and Planning A*, 33:901–920.

Fotheringham, A. S. and O'Kelly, M. E. (1989). *Spatial Interaction Models: Formulations and Applications*. Kluwer Academic., London.

Fotheringham, A. S. and Pitts, T. C. (1995). Directional Variation in Distance-Decay. *Environment and Planning A*, 27:715–729.

Fotheringham, A. S. and Rogerson, P. A. (1993). GIS and Spatial Analytical Problems. *International Journal of Geographic Information Systems*, 7(1):3–19.

Fotheringham, A. S. and Trew, R. (1993). Chain image and store-choice modeling: the e ects of income and race. *Environment and Planning A*, 25(2):179–196.

Fotheringham, A. S. and Webber, M. J. (1980). Spatial Structure and the Parameters of Spatial Interaction Models. *Geographical Analysis*, 12(1):33–46.

Fotheringham, A. S. and Williams, A. (1983). Research notes and comments-further discussion on the Poisson interaction model. *Geographical Analysis*, 15(4):343–347.

Fotheringham, A. S. and Zhan, F. (1996). A Comparison of Three Exploratory Methods for Cluster Detection in Point Patterns. *Geographical Analysis*, 28(3):200–218.

Frankel, J. (1997). *Regional Trading Blocs in the World Economic System*. Institute for International Economics, Washington, DC.

Frick, M. and Axhausen, K. W. (2004). Generating Synthetic Populations using IPF and Monte Carlo Techniques: Some New Results. In *4th Swiss Transport Research Conference, Monte Verità, Ascona, Switzerland*.

Frick, R., Wüthrich, P., Zbinden, R., and Keller, M. (2004). *La pendularité en Suisse*. Recensement fédéral de la population 2000. Relevé structurel de la Suisse. Neuchâtel : Office fédéral de la statistique, edition allemande sous le titre: pendlermobilität in der schweiz edition.

Friendly, M. (2009). The history of the cluster heat map. *The American Statistician*, 63(2):179–184.

Geary, R. C. (1954). The Contiguity Ratio and Statistical Mapping. *The Incorporated Statistician*, 5(3):115–145.

Georgescu-Roegen, N. (1971). *The Entropy Law and the Economic Process*. Harvard University Press, Cambridge.

Georgescu-Roegen, N. (1986). The entropy law and the economic process in retrospect. *Eastern Economic Journal*, 12(1):3–25.

Getis, A. and Aldstadt, J. (2004). Constructing the Spatial Weights Matrix Using a Local Statistic. *Geographical Analysis*, 36(2):90–104.

Getis, A. and Griffith, D. A. (2002). Comparative spatial filtering in regression analysis. *Geographical Analysis*, 34:130–140.

Getis, A. and Ord, K. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis*, 24(3):189–206.

Ghareib, A. (1996). Different travel patterns: interzonal, intrazonal, and external trips. *Journal of Transportation Engineering*, 122:67–75.

Ghoniem, M., Fekete, J., and Castagliola, P. (2004). A comparison of the readability of graphs using node-link and matrix-based representations. In *IEEE Symposium on Information Visualization, 2004. INFOVIS 2004*, pages 17–24.

Gitlesen, J. P. and Thorsen, I. (2000). A competing destinations approach to modeling commuting flows: a theoretical interpretation and an empirical application of the model. *Environment and Planning A*, 32:2057–2074.

Gordon, I. R. (1985). Economic explanations of spatial variation in distance deterrence. *Environment and Planning A*, 17(1):59–72.

Greenacre, M. (2008). Measures of distance between samples: non-euclidean.

Greenwald, M. (2006). The relationship between land use and intrazonal trip making behaviors: evidence and implications. *Transportation Research Part D*, 11:432–446.

Greenwood, M. J. and Sweetland, D. (1972). The Determinants of Migration between Standard Metropolitan Statistical Area. *Demography*, 9:665–681.

Greig, D. M. (1980). *Optimisation*. Longman, London.

Griffith, D. A. (2007). Spatial structure and spatial interaction: 25 years later. *The Review of Regional Studies*, 37(1):28–38.

Griffith, D. A. (2009). Modeling spatial autocorrelation in spatial interaction data: Empirical evidence from 2002 germany journey-to-work flows. *Journal of Geographical Systems*, 11:117–140.

Guldmann, J. M. (1999). Competing destinations and intervening opportunities interaction models of inter-city telecommunication flows. *Papers in Regional Science*, 78(2):179–194.

Guo, D. (2007). Visual analytics of spatial interaction patterns for pandemic decision support. *International Journal of Geographical Information Science*, 21(8):859–877.

Guo, D. (2009). Flow mapping and multivariate visualization of large spatial interaction data. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1041–1048.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall, London.

Haynes, K. and Fotheringham, A. (1984). *Gravity and spatial interaction models*. Number 2 in Scientific Geography Series. Sage, London.

Hoaglin, D. C. and Welsch, R. O. Y. E. (1978). The Hat Matrix in Regression and ANOVA. *The American Statistician*, 32(1):17–22.

Holt, B. H. and Lo, C. P. (2008). The geography of mortality in the Atlanta metropolitan area. *Computers, Environment and Urban Systems*, 32:149–164.

Holten, D. and Van Wijk, J. (2009). Force-directed edge bundling for graph visualization. *Computer Graphics Forum*, 28(3):983–990.

Hu, B., Shao, J., and Palta, M. (2006). Pseudo-r2 in logistic regression model. *Statistica Sinica*, 16:847–860.

Huang, B., Wu, B., and Barry, M. (2010). Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. *International Journal of Geographical Information Science*, 24(3):383–401.

Huff, D. L. (1959). *Geographical Aspects of Consumer Behavior*, volume 18, pages 27–37. University of Washington Business Review.

Hurvich, C. M., Simonoff, J. S., and Tsai, C. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society B*, 60(2):271–293.

Hurvich, C. M. and Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307.

Ireland, C. and Kullback, S. (1968). Contingency tables with given marginals. *Biometrica*, 55:179–188.

Ishikawa, Y. (1987). An empirical study of the competing destinations model using Japanese interaction data. *Environment and Planning A*, 19:1359–1373.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review*, 106(4):620–630.

Jones, J. P. and Casetti, E. (1992). *Applications of the expansion method*. Routledge, London.

Joost, s. and Kalbermatten, M. (2010). *MatSAM, Version 2Beta*.

Kaiser, C. (2011). Dynamic visualisation of journey-to-work flows using map morphing. In *European Conference on Quantitative and Theoretical Geography ECTQG'11, Athens, Greece, 2-5 September 2011*. Harokopio Univeristy, Athens.

Kaiser, C., Kordi, M., Fotheringham, A. S., and Pozdnoukhov, A. (2011). Enabling Spatial Dynamics Modelling with Scalable Local Regression. In *International Conference for Free Open Source Software for Geospatial (FOSS4G)*, Denver, Colorado.

Kamo, K., Yanagihara, H., and Satho, K. (2009). Bias-corrected AIC for selecting variables in Poisson regression models. Technical Report 09-04, Hiroshima Statistical Research Group.

Kanevski, M., Foresti, L., Kaiser, C., Pozdnoukhov, A., Timonin, V., and Tuia, D. (2009). Machine learning models for geospatial data. In Bavaud, F. and Mager, C., editors, *Handbook of Theoretical and Quantitative Geography*, number 2 in FGSE Workshop, pages 175–227. FGSE - University of Lausanne - Faculty of geosciences and environment.

Khawaldah, H., Birkin, M., and Clarke, G. (2012). A review of two alternative retail impact assessment techniques: the case of Silverburn in Scotland. *Town Planning Review*, 83(2):233–260.

Killer, V. and Axhausen, K. W. (2010). Mapping overlapping commuting-to-work areas. *Journal of Maps*, 6(1):147–159.

King, G. (1988). Statistical models for political science event counts: bias in conventional procedures and evidence for the exponential Poisson regression model. *American Journal of Political Science*, 32:838–863.

Kolasa, J. and Rollo, C. D. (1991). The heterogeneity of heterogeneity: a glossary. In *Ecological heterogeneity*, pages 1–23. Springer, New York.

Kordi, M. and Fotheringham, A. S. (2011). Origin- and destination-focused local spatial interaction models. In *The 17th European Colloquium on Theoretical and Quantitative Geography (ECQTG11)*, Athens, Greece.

Kordi, M., Kaiser, C., and Fotheringham, A. S. (2012). A possible solution for the centroid-to-centroid and intra-zonal trip length problems. In *15th AGILE International Conference on Geographic Information Science (AGILE'2012)*, Avignon, France.

Krider, R. and Putler, D. (2013). Which birds of a feather flock together? Clustering and avoidance patterns of similar retail outlets. *Geographical Analysis*, 45(2):123–149.

Krings, G., Calabrese, F., Ratti, C., and Blondel, V. (2009). Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment*, 7.

Kulldorf, M. (1997). A spatial scan statistic. *Communications in Statistics: Theory and Methods*, 26:1481–1496.

Kwan, M. P. (1998). Space–time and integral measures of individual accessibility: A comparative analysis using a point-based framework. *Geographical Analysis*, 30(3):191–216.

Kyriakidis, P. (2004). A geostatistical framework for area-to-point spatial interpolation. *Geographical Analysis*, 36(3):259–289.

Lambiotte, R., Blondel, V., De Kerchove, C., Huens, E., Prieur, C., Smoreda, Z., and Van Dooren, P. (2008). Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387(21):5317–5325.

Langford, M., Higgs, G., Radcliffe, J., and White, S. (2008). Urban population distribution models and service accessibility estimation. *Computers, Environment and Urban Systems*, 32(1):66–80.

Lenormand, M., Huet, S., Gargiulo, F., and Deffuant, G. (2012). A universal model of commuting networks. *PLoS ONE*, 7(10):e45985.

LeSage, J. and Fischer, M. (2010). Spatial econometric methods for modeling origin-destination flows. In M.M., F. and A., G., editors, *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*, pages 409–433. Springer Berlin Heidelberg.

LeSage, J. and Pace, R. (2008). Spatial econometric modeling of origin-destination flows. *Journal of regional science*, 48(5):941–967.

LeSage, J. P. (1999). The Theory and Practice of Spatial Econometrics. Department of Economics, University of Toledo.

LeSage, J. P. and Pace, R. K. (2009). *Introduction to Spatial Econometrics*. Taylor and Francis Group.

Leung, Y., Mei, C., and Zhang, W. (2000a). Statistical tests for spatial nonstationarity based on the geographically weighted regression model. *Environment and Planning A*, 32(1):9–32.

Leung, Y., Mei, C., and Zhang, W. (2000b). Testing for spatial autocorrelation among the residuals of the geographically weighted regression. *Environment and Planning A*, 32(5):871–890.

Li, H. and Reynolds, J. F. (1994). A simulation experiment to quantify spatial heterogeneity in categorical maps. *Ecology*, 75(8):2446–2455.

Linders, G. J. M. and de Groot, H. F. L. (2006). Estimation of the gravity equation in the presence of zero flows. Tinbergen Institute Discussion Paper 2006-072/3.

Linneman, H. V. (1966). *An Econometric Study of International Trade Flows*. North-Holland, Amsterdam.

Liu, X., Kyriakidis, P., and Goodchild, M. (2008). Population-density estimation using regression and area-to-point residual kriging. *International Journal of Geographical Information Science*, 22(4):431–447.

Lloyd, C. (2006). *Local Models for Spatial Analysis*. CRC Press.

Lloyd, C. and Shuttleworth, I. (2005). Analysing commuting using local regression techniques: scale, sensitivity, and geographical patterning. *Environment and Planning A*, 37(1):81–103.

Lloyd, C. D. (2011). *Local models for spatial analysis*. CRC Press, Taylor & Francis Group.

Lloyd, C. D., Shuttleworth, I. G., and Catney, G. (2007). Commuting in Northern Ireland:Exploring Spatial Variations through Spatial Interaction Modelling. In *Proceedings of the Geographical Information Science Research UK Conference, NUI Maynooth*.

Lo, L. (1991). Substitutability, spatial structure and spatial interaction. *Geographical Analysis*, 23:132–146.

Loader, C. (1999). *Local Regression and Likelihood*. Springer, New York.

Lovett, A. and Flowerdew, R. (1989). Analysis of count data using Poisson regression. *Professional Geographer*, 41(2):190–198.

Lovett, A. A., Bentham, C. G., and Flowerdew, R. (1986). Analysing geographic variations in mortality using Poisson regression: the example of ischaemic heart disease in England and Wales 1969-1973. *Social Science and Medicine*, 23:935—943.

Mäkinen, E. and Siirtola, H. (2000). Reordering the reorderable matrix as an algorithmic problem. In *Theory and applications of diagrams, First International Conference, Diagrams 2000 Edinburgh, Scotland, UK, September 1–3, 2000 Proceedings*, volume 1889 of *Lecture Notes in Computer Science*, pages 453–468. Springer, Berlin, Heidelberg.

Masucci, A. P., Serras, J., Johansson, A., and Batty, M. (2013). Gravity versus radiation models: on the importance of scale and heterogeneity in commuting flows. *Physical Review E*, 88(2).

McCullagh, P. and Nelder, J. (1989). *Generalized linear models*. Chapman & Hall, London.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In Zarembka, P., editor, *Frontiers in Econometrics*, pages 105–142. Academic Press.

McFadden, D. (1978). Modelling the choice of residential location. In Karlquist, A., Lundquist, L., Snickars, E., and Weibull, I., editors, *Spatial Interaction Theory and Planning Models*, pages 75–96. North-Holland, Amsterdam.

McFadden, D. (1980). Econometric models for probabilistic choice among products. *Journal of Business*, 53:513–529.

Mennis, J. (2009). Dasymetric mapping for estimating population in small areas. *Geography Compass*, 3(2):727–745.

Meyer, R. J. and Eagle, T. C. (1982). Context-induced parameter instability in a disaggregate-stochastic model of store choice. *Journal of Marketing Research*, 19:62–71.

Monmonier, M. (1969). A spatially controlled principal components analysis. *Geographical Analysis*, 1:192–195.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1):17–23.

Müller, K. and Axhausen, K. (2010). *Population synthesis for microsimulation: State of the art*. ETH Zürich, Institut für Verkehrsplanung, Transporttechnik, Strassen-und Eisenbahnbau (IVT).

Nakaya, T. (1995). Spatial interaction modelling using pr-perceptron: a case study of migration in japan. *Human Geography*, 47:521–540.

Nakaya, T. (2001). Local spatial interaction modelling based on the geographically weighted regression approach. *GeoJournal*, 53:347–358.

Nakaya, T. (2003). Local spatial interaction modelling based on the geographically weighted regression approach. In Boots, B., Okabe, A., and Thomas, R., editors, *Modelling geographical systems: statistical and computational applications*, pages 45–68. Dordrecht, Kluwer Academic Publishers.

Nakaya, T. (2007). Geographically Weighted Regression. In Kemp, K., editor, *Encyclopedia of Geographical Information Science*. Sage Publications, Los Angeles.

Nakaya, T., Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2005). Geographically weighted poisson regression for disease association mapping. *Statistics in Medicine*, 24:2695—2717.

Nakaya, T., Fotheringham, A. S., Hanaoka, K., Clarke, C., Ballas, D., and Yano, K. (2007). Combining microsimulation and spatial interaction models for retail location analysis. *Journal of Geographical Systems*, 9:345—369.

Nelder, J. and Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society A*, 135:370–384.

Niedercorn, J. H. and Bechdolt Jr, B. V. (1969). An economic derivation of the 'gravity law' of spatial interaction. *Journal of Regional Science*, 9(2):273–282.

Nissi, E. and Sarra, A. (2011). Detecting Local Variations in Spatial Interaction Models by Means of Geographically Weighted Regression. *Journal of Applied Science*, 11(4):630–638.

Norman, D. A. and Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology*, 7:44–64.

O'Kelly, M. E. (2004). Isard's contributions to spatial interaction modeling. *Journal of Geographical Systems*, 6(1):43–54.

O'Kelly, M. E. (2009). Spatial Interaction Models. In Kitchin, R. and Thrift, N., editors, *International Encyclopedia of Human Geography*, volume 10, pages 365—368. Elsevier, Oxford.

O'Kelly, M. E. (2010). Entropy based spatial interaction models for trip distribution. *Geographical Analysis*, 42(4):472–487.

O'Kelly, M. E. (2012). Models for spatial interaction data: Computation and interpretation of accessibility. *Lecture Notes in Computer Science*, 7334:249–262.

O'Kelly, M. E. and Lee, W. (2005). Disaggregate journey-to-work data: Implications for excess commuting and jobs-housing balance. *Environment and Planning A*, 37(12):2233–2252.

O'Kelly, M. E. and Niedzielski, M. (2007). Computing and calibrating disaggregated spatial interaction models. In *Proceedings of the Geographical Information Science Research UK Conference, NUI Maynooth*.

O'Kelly, M. E. and Niedzielski, M. A. (2008). Efficient spatial interaction: attainable reductions in metropolitan average trip length. *Journal of Transport Geography*, 16:313–323.

O'Kelly, M. E., Niedzielski, M. A., and Gleeson, J. (2012). Spatial interaction models from irish commuting data: variations in trip length by occupation and gender. *Journal of Geographical Systems*, 14:357–387.

Openshaw, S. (1984). *The modifiable areal unit problem*. Number 38 in CATMOG - Concepts and Techniques in Modern Geography. Geo Abstracts University of East Anglia, Norwich.

Openshaw, S. (1993). Exploratory Space-Time-Attribute Pattern Analysers. In Fotheringham, A. S. and Rogerson, P. A., editors, *Spatial Analysis and GIS*. Taylor and Francis.

Openshaw, S., Charlton, M., Wymer, C., and Craft, A. (1987). A mark 1 geographical analysis machine for the automated analysis of point datasets. *International Journal of Geographical Information Systems*, 1(4):335–358.

Openshaw, S. and Craft, A. (1991). Using geographical analysis machines to search for evidence of clusters and clustering in childhood leukaemia and non-hodgkin lymphomas in britain. In Draper, G., editor, *The Geographical Epidemiology of Childhood Leukaemia and Non-Hodgkin tympt oma. in Great Britain*, number 53 in Studies in Medical and Population Subject, pages 109–122. OPCS, HMSO, London.

Openshaw, S. and Taylor, P. (1981). The modifiable areal unit problem. In Wrigley, N. and Bennet, R., editors, *Quantitative geography*, pages 60–69. Routledge, London.

Ord, J. K. and Getis, A. (1995). Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis*, 27(4):286–305.

Ord, J. K. and Getis, A. (2001). Testing for local spatial autocorrelation in the presence of global autocorrelation. *Journal of Regional Science*, 41:411—432.

Páez, A. and Wheeler, D. C. (2009). Geographically Weighted Regression. In Kitchin, R. and Thrift, N., editors, *International Encyclopedia of Human Geography*, volume 1, pages 407–414. Elsevier, Oxford.

Pagliara, F. and Timmermans, H. J. P. (2009). Choice set generation in spatial contexts-a retrospective review. In Bavaud, F. and Mager, C., editors, *Handbook of Theoretical and Quantitative Geography*, number 2 in FGSE Workshop, pages 311–334. FGSE - University of Lausanne - Faculty of geosciences and environment.

Pattie, C. and Johnston, R. (2000). People who talk together vote together: an exploration of contextual effects in Great Britain. *Annals of the association of American Geographers*, 90:41–66.

Pellegrini, P. A. and Fotheringham, A. S. (1999). Intermetropolitan migration and hierarchical destination choice: a disaggregate analysis from the US Public Use Microdata Samples. *Environment and Planning A*, 31:1093–1118.

Pellegrini, P. A. and Fotheringham, A. S. (2002). Modelling spatial choice: a review and synthesis in a migration context. *Progress in Human Geography*, 26:487—510.

Pellegrini, P. A., Fotheringham, A. S., and Lin, G. (1997). An empirical evaluation of parameter sensitivity to choice set de nition in shopping destination choice models. *Papers in Regional Science*, 76(2):257–284.

Phan, D., Xiao, L., Yeh, R., Hanrahan, P., and Winograd, T. (2005). Flow map layout. In *IEEE Information Visualization (InfoVis)*, pages 219–224.

Pozdnoukhov, A. and Kaiser, C. (2011). Area-to-point kernel regression on streaming data. In *2nd ACM SIGSPATIAL International Workshop on GeoStreaming (IWGS)*. Association for Computing Machinery ACM, Chicago.

Rae, A. (2009). From spatial interaction data to spatial interaction information? Geovisualisation and spatial structures of migration from the 2001 UK census. *Computers, Environment and Urban Systems*, 33(3):161–178.

Rauch, J. E. (1999). Networks versus markets in international trade. *Journal of International Economics*, 48:7–35.

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

Rogerson, P. A. (1999). The detection of clusters using a spatial version of the chi-square goodness-of-fit test. *Geographical Analysis*, 31:130–147.

Rosenberg, M. (2000). The bearing correlogram: a new method of analyzing directional spatial autocorrelation. *Geographical Analysis*, 32:267–278.

Roy, J. R. and Thill, J. C. (2004). Spatial interaction modelling. *Regional science*, 83:339–361.

Roy, R. R. (2004). *Spatial interaction modelling: a regional science context*. Springer, Berlin.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, 9(2):65–78.

Ruzzenenti, F., Picciolo, F., Basosi, R., and Garlaschelli, D. (2012). Spatial effects in real networks: measures, null models, and applications. *Physical Review E*, 86(6).

Sakamoto, Y., Ishiguro, M., and Kitagawa, G. (1986). *Akaike Information criterion statistics*. Mathematics and its applications. KTK Scientific Publishers, Tokyo.

Sang, S., O'Kelly, M. E., and Kwan, M. (2011). Examining commuting patterns: Results from a journey-to-work model disaggregated by gender and occupation. *Urban Studies*, 48(5):891–909.

Schmitt, R. and Greene, D. (1978). An alternative derivation of the intervening opportunities model. *Geographical Analysis*, 10(1):73–77.

Schuler, M., Dessemontet, P., and Joye, D. (2005). *Die Raumgliederungen der Schweiz*. Bundesamt für Statistik (BFS), Neuchâtel.

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464.

Sen, A. and Smith, T. E. (1995). *Gravity Models of Spatial Interaction Behavior*. Advances in Spatial and Network Economics. Springer-Verlag, Berlin.

Sen, A. and Sööt, S. (1981). Selected procedures for calibrating the generalized gravity model. *Papers in Regional Science*, 48(1):165–176.

Shannon, C. F. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379—423 and 623—656.

Sheppard, E. S. (1978). Theoretical underpinnings of the gravity hypothesis. *Geographical Analysis*, 10:386–402.

Sheppard, E. S. (1979). Gravity Parameter Estimation. *Geographical Analysis*, 11:120–132.

Shuttleworth, I. and Lloyd, C. (2005). Analysing Average Travel-to-Work Distances in Northern Ireland Using the 1991 Census of Population: The Effects of Locality, Social Composition, and Religion. *Regional Studies*, 39(7):909–921.

Simini, F., Gonzales, M., Maritan, A., and Barabasi, A.-L. (2012). A universal model for mobility and migration patterns. *Nature*, 484:96–100.

Simini, F., Maritan, A., and Néda, Z. (2013). Human mobility in a continuum approach. *PLoS ONE*, 8(3):e60069.

Singleton, A. D., Wilson, A. G., and O'Brien, O. (2010). Geodemographics and spatial interaction: an integrated model for higher education. *Journal of Geographical Systems*, 14:223–241.

Slingsby, A., Kelly, M., Dykes, J., and Wood, J. (2012). OD maps for studying historical internal migration in Ireland. In *IEEE Conference on Information Visualization (InfoVis), 14 - 19 Oct 2012, Seattle, Washington, US*.

Song, C., Qu, Z., Blumm, N., and Barabasi, A.-L. (2010). Limits of predictability in human mobility. *Science*, 327:1018–1021.

Stewart, J. Q. (1941). An inverse distance variation for certain social influences. *Science*, 93:89–90.

Stouffer, S. (1940). Intervening opportunities: a theory relating mobility and distance. *American Sociological Review*, 5(6):845–867.

Sugiura, N. (1978). Further analysis of the data by akaike's information criterion and the finite corrections. *Communications in Statistics, theory and methods*, A7:13–26.

Thill, J. C. and Kim, M. (2005). Trip making, induced travel demand, and accessibility. *Journal of Geographical Systems*, 7(2):229–248.

Thompson, D. (1974). Review article: Spatial interaction data. *Annals of the Association of American Geographers*, 64(4):560–575.

Thorsen, I. and Gitlesen, J. P. (1998). Empirical evaluation of alternative model specifications to predict commuting flows. *Journal of Regional Science*, 38(2):273—292.

Thrift, N. J. (1983). On the determination of social action in space and time. *Environment and Planning D*, 1:23–57.

Tobler, W. (1976). Spatial interaction patterns. *Journal of Environmental Systems*, 6:271–301.

Tobler, W. (1979). Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association*, 74(367):519–530.

Tobler, W. (1981). A model of geographical movement. *Geographical Analysis*, 13(1):1–20.

Tobler, W. (1987). Experiments in migration mapping by computer. *The American Cartographer*, 14(2):155–163.

Tuia, D. and Bavaud, F. (2007). Revealing distances hidden in flows: a formalism for mapping pendular geostatistical distances. In *Spatial Econometrics Conference*, Cambridge, UK. University of Cambridge.

Uboe, J. (2004). Aggregation of gravity models for journeys to work. *Environment and Planning A*, 36:715–729.

Unwin, A. (1996a). Exploratory spatial analysis and local statistics. *Computational Statistics*, 11:387—400.

Unwin, A. and Unwin, D. (1998). Exploratory spatial data analysis with local statistics. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47:415—421.

Unwin, D. (1996b). GIS, spatial analysis and spatial statistics. *Progress in Human Geography*, 20:540—541.

U.S. Department of Commerce (1965). *Calibrating and treating a gravity model for any size urban area*. U.S. Government Printing Office, Washington, D.C.

Van-Lierop, W. (1986). *Spatial interaction modelling and residential choice analysis*. Gower publishing company, Hants England.

Venigalla, M., Chatterjee, A., and Bronzini, M. (1999). A specialized equilibrium assignment algorithm for air quality modeling. *Transportation Research Part D: Transport and Environment*, 4(1):29–44.

Webb, A. R. (2002). *Statistical Pattern Recognition*. John Wiley and Sons Ltd, second edition.

Wheeler, D. C. and Páez, A. (2010). Geographically Weighted Regression. In Fischer, M. M. and Getis, A., editors, *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*, pages 461–486. Springer, Berlin.

Wheeler, D. C. and Waller, L. A. (2009). Comparing spatially varying coefficient models: a case study examining violent crime rates and their relationships to alcohol outlets and illegal drug arrests. *Journal of Geographical Systems*, 11:1–22.

Wilkinson, L. (1979). Permuting a matrix to a simple pattern. In *Proceedings of the Statistical and Computing Section of the American Statistical Association*, pages 409–412.

Wilson, A. (2010). Entropy in urban and region modelling: retrospect and prospect. *Geographical Analysis*, 42(4):364—394.

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

Wilson, A. G. (1968). *The use of entropy maximising models in the theory of trip distribution, mode split and route split*. Number 1 in Working papers (Centre for Environmental Studies(Great Britain)). Centre for Environmental Studies, London.

Wilson, A. G. (1970). *Entropy in urban and regional modelling*. Pion Ltd, London, UK.

Wilson, A. G. (1971). A family of spatial interaction models, and associated developments. *Environment and Planning A*, 3(1):1–32.

Wilson, A. G. (1974). *Urban and Regional Models in Geography and Planning*. Wiley, London.

Wilson, A. G. (1975). Some New Forms of Spatial Interaction Models: A Review. *Transportation Research*, 9:167–179.

Wilson, A. G. (2000). The widening access debate: student flows to universities and associated performance indicators. *Environment and Planning A*, 32:2019–2031.

Wong, PC adn Foote, H., Mackey, P., Perrine, K., and G, C. (2006). Generating graphs for visual analytics through interactive sketching. *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1386–1398.

Wood, J., Dykes, J., and Slingsby, A. (2010). Visualisation of origins, destinations and flows with OD maps. *The Cartographic Journal*, 47(2):117–129.

Wood, J., Dykes, J., Slingsby, A., and Radburn, R. (2009). Flow trees for exploring spatial trajectories. In *17th annual GIS Research UK conference (GISRUK)*. University of Durham.

Yan, X.-Y., Zhao, C., Fan, Y., Di, Z., and Wang, W.-X. (2013). Universal predictability of mobility patterns in cities. *ArXiv e-prints*, 1307.7502.

Yano, K., Nakaya, T., Fotheringham, A. S., Openshaw, S., and Ishikawa, Y. (2003). A comparison of migration behaviour in japan and britain using spatial interaction models. *International Journal of Population Geography*, 9:419–431.

Yano, K., Nakaya, T., and Ishikawa, Y. (2000). An analysis of inter-municipal migration flows in japan using GIS and spatial interaction modelling. *Geographical Review of Japan*, 73(B):165–177.

Zheng, Y. and Zhou, X., editors (2011). *Computing with Spatial Trajectories*. Springer, New York.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort.* Addison-Wesley Press, Oxford, England.