Semi-Nonparametric Varying Coefficient Regression: Methodology, Theory and Application in Urban Economics

Xin Wang

School of Social Sciences

Heriot-Watt University

A thesis submitted for the degree of

Doctor of Philosophy

2020

To my wife and parents

Declaration

I, Xin Wang, confirm that the work presented in this thesis is my own, except where explicitly stated otherwise.

Chapter 3 of the thesis is co-authored with Prof. Arnab Bhattarcharjee (Heriot-Watt University), Prof. Tapabrata Maiti (Michigan State University) and Prof. Pingshou Zhong (University of Illinois at Chicago). Prof. Maiti and Prof. Bhattarcharjee proposed the topic and provided the dataset. Prof. Zhong outlined an early version of the model. I made substantial contributions to this chapter, including further development of the methodology, development of algorithm, data analysis, visualisation, empirical application, simulation and writing the paper.

Chapter 4 of the thesis is co-authored with Prof. Wenjie Wu (Jinan University), Prof. Jianghao Wang (Chinese Academy of Sciences) and Dr. Chengyu Li. Prof. Wu proposed the idea of proxying human and economic activity with internet-based data. Dr Li and Prof Wang provided the raw data for this project, including mobile phone positioning data and point of interest data. I conducted the rest of the research, including the development of the methodology and algorithm, data analysis, data visualisation, empirical application, simulation and writing the paper. Earlier version of the paper was archived as SERC Discussion Paper under the title '*The Geography of City Liveliness and Consumption: Evidence from Location Based Big Data*' (see Wu et al., 2016b).

Acknowledgements

First and foremost, I would like to thank my supervisors Arnab Bhattacharjee and Mark Schaffer for their continuous support during my PhD study. Without Mark's econometric lecture notes from SGPE and Arnab's passion in spatial econometrics, I would not have pursued econometrics as my PhD research topic. I also thank Tapabrata Maiti for his guidance in the realm of statistics and his incredible hospitality while I was visiting Michigan State University.

I thank my fellow PhD colleagues, Achim Ahrens, António Carvalho, Jan Ditzen, Kyle McNabb for numerous stimulating discussions and banters that made PhD life more enjoyable.

Many thanks to Wenjie Wu from Jinan University in Guangzhou, Jianghao Wang from Chinese Academy of Sciences for all the invaluable cooperation over the past two years. I would like to thank John Rapa from HSBC, for teaching me how to use Excel properly and for many insightful conversations that shaped my understanding of data analysis, geopolitics and risk management.

I'm eternally grateful to my parents for their unconditional support. I cannot be who I am today without the meals prepared by my mother and the computer bought by my father.

Last but not least, I thank my wife Dr Aodi Tang. 10 years ago we were sophomore year classmates at Nanjing University of Finance and Economics. I asked her out and promised her that one day we would study abroad together for master and PhD. I thank her for not laughing at my outlandish remarks back then and for being with me for the next 10 years pursuing our dreams.

Publications

This page lists the publications and revised and resubmitted (R&R) working papers during my PhD study.

1. Wu, Wenjie, Jianghao Wang, Chengyu Li, and Mark Wang. ''The geography of city liveliness and consumption: evidence from location-based big data." *SERC Discussion Paper* (2016)
2. Wu, Wenjie, Jianghao Wang, Tianshi Dai, and Xin Wang. ''The Geographical Legacies of Mountains: Impacts on Cultural Difference Landscapes." *Annals of the American Association of Geographers* 108, No. 1 (2018): 277-290.
3. Wu, Wenjie, Mark Xin Wang, Ning Zhu, Weiyang Zhang, and Hua Sun. "Residential satisfaction about urban greenness: Heterogeneous effects across social and spatial gradients." *Urban forestry & Urban Greening* 38 (2019): 133-144.
4. Wu, Wenjie, Mark Xin Wang, and Fangni Zhang. "Commuting behavior and congestion satisfaction: Evidence from Beijing, China." *Transportation research part D: transport and environment* 67 (2019): 553-564.
5. Wu, Wenjie, Mark Xin Wang and Guanpeng Dong. "From Access to Usage: Parks, People, and Neighbourhoods in Spatial Contexts." submitted.
6. Wu, Wenjie, Jianghao Wang, Chengyu Li and Mark Xin Wang. "The Geography of City Liveliness and Consumption: A geocomputation approach." *Journal of Regional Science*, R&R.

Abstract

This thesis presents three classes of semi-nonparametric varying coefficient regression for modelling spatial heterogeneity with cross-sectional data, panel data, and functional data, respectively, in the urban context.

Chapter 2 presents a selective review of the nonparametric and semiparametric methodologies. We first examine the estimation of a nonparametric regression using the kernel and the series methods, highlighting the cost of using the nonparametric methods. Next, we review the estimation of a varying coefficient regression and stress its relationship with the popular geographical weighted regression. Finally, we discuss the estimation of a functional linear regression, where the independent variable itself is a function. The functional principal component and Tikhonov regularisation are introduced subsequently to estimate the model.

Chapter 3 considers a spatially varying coefficient regression model over irregularly shaped areas. We develop a novel methodology that combines local polynomials and a non-Euclidean metric, called geodesic distance, to achieve both coefficient smoothing and spatial prediction over complex regions. We implement a series of Monte Carlo simulation studies to test the proposed methodology. The results suggest that our method performs better in the estimated coefficients as well as the prediction than alternative methods. Finally, we apply the method to the housing market in Aveiro, Portugal, a coastal area separated by lagoons and rivers. The results highlight the importance of modelling spatial heterogeneity and dependence in a hedonic regression.

Chapter 4 presents a spatiotemporally varying coefficient regression model which extends the spatially varying coefficient regression model into the temporal dimension. A three-dimensional local polynomial method is applied to estimate the coefficient. The Monte-Carlo simulations show that the proposed methodology outperforms the existing

geographical and temporal weighted regression. Empirically, we apply the methodology to study the relationship between human activities and consumption amenities in Beijing. To measure the human activities and the distribution of the consumption amenities, we collect two unique datasets, a high-resolution mobile phone positioning dataset from *Wechat*, a mobile social-networking application, and a point-of-interest(POI) dataset from *Meituan-Dianping*, a crowd-sourcing review website. The results show that the spatial configurations for the consumption amenities play a significant role in attracting human activities, after controlling for a wide range of location-specific characteristics. However, the effects vary substantially over space and a 24-hour time span. The results provide insights into the geographic contextual uncertainties of local amenities in shaping the rise and fall in the city liveliness.

Chapter 5 proposes a novel methodology called sieve continuum generalised method of moments to estimate a functional linear regression model. The methodology uses the sieve method to achieve dimension reduction and the continuum generalised method of moments to exploit all the moment conditions. It provides a general framework for estimating a functional linear regression with exogenous regressors as well as a functional instrumental variable regression. The proposed estimator has a closed-form which makes it easy to implement and intuitively appealing. Finally, we derive the optimal rate of convergence for the estimator.

Chapter 6 concludes with the summaries, the limitations of the thesis, as well as the directions for future researches.

List of Abbreviations

BLUP Best Linear Unbiased Predictor

CDF Cumulative Density Function

DGP Data Generating Process

FDA Functional Data Analysis

FLR Functional Linear Regression

FIVR Functional Instrumental Variable Regression

FPCA Functional Principal Component Analysis

FPCE Functional Principal Component Estimator

GIS Geographic Information System

GLS Generalised Least Squares

GMM Generalised Method of Moments

GTWR Geographically and Temporally Weighted Regression

GWR Geographically Weighted Regression

IID Independent and Identically Distributed

IV Instrumental Variable

MPP Mobile Phone Positioning

MSE Mean Square Error

OLS Ordinary Least Squares

POI Point of Interest

PDF Probability Density Function

SCGMM Sieve Continuum Generalised Method of Moments

SVCR Spatially Varying Coefficient Regression

STVCR Spatiotemporally Varying Coefficient Regression

VCR Varying Coefficient Regression

WLS Weighted Least Squares

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

The last two decades have witnessed an explosive growth of large geospatial datasets, catalysed by technological advances in areas such as computing and storage, remote sensors and monitors, as well as GIS-enabled mobile devices (Schintler and Fischer, 2018). The rise of big data presents many exciting opportunities to spatial economists. With the data that expand along both observation and variable dimensions[1], existing parametric spatial models can offer more accurate inference and prediction. On the other hand, abundant data availability encourages researchers to apply novel methods such as the nonparametric and semiparametric regressions to model spatial processes for their flexibility and robustness over parametric approaches (Pagan and Ullah, 1999; Horowitz, 2012a; Li and Racine, 2007).

This thesis focuses on a class of semi-nonparametric regressions known as the statistical varying coefficient regression (Hastie and Tibshirani, 1993; Fan and Zhang, 2008) and its applications in the urban context. The varying coefficient regressions relax the core assumption that regression coefficients are held fixed in favour of functional coefficients that are allowed to change with some underlying covariates. When the underlying covariate in a varying coefficient regression is *space*, the model sheds lights on a spatial heterogeneous relationship between the response and the explanatory variables. This regression setup is particularly attractive to spatial and urban economists whose research questions often lie in spatial heterogeneity, the structural differences over space (Anselin, 1988, 2001). Estimating a semi-nonparametric regression is more complicated than estimating

[1] Varian (2014) calls data with a large number of observations *tall* data and data with a large number of variables *fat* data.

a parametric one as the coefficients are unknown functions associated with an infinite-dimensional parameter space. Smoothing or regularisation is commonly employed to estimate the components with finite data (Chen, 2007).

Modelling spatial heterogeneity using semi-nonparametric methods is not a new idea in the spatial literature. Fotheringham and his co-authors proposed the 'Geographically Weighted Regression' (GWR) to estimate the spatially varying coefficients (see, e.g., Brunsdon et al., 1996, 1998; Fotheringham et al., 2002; Wheeler, 2019). The GWR is later extended to the temporal domain to model the spatiotemporal heterogeneity (Fotheringham et al., 2015; Ma et al., 2018). Despite its popularity in the geographic studies, the GWR has many drawbacks. Specifically, the derivation of the GWR is not founded on rigorous statistical theories and fails to relate itself with the statistical varying coefficient regression. We will illustrate using the Taylor approximation that the GWR is essentially a '*local constant estimator*', which is known to have several disadvantages such as excessive bias and poor boundary performance (see Fan and Gijbels, 1996; Li and Racine, 2007). Besides, the GWR cannot incorporate additional elements such as non-varying coefficients or spatially correlated errors, which significantly limits its application in various urban economic contexts. Furthermore, the GWR cannot work with functional spatial data, i.e., when each datum represents a surface over space.

In this thesis, we aim to close this gap by developing a suite of statistically sound varying coefficient models to facilitate the estimation and understanding of spatial heterogeneity with different data structures. In chapter 3, we develop a spatially varying coefficient regression (SVCR) for spatial cross-sectional data. Motivated by the empirical question, the model jointly incorporates four components: spatially heterogeneous effects, spatially correlated errors, irregular spatial domain, and measurement error. In chapter 4, we develop a spatiotemporally varying coefficient regression (STVCR), which extends the SVCR into the temporal dimension. In chapter 5, we study the heterogeneous coefficients in a functional linear regression (FLR) setting, which has been linked to the spatial lag model recently (Bhattacharjee et al., 2016a). We propose two estimators for the FLR with exogenous functional regressors as well as endogenous functional regressors. In both cases, we derive the rate of convergence. Empirically, we study two important issues in urban economics, urban housing and urban consumption, using the proposed methodologies. In chapter 3, we apply the SVCR to study the housing market in Aveiro, Portugal. In chapter 4, the STVCR is applied to study the spatiotemporally heterogeneous relationship between the configuration of consumption amenities and the distribution of human activities in Beijing.

1.2 Chapter 2: Selective Review of Nonparametric and Semiparametric Regression

This chapter conducts a selective review of several nonparametric and semiparametric regression techniques related to this thesis. We first consider the nonparametric regression, characterised by a regression function with an unknown functional form. Two popular methods are introduced to estimate the regression function. The first one is the kernel method, which estimates functional values by smoothing local data. The other one is the series method, which approximates a function with a finite number of basis functions. Next, we review the varying coefficient regression, a semiparametric model with the conditional mean additive in explanatory variables but heterogeneous in their partial effects. A local polynomial method is considered to estimate the model. Finally, we review the functional linear regression, which extends a multiple linear regression with a finite number of regressors to an infinite number. The functional linear regression shares some similarity with the varying coefficient regression as the parameters of interest in both cases are functional coefficients. We introduce two estimators proposed by Hall and Horowitz (2007), Tikhonov regularisation and functional principal component analysis (FPCA).

1.3 Chapter 3: Spatially Varying Coefficient Regression

In many empirical applications with spatial data, interest lies in the spatial heterogeneity of the regression coefficients. In real estate studies, for example, one may not only be interested in explaining property prices with different housing characteristics but also in understanding how the effects of these factors vary with location, as they are related to important policy and business questions.

This chapter develops a new spatial regression technique that incorporates a spatially heterogeneous coefficient over a non-convex spatial domain and a spatially dependent error with an autocorrelation structure unknown up to a finite number of parameters. To estimate the coefficients, we apply a two-dimensional Taylor approximation combined with kernel smoothing over a geodesic distance to measure the closeness between locations. The error autocorrelation, on the other hand, is estimated via variogram fitting. The methodology is suitable for empirical questions that require joint modelling of spatial heterogeneity and spatial dependency in the spatial big data context. We apply the proposed methodology to a hedonic house price model for the Aveiro urban housing market in Portugal – a coastal area divided by lagoons and rivers, and with natural holes and irregular boundaries. We focus on the spatially varying implicit price of living spaces and predicted house prices, the primary objects of inference. Application to the Aveiro housing market provides exciting new inferences on the value of living spaces and price predictions.

1.4 Chapter 4: Spatiotemporally Varying Coefficient Regression

In this chapter, we study the relationship between the spatiotemporal urban human activities and the spatial distribution of consumption amenities in Beijing. Understanding the interaction between the two variables has been a key research theme in urban economics since Glaeser et al. (2001). The human activity is measured by a unique mobile phone positioning (MPP) dataset from China's largest social networking application *WeChat*, whereas the geo-coded consumption amenity data are retrieved from China's largest local reviewing website *meituan-dianping*. These two unique big datasets offer high-resolution details, which allow us to infer the relationship between consumption amenities and human activity in Beijing at a short period (24-hour) and with high spatial granularity.

We develop a spatiotemporal varying coefficient regression (STVCR) model that allows some or all of the coefficients to vary over space and time, offering a channel to assess the spatially and temporally heterogeneous relationship. The model improves upon the geographical and temporal weighted regression (GTWR) (Fotheringham et al., 2015) by allowing the presence of non-varying coefficients and by employing local linear estimators. Simulation analysis is conducted to compare the performances between the two approaches. The results suggest that the proposed methodology outperforms the GTWR.

Empirically, we apply the proposed STVCR model to study the spatiotemporal relationship between human activity and urban consumption, including a set of control variables. We find that human activity is strongly linked to the distribution of consumption amenities in Beijing. The effects, however, are not homogeneous and vary substantially over time and space. The temporal heterogeneity largely depends on the human periodic cycle. It is stronger during the day when people are awake and weaker at night when people are asleep. The spatial heterogeneity depends on the centrality of the location. The effect is higher in the outer suburb where consumption amenities are at a premium than the central city where amenities are abundant. These results provide insights into the geographic contextual uncertainties of the consumption amenities in shaping the rise and fall in the vibrancy of urbanity.

1.5 Chapter 5: Functional Linear Regression

This chapter studies the theory of functional linear regression (FLR). The FLR has long been considered as a method not suitable for economic analysis because 'economic data are not functional data'[2]. Recently Bhattacharjee et al. (2016a) points out that a reduced-form spatial lag model with heterogeneous effects can be considered as a functional linear regression. Analysing spatial data with a functional linear regression is particularly attractive as it is capable of incorporating spatial heterogeneity and spatial dependence in a single regression framework. This chapter studies functional linear regression with exogenous regressor and functional instrumental variable regression. Both models are Fredholm integral equations of the first kind, known to be ill-posed. To estimate the slope function in each model, we propose a unified method that combines the method of sieves (Grenander, 1981; Chen, 2007) and the GMM with a continuum of moments (Carrasco and Florens, 2000). The proposed estimator has a closed-form resembling generalised least square (GLS), making it easy to implement. Under suitable assumptions, we derive the optimal minimax rate of convergence for the proposed estimators.

1.6 Chapter 6: Conclusion

This chapter summarises the thesis, highlighting the key methodological contributions and empirical findings. We also discuss the implications of the study and point to the directions of future researches.

[2] This is a direct quote from Joel Horowitz at London CEMMAP conference in 2015.

Chapter 2

Nonparametric and Semiparametric Regression: A Selective Review

2.1 Introduction

In this chapter, we conduct a selective review of several nonparametric and semiparametric regression methodologies in the literature, including nonparametric regression, varying coefficient regression, and functional linear regression. The term "regression" in this thesis refers to the statistical techniques used to estimate $\mathbb{E}(y|\mathbf{x})$, the mean of a random variable y conditional on a vector of variables $\mathbf{x} = (x_1, \ldots, x_k)^{\mathrm{T}}$. We call y dependent or response variable and $\mathbf{x}$ independent or explanatory variable. Researchers often impose various assumptions about the functional form of $\mathbb{E}(y|\mathbf{x})$. For example, a multiple linear regression assumes the conditional mean is linear in $\mathbf{x}$,

$$\mathbb{E}(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k, \tag{2.1.1}$$

where $\beta_0, \ldots, \beta_k$ are unknown scalars, or *parameters* of the regression. The possible values that these parameters are allowed to take, called the *parameter space*, is a subset of a finite dimensional Euclidean space. We call a regression with a finite number of parameters a *parametric regression*.

Alternatively, researchers can impose a minimal number of assumptions on a regression, such as

$$\mathbb{E}(y|\mathbf{x}) = f(\mathbf{x}), \tag{2.1.2}$$

where $f(\cdot)$ is an arbitrary k-variate smooth function. This type of regression is known as a *nonparametric regression*, and it involves an infinite number of unknown parameters. Finally, a *semiparametric regression* is the one that combines both parametric and nonparametric elements, such as the partially linear regression (Robinson, 1988),

$$\mathbb{E}(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{k-1} x_{k-1} + g(x_k), \tag{2.1.3}$$

where g is a function left unspecified.

Being a long-standing research topic in the statistical community (see, e.g., Stone, 1977, 1980, 1982), nonparametric and semiparametric techniques did not gain popularity in applied econometrics until recent decades. The rise of the popularity is attributed to several factors. First, economists began to recognise the specification error associated with parametric regressions. As the underlying functional relationships among economic variables are rarely known, a parametric regression often represents an 'educated guess' by a researcher. If the model assumptions deviate significantly from the actual data generating process (DGP), statistical inferences drawn from a parametric model will suffer from a sizeable bias (Wooldridge, 2010). Using motorcycle data, Fan and Gijbels (1996) illustrates in a simple bivariate setting that when the data generating process is highly nonlinear, the bias from a simple linear regression model is exceptionally large and does not diminish, even with a fourth-degree polynomial. With multiple variables, the nonlinearity effect is aggravated by the possible interactions among them. Concerns over misspecification have encouraged econometricians to develop and apply techniques that are more robust to misspecification and rely on less restrictive assumptions (see, e.g., Robinson, 1988; Ichimura, 1993).

Second, the advance of the nonparametric and semiparametric econometrics is partly driven by the need to estimate and infer intrinsically functional parameters. In spatial and urban economics, research interest not only lies in the interaction between spatial units (spatial autocorrelation) but also in spatial heterogeneity, the structural differences across space (Anselin, 1988, 2001). A real-estate economist, for example, is interested in understanding how much people are willing to pay for an extra square meter of living space in different neighbourhoods, as it helps price a property more accurately. Nonparametric and semiparametric techniques offer a flexible way to model smooth spatial heterogeneity with limited data (see Fotheringham et al., 2002, 2015).

Third, the growth of big data and computing technology also facilitates the adoption of nonparametric and semiparametric regressions. With the emergence of big economic data such as high-frequency trading data (Engle, 2000), sensor data

(Henderson et al., 2012), GPS positioning data (Wu et al., 2016a) and internet-based data (Glaeser et al., 2017), it becomes feasible to estimate more complex non-semiparametric models that require a substantial amount of data to achieve reasonable accuracy. On the other hand, estimating a semi-nonparametric regression is more computationally demanding than estimating a parametric regression. To be specific, most nonparametric and semiparametric estimators depend on one or more tuning parameters which control the smoothness of the nonparametric components. Common procedures such as 'cross-validation' require searching through all the possible grid values for the tuning parameter and compute the out-of-sample performance (Craven and Wahba, 1978; Arlot et al., 2010). The computational burden has been significantly alleviated with the improvement of computing capacity and the development of distributed computing.

This chapter does not intend to provide a comprehensive review of all the nonparametric and semiparametric methods. We will focus on the semi-nonparametric models that are *directly* related to the proposed methodologies in this thesis and highlight the advantages as well as the cost associated with these models. The rest of the chapter is organised as follows. In section 2.2, we review the nonparametric regression focusing on the kernel method. In section 2.3, we review the series estimation of a nonparametric regression. In section 2.4, we review a class of semiparametric regression known as the varying coefficient regression, and in section 2.5, we look into the functional linear regression. Section 2.6 discusses the role of the tuning parameter in semi-nonparametric models and section 2.7 concludes.

2.2 Nonparametric Regression and Kernel Method

2.2.1 Nonparametric Regression

Consider a random sample $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ that satisfy the following model,

$$y_i = g(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.2.4}$$

where $x_1, \ldots, x_n$ are the realisations of a random variable x with support $\mathcal{D}$ and cumulative density function (CDF) F; ε_i, $i = 1, \ldots, n$ are i.i.d random variables with 0 mean, and g is an unknown function defined on $\mathcal{D}$.

Regression (2.2.4) is nonparametric as $g(\cdot)$ is completely left unspecified, and the parameter space for the function is infinite-dimensional. As the sample size is always less than the dimension of the parameter space, minimising a loss function like the mean square error (MSE) without additional constraint would lead to overfitting (Chen, 2007; Li and Racine, 2007). As a result, estimating a nonparametric regression requires smoothing or regularisation. In this section, we review the kernel estimation of the nonparametric regression.

2.2.2 Nadaraya–Watson Estimator

Nadaraya (1965) and Watson (1964) proposed a kernel estimator later known as the Nadaraya–Watson estimator. Motivated by the fact that $g(\cdot)$ is the conditional mean of y given x, they suggest that $g(\cdot)$ be estimated by taking a weighted average of y_i around each $x \in \mathcal{D}$, divided by the density estimate of x,

$$\hat{g}(x) = \frac{\sum_{i=1}^{n} K_h(x - x_i) y_i}{\sum_{i=1}^{n} K_h(x - x_i)}, \tag{2.2.5}$$

where $K_h(u) = K(u/h)$, and $K : \mathbb{R} \mapsto \mathbb{R}$ is a weighting function called 'Kernel' that assigns weights to observations based on the distances between data x_is and x. Data that are close to x receive larger weights whereas those that are further away receive smaller weights.

Popular kernel functions include the Epanechnikov kernel,

$$K(v) = \frac{3}{4}(1 - v^2)\mathbb{1}(|v| \leq 1), \tag{2.2.6}$$

and the Gaussian kernel

$$K(v) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}v^2}. \tag{2.2.7}$$

h is a scalar called 'bandwidth', which controls the smoothness of the estimate.

The Nadaraya–Watson estimator is easy to implement. By evaluating the equation (2.2.5) pointwise with different x, we obtain an estimate of an entire curve. Despite its simplicity, the estimator suffers from non-trivial bias issue. Firstly, by directly taking the weighted average of local data around a point, the Nadaraya-Watson estimator does not consider the discrepancy of functional values between local data and that at the point. Secondly, the Nadaraya-Waston estimator does not perform well at areas close to the boundary of the support due to the use of disproportionately one-sided data (Fan and Gijbels, 1996). Consequently, boundary correction is often required for this type of estimator (Müller, 1991; Jones, 1993).

2.2.3 Local Polynomial Regression

Following Nadaraya (1965) and Watson (1964), alternative kernel methods have been developed to remedy the drawbacks. Cleveland (1979) introduced a smoother called local polynomial regression to smooth a scatter plot. Fan and his co-author subsequently investigated the asymptotic properties of the smoother, particularly the first-order local polynomial (local linear). They found that the local polynomial (linear) possesses superior statistical performance and does not suffer from the boundary bias compared with the Nadaraya-Watson estimator (Fan, 1992, 1993; Fan and Gijbels, 1992, 1995, 1996). Furthermore, using the Epanechnikov kernel, the local linear estimator achieves the optimal linear minimax risk (see, e.g., Fan, 1993).

The basic idea of a local polynomial regression is to combine the kernel smoothing with the Taylor approximation. For data x_is in the neighbourhood of an arbitrary $x \in \mathcal{D}$, the following p-th order Taylor approximation holds,

$$g(x_i) \approx \sum_{j=0}^{p} \frac{g^{(j)}(x)}{j!}(x_i - x)^j \tag{2.2.8}$$

where $g^{(j)}$ denote j-th order derivative with $g^{(0)}(x) := g(x)$, and $j!$ means j-th order factorial. The formula quantifies the relationship between the gradient of an arbitrary point x and the function values of the data in the neighbourhood. Define $\beta_j(x) := \frac{g^{(j)}(x)}{j!}$ and substitute the Taylor approximation of $g(x_i)$ into the nonparametric regression (2.2.4),

$$y_i \approx \sum_{j=0}^{p} \beta_j(x)(x_i - x)^j + \varepsilon_i, \quad i = 1, \ldots, n. \tag{2.2.9}$$

The equation only involves a fixed number of parameters, hence is estimable using finite data. The Taylor approximation works well when x and x_i are close and less so when they are far apart, suggesting each data point be weighted based on its relative distance to x. We consider the following kernel weighted least square objective function,

$$\sum_{i=1}^{n} \left[y_i - \sum_{j=0}^{p} \beta_j(x)(x_i - x)^j \right]^2 K_h(x_i - x), \tag{2.2.10}$$

where weights are assigned by passing the Euclidean distance between x_i and x into a kernel function. Minimising the objective functions with respect to β_js leads to the following closed-form weighted least square (WLS) estimator,

$$\hat{g}(x) := \hat{\beta}_0(x) = \mathbf{e}_{1,p}^{\mathrm{T}}(\mathbf{X}_x^{\mathrm{T}}\mathbf{W}_x^h\mathbf{X}_x)^{-1}\mathbf{X}_x^{\mathrm{T}}\mathbf{W}_x^h\mathbf{Y}, \tag{2.2.11}$$

where

$$\mathbf{X}_x = \begin{bmatrix} 1 & (x_1 - x) & \dots & (x_1 - x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (x_n - x) & \dots & (x_n - x)^p \end{bmatrix}, \tag{2.2.12}$$

$\mathbf{Y} = [y_1, \dots, y_n]^{\mathrm{T}}$, $\mathbf{W}_x^h = \mathrm{diag}(K_h(x_1 - x), \dots, K_h(x_n - x))$, and $\mathbf{e}_{1,p}^{\mathrm{T}} = [1, 0, \dots]^{\mathrm{T}}$, a $p \times 1$ vector with the first element as 1 and remaining elements as 0.

With a local polynomial regression, one needs to choose the order of the Taylor approximation, p. A larger p would reduce the approximation bias, but introduce additional unknown parameters. A smaller p, on the other hand, necessitates a smaller number of parameters. A popular choice of the order is $p = 1$, when the approximation becomes linear,

$$g(x_i) \approx g(x) + g'(x)(x_i - x), \tag{2.2.13}$$

which gives rise to the 'local linear estimator'. It is also interesting to note that when $p = 0$, the Taylor approximation takes the form $g(x_i) \approx g(x)$, and the estimator becomes

$$\hat{g}(x) = (\mathbf{u}^{\mathrm{T}}\mathbf{W}_x^h\mathbf{u})^{-1}\mathbf{u}^{\mathrm{T}}\mathbf{W}_x^h\mathbf{Y} = \frac{\sum_{i=1}^n K_h(x - x_i)y_i}{\sum_{i=1}^n K_h(x - x_i)}, \tag{2.2.14}$$

where $\mathbf{u}$ is a $n \times 1$ vector of 1. In this case, the local polynomial estimator degenerates to the Nadaraya–Watson estimator. Therefore the Nadaraya-Watson estimator is known as a type of a local constant estimator because of the 0-th order Taylor approximation.

2.2.4 Asymptotic Properties

In this section, we investigate the asymptotic properties of the local linear estimator in a univariate nonparametric regression, following Fan (1993). The local polynomial and multivariate cases are generalised in Ruppert and Wand (1994) and Masry (1996).

Assumption 2.2.1. *(Fan, 1993)*

- $g^{(2)}(\cdot)$ *is bounded.*
- $f_x(\cdot)$*, the marginal density of* x *satisfies* $|f_x(u) - f_x(v)| \leq c|u - v|^{\alpha}$*, for* $0 < \alpha < 1$.

- $\sigma^2(u) = \text{Var}(y|x = u)$ *is bounded and continuous.*
- $K(\cdot)$ *is bounded and continuous satisfying*
    1. $\int_{-\infty}^{\infty} K(v)\mathrm{d}v = 1$,
    2. $\int_{-\infty}^{\infty} vK(v)\mathrm{d}v = 0$,
    3. $\int_{-\infty}^{\infty} v^2K(v)\mathrm{d}v \neq 0$,
    4. $\int_{-\infty}^{\infty} v^{2r}K(v)\mathrm{d}v < \infty$, $r = 1, 2, \ldots$.

Theorem 2.2.2. *(Fan, 1993) Under assumption 2.2.1, the MSE of the estimator is*

$$\mathbb{E}[\hat{g}(u) - g(u)]^2 = a(u)h^4 + b(u)(nh)^{-1} + o\left(h^4 + \frac{1}{nh}\right), \tag{2.2.15}$$

where $a(u) = \frac{1}{4}\left(g^{(2)}(u)\int_{-\infty}^{\infty} v^2K(v)\mathrm{d}v\right)^2$, $b(u) = \frac{\sigma^2(u)}{f_x(u)}\int_{-\infty}^{\infty} K^2(v)\mathrm{d}v$.

The theorem shows that how fast the local linear estimator converges to the true function depends on two parameters: the sample size n and the bandwidth size h. A large h inflates the bias term $a(u)h^4$ of the MSE and diminishes the variance term $b(u)(nh)^{-1}$, whereas a small h enjoys a small bias but loses on the variance side. The success of a nonparametric estimator depends critically on how the tuning parameter is selected. The theorem also informs that the rate of convergence is strictly less than $\sqrt{n}$, a rate that is normally achieved by parametric estimators.

Fan (1993)'s theorem is established on a set of relatively restrictive assumptions such as i.i.d error terms. Recent researches have since relaxed the assumptions and extended the result to allow for non-i.i.d cases. Martins-Filho and Yao (2009) and Su et al. (2013) show how to efficiently apply local linear estimator when the errors have a non-spherical parametric covariance structure. Henderson et al. (2008) extends the nonparametric kernel estimation into the panel data setting with fixed effects. Linton and Xiao (2019) introduce a weighted local polynomial regression to take account of the dynamic heteroskedasticity in a nonparametric regression.

2.3 Nonparametric Regression and Series Method

We now turn to the series method for estimating a nonparametric regression. Unlike the kernel smoothing which models an unknown function 'locally', the series method aims to model a function 'globally' using the basis function expansion (Newey, 1997; Chen, 2007; Li and Racine, 2007).

2.3.1 Basis Functions

A basis is a collection of functions defined on some domain such that an arbitrary function f defined on the same domain satisfying certain regularity conditions can be expressed in terms of a linear combination of these basis functions,

$$f(x) = \sum_{j=1}^{\infty} \delta_j \phi_j(x), \tag{2.3.16}$$

where $\delta_j,\ j = 1, 2, \ldots$ are the coefficients associated with the basis functions. Classic examples include polynomial series, Fourier series, Spline series, Wavelet series (Newey, 1997; Belloni et al., 2015). Basis functions that satisfy the following conditions are known as the orthonormal basis,

1. $\int_{\mathcal{D}} \phi_j(x)\phi_i(x)\mathrm{d}x = 0, \quad i \neq j,$

2. $\int_{\mathcal{D}} \phi_j(x)\phi_j(x)\mathrm{d}x = 1.$

2.3.2 Series Estimator

Let $\phi_j(x),\ j = 1, 2, 3, \ldots$ be a sequence of basis functions defined on $\mathcal{D}$ such that $g(\cdot)$ can be represented as a linear combination of these functions,

$$g(x) = \sum_{j=1}^{\infty} \delta_j \phi_j(x), \tag{2.3.17}$$

where δ_js are the unknown coefficients.

Substitute the basis function expansion of $g(x_i)$ into the nonparametric regression (2.2.4) and take the first k elements of the expansion,

$$\begin{aligned} y_i &= \sum_{j=1}^{\infty} \delta_j \phi_j(x_i) + \varepsilon_i, \quad i = 1, \ldots, n \qquad (2.3.18) \\ &\approx \sum_{j=1}^{k} \delta_j \phi_j(x_i) + \varepsilon_i. \end{aligned}$$

By approximating an unknown function with k basis functions, we immediately reduce the dimension of the parameter space from infinity to k. The approximation error is the remainder of the expansion $\sum_{j=k+1}^{\infty} \delta_j \phi_j$. When $k < n$, δ_js can be estimated by minimising the following objective function with respect to these coefficients,

$$\sum_{i=1}^{n}\left[y_i - \sum_{j=1}^{k}\delta_j\phi_j(x_i)\right]^2. \tag{2.3.19}$$

This gives a least square type estimator,

$$\hat{\Delta}_k := (\hat{\delta}_1, \dots, \hat{\delta}_k)^{\mathrm{T}} = (\mathbf{\Phi}_k^{\mathrm{T}}\mathbf{\Phi}_k)^{-1}\mathbf{\Phi}_k^{\mathrm{T}}\mathbf{Y}, \tag{2.3.20}$$

where

$$\mathbf{\Phi}_k = \begin{bmatrix} \phi_1(x_1) & \dots & \phi_k(x_1) \\ \vdots & \ddots & \vdots \\ \phi_1(x_n) & \dots & \phi_k(x_n) \end{bmatrix}. \tag{2.3.21}$$

Associate each $\hat{\delta}_j$ with its basis function, and we obtain an estimator for $g(\cdot)$,

$$\begin{aligned} \hat{g}(x) &= \sum_{j=1}^{k}\hat{\delta}_j\phi_j(x) \\ &= \boldsymbol{\phi}_k^{\mathrm{T}}(x)(\mathbf{\Phi}_k^{\mathrm{T}}\mathbf{\Phi}_k)^{-1}\mathbf{\Phi}_k^{\mathrm{T}}\mathbf{Y}, \end{aligned}$$

where $\boldsymbol{\phi}_k^{\mathrm{T}}(x) = [\phi_1(x), \dots, \phi_k(x)]$.

2.3.3 Splines

In this section, we discuss a popular choice of the basis function system known as 'splines', widely used for modelling non-periodic functions.

A spline is essentially a function built from piecewise polynomial functions. For a real interval $[a, b]$, where $a < b$, we could partition the interval into m sub-intervals: $[t_0, t_1]$, $[t_1, t_2]$, $\dots$, $[t_{m-1}, t_m]$, where $a = t_0 < t_1 < \cdots < t_m = b$. Then a basis function ϕ is a k-th order spline with knots at $t_1, \dots, t_{m-1}$ if the following two conditions hold,

1. ϕ is a polynomial of degree k on each of the sub-interval.

2. The j-th derivative of ϕ is continuous at knots $t_1, \dots, t_{m-1}$, for $j = 1, 2, \dots, k-1$.

For example, when $k = 3$, ϕ is known as a 'cubic spline', composed of piecewise cubic functions with continuous first and second order derivatives. For a given set of knots, a spline could be represented in terms of truncated power series, $\phi_1(x) = 1$, $\phi_2(x) = x$, $\dots$, $\phi_{k+1} = x^k$, $\phi_{k+2} = (x - t_1)_+^k$, $\dots$, $\phi_{k+m+1} = (x - t_m)_+^k$, where $x_+ := max\{x, 0\}$. See De Boor et al. (1978) and Schumaker (2007) for a detailed discussion.

Perform a regression of y on the spline functions yields the so-called 'regression spline' estimate of the regression function. However, the estimate is prone to have high variance, particularly at boundaries. Furthermore, one needs to decide where to place the knots.

The smoothing spline method is designed to tackle the issues. Interestingly, it can be motivated from the perspective of functional optimisation. Smoothing spline aims to find a function g, which minimises the following objective function,

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int (g''(x_i))^2 \mathrm{d}x \tag{2.3.22}$$

where g'' is the second order derivative of g, and λ is a tuning parameter that controls the weight assigned to the second term of the objective function. The objective function creates a trade-off between the least square of error and the smoothness of g. Functions that are wiggly attracts more penalty through the second derivative, whereas smooth functions receive less penalty. It turns out that the objective function has a unique solution, and the solution is a cubic spline with knots at data points $x_1, \ldots, x_n$ (De Boor et al., 1978).

2.3.4 Rate of Convergence

In this section, we discuss the asymptotic properties of the series method. We follow Newey (1997) and focus on i.i.d data.

Assumption 2.3.1. *$(x_i, y_i), i = 1, \ldots, n$ are i.i.d data, conditional variance* $\mathrm{Var}(y|x)$ *is bounded. $x_i \in \mathcal{D}$, which is compact.*

Assumption 2.3.2. *Uniformly in k, the eigenvalues of $\mathbb{E}[\Phi_k(x_i)\Phi_k(x_i)^{\mathrm{T}}]$ are bounded away from 0, where $\Phi_k(x_i) = (\phi_1(x_i), \ldots, \phi_k(x_i))^{\mathrm{T}}$*

Assumption 2.3.3. *There is a sequence of constants $\zeta(k)$ satisfying $\sup_{x\in\mathcal{D}} |\phi_k(x)| \leq \zeta(k)$, where $k = k(n)$ such that $\zeta(k)^2/n \to 0$ as $n \to \infty$.*

Assumption 2.3.4. *There exists α and $\Delta_k = (\delta_1, \ldots, \delta_k)$ such that $\sup_{x\in\mathcal{D}} |g(x) - \Phi_k^{\mathrm{T}}\Delta_k| = O(k^{-\alpha})$ as $k \to \infty$*

Assumption 2.3.5. *As $n \to \infty$, $k \to \infty$ and $k/n \to \infty$*

Theorem 2.3.6. *Under assumption (2.3.1),(2.3.2),(2.3.3),(2.3.4) and (2.3.5), we have*

$$\int [\hat{g}(x) - g(x)]^2 \mathrm{d}F(x) = O_p(k/n + k^{-2\alpha}) \tag{2.3.23}$$

Theorem 2.3.6 suggests that the error of the series estimator can be decomposed into two parts, the approximation error ($k^{-2\alpha}$) which governs the bias of the estimator and the estimation error (k/n), associated with the variance. A larger k brings more basis functions into the equation and leads to a smaller approximation bias but loses on variance side. A smaller k improves the variance but suffers a higher approximation bias. The number of the basis functions k plays the same role as the bandwidth h in a kernel estimation.

Building on the theory of the series method pioneered by Newey (1997), Belloni et al. (2015) recently developed some new asymptotic results that considerably weaken the assumptions required for the convergence of the series estimator. In particular, they establish that the condition $\zeta(k)^2/n \to 0$ can be relaxed to $\zeta(k)/n \to 0$.

2.4 Varying Coefficient Regression

In this section, we move from the nonparametric regression to the semiparametric regression which imposes some structures on the regression function. A popular type of the semiparametric regression is known as the 'varying coefficient regression' (Hastie and Tibshirani, 1993), specified as follows,

$$y_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}(z_i) + \varepsilon_i, \tag{2.4.24}$$

where $\mathbf{x}_i = (x_{1,i}, \ldots, x_{k,i})^{\mathrm{T}}$ and $\boldsymbol{\beta}(z_i) = (\beta_1(z_i), \ldots, \beta_k(z_i))^{\mathrm{T}}$, a vector of unknown functions of the covariate z_i. For example, z_i could represent spatial coordinates and $\boldsymbol{\beta}(z_i)$ reflects spatially varying effects.

The varying coefficient regression offers many useful features by combining both parametric and nonparametric components. As the partial effect of each variable $x_{k,i}$ is solely captured by $\beta_k(z_i)$, a VCR has better interpretability than a nonparametric regression. $\beta_k(z_i)$, on the other hand, is allowed to change with some covariate z_i, providing a channel to model non-stationary coefficient heterogeneity.

2.4.1 Local linear estimator

The local linear estimator has been a popular method to estimate a varying coefficient regression and the mechanism is similar to estimating a nonparametric regression. To start with, we apply the first order Taylor approximation to $\boldsymbol{\beta}$,

$$\boldsymbol{\beta}(z_i) \approx \boldsymbol{\beta}(z) + \boldsymbol{\beta}^{(1)}(z)(z_i - z), \tag{2.4.25}$$

where $\boldsymbol{\beta}^{(1)}(z) = (\frac{\mathrm{d}\beta_1}{\mathrm{d}z}, \ldots, \frac{\mathrm{d}\beta_k}{\mathrm{d}z})$, and substitute the approximation into the VCR model,

$$y_i \approx \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}(z) + \mathbf{x}_i^{\mathrm{T}}(z_i - z)\boldsymbol{\beta}^{(1)}(z) + \varepsilon_i. \tag{2.4.26}$$

Both $\boldsymbol{\beta}(z)$ and $\boldsymbol{\beta}^{(1)}(z)$ can be estimated by minimising the following kernel weighted sum of squares of errors,

$$\sum_{i=1}^{n} \left\{y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}(z) - \mathbf{x}_i^{\mathrm{T}}(z_i - z)\boldsymbol{\beta}^{(1)}\right\}^2 K_h(z_i - z).$$

We could write the estimator compactly as

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}(z) \\ \hat{\boldsymbol{\beta}}^{(1)}(z) \end{bmatrix} = (\mathbf{G}^{\mathrm{T}}\mathbf{W}_h\mathbf{G})^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{W}_h\mathbf{Y}, \tag{2.4.27}$$

where $\mathbf{Y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, $\mathbf{W}_h = \mathrm{diag}(K_h(z_1 - z), \ldots, K_h(z_n - z))$, and $\mathbf{G} = (\mathbf{x}_i^{\mathrm{T}}, \mathbf{x}_i^{\mathrm{T}}(z_i - z))_{i=1,\ldots,n}$.

The asymptotic properties of the local linear estimator in the VCR model are similar to that in the nonparametric regression and have been substantially studied and reviewed in Fan and Zhang (2008).

2.4.2 Geographical Weighted Regression

The varying coefficient regression is widely used in various subjects to model coefficient heterogeneity. In geography, the model is applied under the name 'geographical weighted regression' (GWR), and often without mentioning its statistical foundation (Brunsdon et al., 1996, 1998; Fotheringham et al., 2002; Wheeler, 2019).

A standard GWR is specified as follows,

$$y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}(a_i, b_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.4.28}$$

where z_i in equation (2.4.24) is replaced by a_i, b_i, the latitude and longitude defined over a two-dimensional spatial domain $\mathcal{S}$. To estimate $\boldsymbol{\beta}(a, b)$ for each $(a, b) \in \mathcal{S}$, the GWR suggests using the following kernel estimator,

$$\hat{\boldsymbol{\beta}}(a, b) = (\mathbf{X}^{\mathrm{T}}\mathbf{W}_h\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}_h\mathbf{Y}. \tag{2.4.29}$$

It could be shown that the estimator corresponds to a local constant nonparametric regression where the Taylor approximation takes the following form,

$$\boldsymbol{\beta}(a_i, b_i) \approx \boldsymbol{\beta}(a, b), \tag{2.4.30}$$

for (a_i, b_i) close to (a, b). As a result, the GWR inherits all the disadvantages of the local constant estimator such as excessive bias and poor performance at boundaries.

2.4.3 Varying Coefficient Regression with Discontinuity

A key assumption of the varying coefficient regression is that the coefficients are smooth functions of the underlying domain. Several recent researches have explored varying coefficient regressions with a discontinuity on the domain. Zhu et al. (2014) propose a spatially varying coefficient regression model, which simultaneously captures (1) spatially varying coefficients, (2) multiple piecewise smooth regions with unknown edges and jumps, (3) substantial spatial correlations. The model is specified as follows,

$$y_i(\mathbf{d}) = \mathbf{x}_i^T \boldsymbol{\beta}(\mathbf{d}) + \eta_i(\mathbf{d}) + \epsilon_i(\mathbf{d}) \tag{2.4.31}$$

where $\mathbf{d} \in \mathcal{S}$ is a spatial location. Zhu et al. (2014) develop a multiscale adaptive and sequential smoothing (MASS) method which incorporates the propagation-separation, functional principal component analysis, and jumping surface model. The approach is capable of estimating the coefficients while simultaneously identifying the edges of the discontinuous regions.

2.5 Functional Linear Regression

The nonparametric and semiparametric regressions covered in the previous sections focus on fitting data with finite dimensions. In this section, we consider a special class of the semiparametric models called functional linear regression (FLR), where the independent variable X itself is a function (Wang et al., 2016).

Consider i.i.d data (X_i, y_i), $i = 1, \ldots, n$ that are generated by the following model,

$$y_i = \int_0^1 X_i(s)\beta(s)\mathrm{d}s + \varepsilon_i. \tag{2.5.32}$$

y_i is a scalar, whereas X_i is a random function defined on $[0,1]$. Without loss of generality, we assume X_i is centred in the sense that $\mathbb{E}(X_i)=0$. ε_i is i.i.d with 0 mean and finite variance σ^2.

2.5.1 Inverse Problem and Spectral Decomposition

The independence between ε and X suggests that for $t \in [0,1]$, the covariance between X and ε is 0,

$$\begin{aligned}\mathbb{E}[X(t)\varepsilon] &= \mathbb{E}[X(t)Y] - \int_0^1 \mathbb{E}[X(s)X(t)]\beta(s)\mathrm{d}s \\ &= 0\end{aligned}$$

Let $g(t) := \mathbb{E}[X(t)Y]$, the covariance between Y and X, and let $Kf(t) := \int_0^1 [X(s)X(t)]f(s)\mathrm{d}s$, the covariance operator of X. The relationship of X and ϵ can be summarised into

$$K\beta = g. \tag{2.5.33}$$

The equation is known as the Fredholm integral equation of the first kind (Kress et al., 1989). Assuming K is invertible and K^{-1} is the inverse of K, then β is solved by

$$\beta = K^{-1}g. \tag{2.5.34}$$

Intuitively, if K^{-1} and g are to be estimated from the data, then the β estimator would follow by equation (2.5.34). However, this turns out to be challenging as K^{-1} is not a continuous operator, and a small perturbation in g would be translated into a large deviation in β estimate.

To appreciate the challenge, we conduct a spectral decomposition of the equation. Note that the covariance operator K is symmetric and positive definite, therefore there exists a sequence of positive and non-increasing eigenvalues k_js, and a sequence of orthonormal[1] eigenfunctions $\phi_j(\cdot)$s satisfying,

$$K\phi_j = k_j\phi_j, \quad j = 1,2,3,\dots. \tag{2.5.35}$$

The eigenfunctions serve as an orthonormal basis for the functional space where they are defined. Specifically, any function f defined on $[0,1]$ with finite norm can be expressed in terms of ϕ_js,

$$f(s) = \sum_{j=1}^{\infty} \alpha_j\phi_j(s), \tag{2.5.36}$$

[1]Orthonormal means that $\int_0^1 \phi_i\phi_j = \delta_{ij}$ where $\delta_{ij}=0$ if $i \neq j$ and 1 if $i=j$.

where $\alpha_j = \int_0^1 f(s)\phi_j(s)\mathrm{d}s,\ j = 1, 2, 3, \ldots$ are known as the generalised Fourier coefficients. Likewise, we can expand β and g function using the eigenfunctions,

$$\beta(s) = \sum_{j=1}^{\infty} b_j \phi_j(s), \tag{2.5.37}$$

$$g(s) = \sum_{j=1}^{\infty} g_j \phi_j(s), \tag{2.5.38}$$

where b_js and g_js are the generalised Fourier coefficients for β and g respectively.

Substitute (2.5.37) and (2.5.38) into (2.5.33), we obtain

$$\begin{aligned}\sum_{j=1}^{\infty} b_j K\phi_j(s) &= \sum_{j=1}^{\infty} b_j k_j \phi_j(s) \\ &= \sum_{j=1}^{\infty} g_j \phi_j(s),\end{aligned}$$

suggesting $b_j k_j = g_j$ for each j. Substitute $b_j = g_j/k_j$ back into (2.5.37), we obtain the solution of β based on expanded series,

$$\beta(s) = \sum_{j=1}^{\infty} \frac{g_j}{k_j} \phi_j(s). \tag{2.5.39}$$

Equation (2.5.39) highlights the challenge of estimating a functional linear regression from a different perspective. Note that k_js appears in the denominator and is a sequence of scalars converging towards 0. For a large j, a small estimation error in g_j would be inflated dramatically through k_j^{-1}. Furthermore, k_j and ϕ_j themselves are estimated from data, adding additional uncertainties into the estimator of β.

2.5.2 Functional Principal Component Estimator

Hall and Horowitz (2007) suggest taking the first k components of the β expansion and replacing k_j, ϕ_j, g_j with their empirical counterparts $\hat{k}_j, \hat{\phi}_j, \hat{g}_j$ to form the functional principal component estimator (FPCE),

$$\hat{\beta}(s) = \sum_{j=1}^{k} \frac{\hat{g}_j}{\hat{k}_j} \hat{\phi}_j(s), \tag{2.5.40}$$

where $\hat{k}_j$ and $\hat{\phi}_j$ are estimated from the empirical covariance operator $\hat{K}$, with $\hat{K}\hat{\phi}_j = \hat{k}_j\hat{\phi}_j$. $\hat{g}_j$ is estimated via $\int_0^1 \hat{g}(x)\hat{\phi}_j(x)\mathrm{d}x$. Terms after k are excluded from the estimator, making the result more stable. k is a tuning parameter picked by a researcher. A smaller k means more items are excluded, introducing extra biases, whereas a larger k would lead to smaller eigenvalues being included, increasing the variance.

The FPCE has a similar structure to the series estimator. Both are represented using a finite number (k) of basis functions, where k plays a critical role in balancing the variance and bias. With a series estimator, the basis functions are pre-determined and non-stochastic, whereas the FPCE applies the empirical eigenfunctions that are data-dependent.

2.5.3 Tikhonov Regularisation

Another popular method to estimate the FLR is known as the Tikhonov regularisation, or Ridge regression, due to Hall and Horowitz (2005, 2007). Tikhonov regularisation could be motivated in several ways, and here we start from the spectral decomposition of the β (equation 2.5.39). To avoid k_j^{-1} from diverging, Tikhonov regularisation suggests modifying k_j^{-1} with $(k_j + \lambda)^{-1}$, where λ is a positive number. Choosing λ properly, we keep $k_j + \lambda$ away from 0 for large j. The Tikhonov regularisation estimator is defined as follows,

$$\hat{\beta}(s) = \sum_{j=1}^{\infty} \frac{\hat{g}_j}{\hat{k}_j + \lambda} \hat{\phi}_j(s). \tag{2.5.41}$$

λ is the tuning parameter in Tikhnov regularisation. As n goes to infinite, λ moves slowly towards 0 so that $(\hat{k}_j + \lambda)$ converges to k_j.

2.5.4 Rate of Convergence

Now we discuss the rate of convergence of the FLR estimators. We focus on the functional principal component estimator, whereas results for Tikhonov regularisation can be found in Hall and Horowitz (2005).

Assumption 2.5.1. *Let $C > 1$ be a constant.*

1. *X has finite 4th moment, $\int_0^1 \mathbb{E}[X^4(s)]ds < \infty$.*

2. *$\mathbb{E}(\xi_j^4) \leq C k_j^2$ uniformly in j, where ξ_j is the jth coefficient of the basis function expansion of X with eigenfunctions.*

3. *ε_i follows i.i.d distribution with 0 mean and variance $\mathrm{Var}(\varepsilon) < C$*

Assumption 2.5.2 (Eigenvalue Spacing).

$$k_j - k_{j+1} \geq C^{-1} j^{-\alpha-1}, \quad \alpha > 1, \quad j \geq 1 \tag{2.5.42}$$

Assumption (2.5.2) outlines the spacing between consecutive eigenvalues. It prevents the difference between two consecutive eigenvalues from being too small. In particular, it rules out the existence of tied eigenvalues.

Assumption 2.5.3 (Parameter Smoothness).

$$|b_j| \geq C j^{-\beta}, \quad 0.5\alpha + 1 < \beta \tag{2.5.43}$$

Assumption (2.5.3) sets out the smoothness of the functional parameter with respect to the covariance operator.

Assumption 2.5.4 (Tuning Parameter).

$$k \asymp n^{1/(\alpha+2\beta)} \tag{2.5.44}$$

Assumption (2.5.4) decides how fast the tuning parameter should converge to infinity.

Theorem 2.5.5. *(Hall and Horowitz, 2007) For each $\mathcal{F}$ be a class joint distributions of (X, Y) satisfying (2.5.1), and if (2.5.1)– (2.5.4) holds, for each $F \in \mathcal{F}$*

$$\int_0^1 (\hat{b} - b)^2 = O_p(n^{-(2\beta-1)/(\alpha+2\beta)}) \tag{2.5.45}$$

Theorem (2.5.5) states that the rate of convergence depends negatively on the smoothness of the covariance operator and positively on the smoothness of the β and strictly less than n^{-1}.

2.6 Tuning Parameter

All of the nonparametric and semiparametric estimators introduced in this chapter require some tuning parameters. With local polynomial, linear and constant estimators, the tuning parameter h determines the window size in local smoothing. With series and functional principal component estimators, the tuning parameter k represents the number of basis functions to take. With Tikhonov regularisation, the tuning parameter λ measures the deviation from the original eigenvalues. h, k, λ control the systematic bias introduced into the estimation process. As the sample size n goes to infinity, these tuning parameters need to move towards the direction with less bias to achieve the consistency of the estimators. For example, the number of the basis functions k in a series estimator needs to grow with n to diminish the approximation error. For a given sample size n, it is critical to pick an appropriate value for a tuning parameter to balance the bias and variance.

In this section, we discuss a tuning parameter selection procedure called cross-validation (Arlot et al., 2010). With a cross-validation process, a tuning parameter is selected that optimises the out-of-sample performance of the model. To illustrate the concept, let $y_1, \ldots, y_n$ be a set of actual values of y and $\hat{y}_1(m), \ldots, \hat{y}_n(m)$ be the out-of-sample estimated values using a semi-nonparametric model with a tuning parameter value m. The out-of-sample performance is then summarised by the mean square error,

$$\mathrm{MSE}(m) = \frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j(m))^2 \tag{2.6.46}$$

The best tuning parameter is chosen by minimising the MSE with respect to m

$$m^* = \arg\min_m \mathrm{MSE}(m) \tag{2.6.47}$$

To estimate the out-of-sample $\hat{y}_j$, it is common to split the data into a training sample and a test sample[2], then evaluate the out-of-sample prediction for the test sample using models estimated by the training sample.

2.7 Conclusion

The nonparametric and semiparametric econometrics has experienced rapid growth during the past decades. The core principles of estimating a nonparametric and semiparametric regression, on the other hand, have largely remained consistent. It involves applying regularisation to balance the variance and the bias of an estimator. This chapter reviews a set of methodologies and regularisation approaches that are related to this thesis. For a more comprehensive review on econometric applications, please refer to Pagan and Ullah (1999), Horowitz (2012a) and Li and Racine (2007). For a comprehensive review on functional data analysis (FDA) and functional linear regression, please refer to Ramsay and Silverman (2007), Ramsay and Silverman (2005) and Wang et al. (2016).

[2]Train-test split is a term borrowed from the machine learning literature, see Hastie et al. (2015).

Despite the rapid theoretical and methodological development in nonparametric and semiparametric regressions, several challenges are still present. The first issue is related to high dimensionality. With a large dimension of explanatory variables, nonparametric regressions suffer from the 'curse of dimensionality' when the data required to accurately estimate a model become astronomically large[3]. The phenomenon is aggravated when the number of variables exceeds the number of observations (Yang and Tokdar, 2015). The second challenge is associated with endogeneity. In consumer behaviour studies, economists are interested in estimating the nonparametric Engel curve, which describes the relationship between a household's total expenditure and spending on particular goods (Chai and Moneta, 2010; Blundell et al., 2003, 2008). As the total expenditure has been considered as endogenous, i.e., correlated with the error term, standard semi-nonparametric estimators suffer from an asymptotic bias. To tackle the endogeneity in semi-nonparametric regressions, nonparametric instrumental variable estimators have been developed and received significant attention in the recent literature (Newey and Powell, 2003; Hall and Horowitz, 2005; Blundell et al., 2007; Darolles et al., 2011; Horowitz, 2011; Chen and Pouzo, 2012; Gagliardini and Scaillet, 2012).

[3]Stone (1982) finds that the convergence rate of nonparametrically estimated function is $n^{-\frac{p}{2p+d}}$, where p measures the smoothness of the function and d is the dimension of the function. The convergence rate drops dramatically as d increases.

Chapter 3

Spatially Varying Coefficient Regression Over Irregularly Shaped Regions: Application to a Hedonic House Price Model

3.1 Introduction

The central object of this chapter is to develop a new spatial regression methodology called spatially varying coefficient regression (SVCR) that simultaneously models spatial heterogeneity and spatial autocorrelation over an irregular domain and in the presence of a large dataset. The methodology is applied to a hedonic house price model, a popular tool in both applied research and business to understand the variation of house prices. Pioneered by Lancaster (1966) and Rosen (1974), a hedonic model regresses the value of a property on its attributes, such as structural and neighbourhood factors, therefore provides a way to characterise markets for heterogeneous goods. The regression coefficient for each attribute gives rise to the estimated price of, or the willingness to pay for, the attribute. It is also known as the 'shadow price', as it is not directly observed from the market. Empirical examples of hedonic models include the price of air quality (Anselin and Lozano-Gracia, 2008), clean water (Anselin et al., 2010) and living space (Bhattacharjee et al., 2016a). Effective use of the hedonic models in business and policy requires an accurate and fast estimation of the regression. However, available methods in the literature have not placed adequate attention to several essential features.

[0]Earlier version of this chapter was co-authored with Arnab Bhattacharjee, Taps Maiti, Pingshou Zhong under the title '*Spatially Varying Regression Over Irregularly Shaped Regions: Application to a Hedonic House Price Model*' (see Bhattacharjee et al., 2016b)

First of all, there exists substantial spatial heterogeneity in the implicit prices. Urban housing markets are usually segmented with complicated spatial patterns (Bhattacharjee et al., 2016a). Residents with different preferences and socioeconomic backgrounds tend to be sorted into different spatial clusters (Galster, 2001). As a result, there exist several housing submarkets, each with a different equilibrium (Rothenberg et al., 1991). This suggests the shadow prices are varying over spatial domains (Bhattacharjee et al., 2016a). For example, in the context of our empirical application, we expect the implicit price of the living spaces to be higher in the central areas where larger houses are at a premium, as opposed to the suburban areas where space is less scarce. Second, there exist strong spatial spillovers or contagion between different houses in the same submarket, and between submarkets (Gillen et al., 2001; Anselin et al., 2010). The recent literature has discussed the potential bias and loss of efficiency when such spatial effects are ignored (see, e.g., Anselin and Lozano-Gracia, 2008; LeSage and Pace, 2009; Anselin et al., 2010). Therefore, the spatial autocorrelation in house prices needs to be adequately accounted for. Third, in many spatial applications with housing markets, the Euclidean distance is often implicitly applied to measure the closeness of two locations for the purpose of modelling spatial heterogeneity or spatial autocorrelation. However, this is not appropriate with a non-convex spatial domain. In regions with irregular boundaries, peninsulas, and interior, closeness in the Euclidean sense does not necessarily mean similarity between two locations (Wang and Ranalli, 2007). Fourth, whereas the literature primarily focuses on the estimation of hedonic models, we place special emphasis on the prediction of house prices which is of key policy and business interest. When the response variable is modelled as a linear function of several covariates, the best linear unbiased predictor (BLUP) can be obtained using a covariance function that may be unknown up to a finite number of parameters (Cressie, 1990). BLUP formula requires inverting an $n \times n$ covariance matrix with n being the sample size. With a large n, this becomes computationally challenging.

To address these concerns, we develop a spatially varying coefficient regression (SVCR) that jointly deals with spatial heterogeneity, spatial autocorrelation, irregular domains and computational issues with large datasets. First, the regression function is specified as semiparametric with some coefficients changing smoothly over space. This feature captures spatially heterogeneous effects. Second, the spatial autocorrelation is embedded in the spatial error term, which is modelled as a parametric function of distance with finite parameters. To measure the (dis-)similarity between two locations, we replace the Euclidean distance with the geodesic distance (Wang and Ranalli, 2007). The geodesic distance calculates the closeness between two locations using the shortest path inside the domain, rather than a straight line. To estimate the model, we apply the first-order Taylor approximation with a kernel and inverse covariance weighting scheme. The estimation and prediction are conducted locally, which substantially reduces the computational burden of inverting a covariance matrix.

The methodology is applied to the housing market of Aveiro in central Portugal, including the twin municipality of Ílhavo and the adjoining peri-urban and rural areas (Figure 3.1). We regress the logarithm of the house price per square meter on a large collection of housing characteristics, including our key covariate, the living space. The other covariates include the internal and external features of the house, the quality of the residential neighbourhood, and access to local and central facilities. These additional covariates are combined, using factor analysis, into five factors, each having a distinct economic interpretation. Then, the estimates of the regression coefficients are interpreted as the shadow (or implicit) prices for the relevant characteristics, that is, the willingness to pay for each such feature.

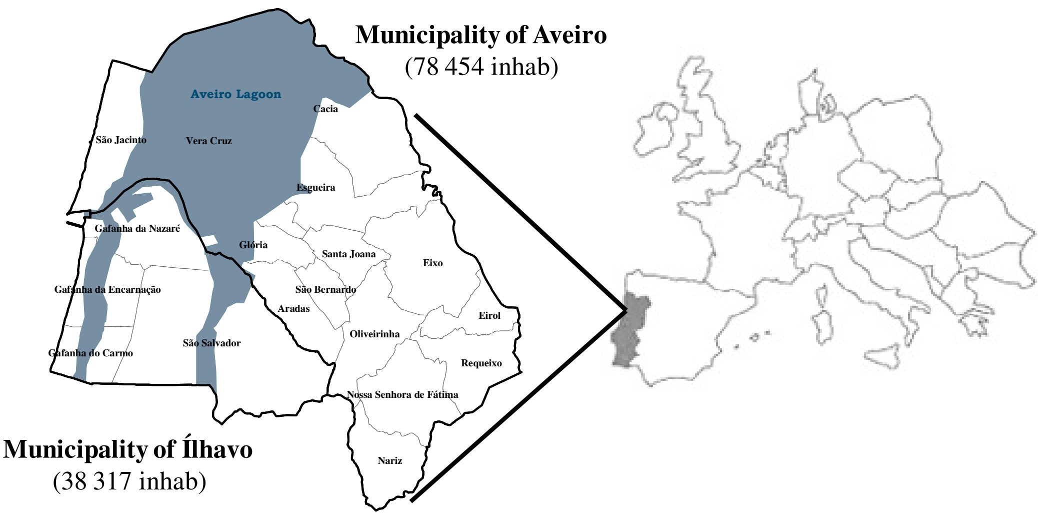

Figure 3.1: Municipalities of Aveiro and Ílhavo

This chapter makes methodological contributions to three domains of spatial researches. The first area is spatial prediction (or Kriging, see Cressie (1993)). The SVCR relaxes the assumption that the mean structure is parametric in favour of a semiparametric form. The errors can be spatially correlated and potentially heteroskedastic over the domain. Compared with the existing Kriging regression, the methodology is more flexible and suffers from less specification bias.

Besides, the estimation and prediction do not require the inversion of an $n \times n$ covariance matrix, but only matrices of the order $n(s(h))$, where n is the total sampled locations and $n(s(h))$ is the local sample size. This enriches researches of Kriging with big data (see, e.g., Cressie and Johannesson, 2008). The second area is related to the modelling of spatial heterogeneity. The proposed methodology improves significantly upon the popular geographically weighted regression (GWR), commonly used in geography to estimate spatial heterogeneity (Brunsdon et al., 1996, 1998; Fotheringham et al., 2002). To estimate the spatially heterogeneous coefficients, we extend the local linear approach to a two-dimensional space, and provide statistical inference on the estimated functional coefficient as well as predicted response. The approach enjoys many statistical advantages over the GWR-type estimator, discussed in Fan and Gijbels (1996) and Fan and Zhang (2008). A simulation study is carried out to compare the performances between the two approaches. Finally, we contribute to the literature of spatial smoothing over irregular domains. Wang and Ranalli (2007) suggested a modified version of low-rank thin plate splines (LTPS) where the Euclidean distance is replaced by the geodesic distance. While the distance-based approaches are conceptually similar to the modified kernel regression, this connection is not highlighted in the literature. We make the link explicit by building a kernel-based local linear regression methodology, where the kernel is based on a distance function that takes the complex nature of the spatial domain into account. Our methodology shares some similarity with Zhu et al. (2014) which proposed a spatially varying coefficient model for smoothing neuroimaging data with unknown edges and jumps. However, they consider only smoothing not spatial prediction. Further, they did not use the geodesic distances which restrict applications in irregular domains.

The chapter is organised as follows. Section 3.2 develops our models and methodology followed in section 3.3 by simulation studies. Section 3.4 develops the application to the urban housing market of Aveiro, Portugal. Finally, section 3.5 concludes.

3.2 Spatially Varying Coefficient Regression

In this section, we develop the methodology for the spatially varying coefficient regression over an irregular domain. Consider a spatial sample $\{(Y_i, \mathbf{x}_i, \mathbf{s}_i), i = 1, \ldots, n\}$ generated by the following model,

$$Y_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}(\mathbf{s}_i) + \epsilon_i + \eta_i, \quad i = 1, \ldots, n, \tag{3.2.1}$$

where $\mathbf{s}_i = (a_i, b_i) \in D \subset \mathbb{R}^2$ represents the x coordinate and y coordinate in a two-dimensional spatial domain D. D is a non-convex region equipped with a suitable distance measure (e.g., geodesic distance, see Wang and Ranalli (2007)), denoted by $d(\cdot, \cdot)$. $\mathbf{x}_i = [x_{1,i}, \ldots, x_{p,i}]^{\mathrm{T}}$ is a vector of $p-$dimensional explanatory variables and $\boldsymbol{\beta}(\mathbf{s}) = [\beta_1(\mathbf{s}), \ldots, \beta_p(\mathbf{s})]^{\mathrm{T}}$ is the vector of $p-$dimensional functional coefficients. Each β_j, $j = 1, \ldots, p$ is a smooth function over D. η_i is a measurement error term which follows a Gaussian process with 0 mean, known variance $\sigma^2_\eta(\mathbf{s}_i)$ and $\mathrm{Cov}(\eta_i, \eta_j) := \Sigma_\eta(\mathbf{s}_i, \mathbf{s}_j)$[1]. ϵ_i is an independent Gaussian process with 0 mean, variance $\sigma^2_\epsilon(\mathbf{s}_i)$, and covariance $\mathrm{Cov}(\epsilon_i, \epsilon_j) = \sigma_\varepsilon(\mathbf{s}_i)\sigma_\varepsilon(\mathbf{s}_j)\rho(d(\mathbf{s}_i, \mathbf{s}_j))$ where $\rho(d(\mathbf{s}_i, \mathbf{s}_j))$ is the correlation function depending on distance.

We are interested in estimating $\boldsymbol{\beta}(\mathbf{s}_i)$ as well as predicting Y at unsampled location $\mathbf{s}_0$ with $\mathbf{x}_0$.

3.2.1 Estimation

Two-Dimensional Taylor Approximation

For a generic location $\mathbf{s} = (a, b) \in D$, define the neighbourhood of $\mathbf{s}$ as a collection of nearby sample locations,

$$s(h) = \{\mathbf{s}_i = (a_i, b_i), i = 1, \ldots, n : d(\mathbf{s}, \mathbf{s}_i) \leq h \text{ for some small } h > 0\}, \quad (3.2.2)$$

where h is a bandwidth that controls the size of the neighbourhood. Let $n(s(h))$ be the number of locations in the set $s(h)$. For a small h and for an $\mathbf{s}_i$ in $s(h)$, the following two-dimensional Taylor approximation holds,

$$\begin{aligned} \boldsymbol{\beta}(\mathbf{s}_i) &\approx \boldsymbol{\beta}(\mathbf{s}) + \boldsymbol{\beta}^{(a)}(\mathbf{s})(a_i - a) + \boldsymbol{\beta}^{(b)}(\mathbf{s})(b_i - b) \\ &:= \boldsymbol{\beta}_{0s} + \boldsymbol{\beta}_{1s}(a_i - a) + \boldsymbol{\beta}_{2s}(b_i - b). \end{aligned} \quad (3.2.3)$$

where $\boldsymbol{\beta}^{(a)}(s) = [\frac{\partial \beta_1}{\partial a}, \ldots, \frac{\partial \beta_p}{\partial a}]$ and $\boldsymbol{\beta}^{(b)}(s) = [\frac{\partial \beta_1}{\partial b}, \ldots, \frac{\partial \beta_p}{\partial b}]$. The Taylor approximation is a particularly useful formula in the nonparametric estimation of spatial models, as it relates $\boldsymbol{\beta}$ at arbitrary locations $\mathbf{s}$ to $\boldsymbol{\beta}$ in the neighbourhood locations where we sample data.

Substitute the Taylor approximation 3.2.3 into the model 3.2.1 and we obtain

$$Y_i \approx \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{0s} + \mathbf{x}_i^{\mathrm{T}}(a_i - a)\boldsymbol{\beta}_{1s} + \mathbf{x}_i^{\mathrm{T}}(b_i - b)\boldsymbol{\beta}_{2s} + \epsilon_i + \eta_i, \quad \mathbf{s}_i \in s(h). \quad (3.2.4)$$

[1]The requirement for known covariance $\Sigma_\eta(\mathbf{s}_i, \mathbf{s}_j)$ is necessary for parameter identification but is not necessary if repeated measurements are available at each location s. Please refer to our application example for further details.

In doing so, we transform the original regression with the heterogeneous coefficient for each observation into a regression with the same coefficient value $\boldsymbol{\beta}_{0s}$, after adjustment for the first-order derivatives. Note that the Taylor approximation works well if $\mathbf{s}_i$ is close to $\mathbf{s}$, and less so if two locations are further apart. Therefore, the importance of each datum to the estimation of $\boldsymbol{\beta}(\mathbf{s})$ shall vary based on their proximity to $\mathbf{s}$. To incorporate this feature, we define a kernel function to assign weight to each observation according to the its distance to $\mathbf{s}$,

$$K_h(d(\mathbf{s}_i, \mathbf{s})) = K(d(\mathbf{s}_i, \mathbf{s})/h) \tag{3.2.5}$$

where $K(\cdot)$ is a kernel function such as the Epanechnikov kernel, defined as $K(u) = 0.75(1 - u^2)\mathbb{1}(|u| \leq 1)$ (see Wasserman, 2006, p. 3).

First Stage Estimation

In the first stage, the covariance of ϵ is unknown, therefore an initial estimator of $\boldsymbol{\beta}_{0s}$ is needed to estimate the error covariance. We minimise the following kernel weighted objective function with respect to $\boldsymbol{\beta}_{0s}, \boldsymbol{\beta}_{1s}, \boldsymbol{\beta}_{2s}$,

$$\sum_{i=1}^{n} \left[Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{0s} - \mathbf{x}_i^{\mathrm{T}}(a_i - a)\boldsymbol{\beta}_{1s} - \mathbf{x}_i^{\mathrm{T}}(b_i - b)\boldsymbol{\beta}_{2s}\right]^2 K_h(d(\mathbf{s}_i, \mathbf{s})), \tag{3.2.6}$$

which leads to a weighted-least-square type estimator written as follows,

$$\hat{\boldsymbol{\beta}}_h(\mathbf{s}) = [\mathbb{I}_p, \mathbf{0}_{p,2p}](\mathbf{G}_{s(h)}^{\mathrm{T}}\mathbf{W}_{s(h)}\mathbf{G}_{s(h)})^{-1}\mathbf{G}_{s(h)}^{\mathrm{T}}\mathbf{W}_{s(h)}\mathbf{Y}_{s(h)} \tag{3.2.7}$$

where $\mathbb{I}_p$ is a $p \times p$ identity matrix, $\mathbf{0}_{p,2p}$ is a $p \times 2p$ matrix of 0; $\mathbf{Y}_{s(h)} = [Y_i]_{\mathbf{s}_i \in s(h)}$, $\mathbf{G}_{s(h)} = [\mathbf{x}_i^{\mathrm{T}}, \mathbf{x}_i^{\mathrm{T}}(a_i - a), \mathbf{x}_i^{\mathrm{T}}(b_i - b)]_{\mathbf{s}_i \in s(h)}$ and $\mathbf{W}_{s(h)} = \mathrm{diag}([K_h(d(\mathbf{s}_i, \mathbf{s}))]_{\mathbf{s}_i \in s(h)})$.

$\hat{\boldsymbol{\beta}}_h(\mathbf{s})$ can be evaluated for each $\mathbf{s} \in D$. Particularly for each $\mathbf{s}_i, i = 1, \ldots, n$, we obtain the residual e_i, defined as $e_i = Y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_h(\mathbf{s}_i)$.

Covariance Estimation

In the first stage estimation, we construct an initial $\boldsymbol{\beta}$ estimator and obtain the residual e_i. Now we focus on the estimation of the covariance of ϵ via the variogram fitting (Opsomer et al., 1999; Zimmerman and Zimmerman, 1991).

Define $r(\mathbf{s}_i) = (\epsilon_i + \eta_i)/\sigma_\epsilon(\mathbf{s}_i)$. We note that, for any location pair $(\mathbf{s}_i, \mathbf{s}_j) \in D$,

$$\begin{aligned}\Gamma(d(\mathbf{s}_i, \mathbf{s}_j)) &:= \mathbb{E}\left(r(\mathbf{s}_i) - r(\mathbf{s}_j)\right)^2 \\ &= 2 - 2\rho_\epsilon(d(\mathbf{s}_i, \mathbf{s}_j)) + \frac{\sigma_\eta^2(\mathbf{s}_i)}{\sigma_\epsilon^2(\mathbf{s}_i)} + \frac{\sigma_\eta^2(\mathbf{s}_j)}{\sigma_\epsilon^2(\mathbf{s}_j)} - \frac{2\Sigma_\eta(\mathbf{s}_i, \mathbf{s}_j)}{\sigma_\epsilon(\mathbf{s}_i)\sigma_\epsilon(\mathbf{s}_j)}.\end{aligned} \tag{3.2.8}$$

Then

$$\rho_\epsilon(d(\mathbf{s}_i,\mathbf{s}_j)) = 1 - \frac{1}{2}\left\{\Gamma\ (d(\mathbf{s}_i,\mathbf{s}_j)) - \frac{\sigma_\eta^2(\mathbf{s}_i)}{\sigma_\epsilon^2(\mathbf{s}_i)} - \frac{\sigma_\eta^2(\mathbf{s}_j)}{\sigma_\epsilon^2(\mathbf{s}_j)} + \frac{2\Sigma_\eta(\mathbf{s}_i,\mathbf{s}_j)}{\sigma_\epsilon(\mathbf{s}_i)\sigma_\epsilon(\mathbf{s}_j)}\right\}. \quad (3.2.9)$$

It follows that $\rho_\epsilon(d(\mathbf{s}_i,\mathbf{s}_j))$ can be estimated as

$$\hat{\rho}_\epsilon(d(\mathbf{s}_i,\mathbf{s}_j) = 1 - \frac{1}{2}\left\{\hat{\Gamma}(d(\mathbf{s}_i,\mathbf{s}_j)) - \frac{\sigma_\eta^2(\mathbf{s}_i)}{\hat{\sigma}_\epsilon^2(\mathbf{s}_i)} - \frac{\sigma_\eta^2(\mathbf{s}_j)}{\hat{\sigma}_\epsilon^2(\mathbf{s}_j)} + \frac{2\Sigma_\eta(\mathbf{s}_i,\mathbf{s}_j)}{\hat{\sigma}_\epsilon(\mathbf{s}_i)\hat{\sigma}_\epsilon(\mathbf{s}_j)}\right\},$$

where $\hat{\sigma}_\epsilon^2(\mathbf{s})$ can be estimated via

$$\hat{\sigma}_\epsilon^2(\mathbf{s}) = \frac{\sum_i K_h(d(\mathbf{s}_i,\mathbf{s}))e_i^2}{\sum_i K_h(d(\mathbf{s}_i,\mathbf{s}))} - \sigma_\eta^2(\mathbf{s}),$$

and $\hat{\Gamma}(d(\mathbf{s}_i,\mathbf{s}_j))$ is estimated via

$$\hat{\Gamma}(d(\mathbf{s}_i,\mathbf{s}_j)) = \frac{\sum_{l,l',l\neq l'} K_h\left([d(\mathbf{s}_l,\mathbf{s}_{l'}) - d(\mathbf{s}_i,\mathbf{s}_j)]\right)(\hat{r}(\mathbf{s}_i) - \hat{r}(\mathbf{s}_j))^2}{\sum_{l,l',l\neq l'} K_h\left([d(\mathbf{s}_l,\mathbf{s}_{l'}) - d(\mathbf{s}_i,\mathbf{s}_j)]\right)}, \quad (3.2.10)$$

where $\hat{r}(\mathbf{s}_i) = e_i/\hat{\sigma}_\epsilon(\mathbf{s}_i)$.

Now, consider $\{d(\mathbf{s}_i,\mathbf{s}_j), \hat{\rho}_\epsilon(d(\mathbf{s}_i,\mathbf{s}_j)), i,j = 1,\ldots,n\}$ as data points and assume the correlation function has a parametric form $\rho_\epsilon(d(\mathbf{s}_i,\mathbf{s}_j);\boldsymbol{\theta})$ up to a finite number of parameters $\boldsymbol{\theta} := (\theta_1,\theta_2,\ldots,\theta_r)$. $\boldsymbol{\theta}$ can be estimated by minimising the following objective function,

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i\neq j,i,j=1,\ldots,n} \left[\hat{\rho}_\epsilon(d(\mathbf{s}_i,\mathbf{s}_j)) - \rho_\epsilon(d(\mathbf{s}_i,\mathbf{s}_j);\boldsymbol{\theta})\right]^2. \quad (3.2.11)$$

For an arbitrary location pair $(\mathbf{s}_l,\mathbf{s}_m) \in D$, the covariance $\Sigma_\epsilon(\mathbf{s}_l,\mathbf{s}_m)$ is then estimated as $\hat{\Sigma}_\epsilon(\mathbf{s}_l,\mathbf{s}_m) = \hat{\sigma}_\epsilon(\mathbf{s}_l)\hat{\sigma}_\epsilon(\mathbf{s}_m)\rho_\epsilon(d(\mathbf{s}_l,\mathbf{s}_m);\hat{\boldsymbol{\theta}})$.

Second Stage Estimation

With the estimated error covariance, we move to the second stage estimation, which incorporates the covariance structure into the $\boldsymbol{\beta}$ estimator.

Let $\hat{\boldsymbol{H}}_{s(h)}$ be the estimated covariance matrix of $\epsilon+\eta$ with element (i,j) being $\hat{\Sigma}(\mathbf{s}_i,\mathbf{s}_j) = \hat{\Sigma}_\epsilon(\mathbf{s}_i,\mathbf{s}_j) + \Sigma_\eta(\mathbf{s}_i,\mathbf{s}_j)$ for $\mathbf{s}_i,\mathbf{s}_j$ in $s(h)$. We also define $\hat{\boldsymbol{H}}_{s(h)}^{-1/2}$ such that $(\hat{\boldsymbol{H}}_{s(h)}^{-1/2})^{\mathrm{T}}\hat{\boldsymbol{H}}_{s(h)}^{-1/2} = \hat{\boldsymbol{H}}_{s(h)}^{-1}$ and $\mathbf{W}_{s(h)}^{1/2} = diag([K_h^{1/2}(d(\mathbf{s}_i,\mathbf{s}))]_{\mathbf{s}_i\in s(h)})$.

Note that the local regression 3.2.4 can be written in a matrix format,

$$\mathbf{Y}_{s(h)} = \mathbf{G}_{s(h)}\boldsymbol{\beta}_s + \boldsymbol{\epsilon}_{s(h)} + \boldsymbol{\eta}_{s(h)} \quad (3.2.12)$$

where $\boldsymbol{\beta}_s = [\boldsymbol{\beta}_{0s}^{\mathrm{T}},\boldsymbol{\beta}_{1s}^{\mathrm{T}},\boldsymbol{\beta}_{2s}^{\mathrm{T}}]^{\mathrm{T}}$, $\boldsymbol{\epsilon}_{s(s)} = [\epsilon_i]_{i\in s(h)}^{\mathrm{T}}$ and $\boldsymbol{\eta}_{s(s)} = [\eta_i]_{i\in s(h)}^{\mathrm{T}}$. Left multiply both sides by $\hat{\boldsymbol{H}}_{s(h)}^{-1/2}\mathbf{W}_{s(h)}^{1/2}$

$$\hat{\boldsymbol{H}}_{s(h)}^{-1/2}\mathbf{W}_{s(h)}^{1/2}\mathbf{Y}_{s(h)} = \hat{\boldsymbol{H}}_{s(h)}^{-1/2}\mathbf{W}_{s(h)}^{1/2}\mathbf{G}_{s(h)}\boldsymbol{\beta}_s + \hat{\boldsymbol{H}}_{s(h)}^{-1/2}\mathbf{W}_{s(h)}^{1/2}\boldsymbol{\epsilon}_{s(h)} + \hat{\boldsymbol{H}}_{s(h)}^{-1/2}\mathbf{W}_{s(h)}^{1/2}\boldsymbol{\eta}_{s(h)}. \tag{3.2.13}$$

Then, minimise the following objective function with respect to $\boldsymbol{\beta}_s$,

$$(\hat{\boldsymbol{H}}_{s(h)}^{-1/2}\mathbf{W}_{s(h)}^{1/2}\mathbf{Y}_{s(h)} - \hat{\boldsymbol{H}}_{s(h)}^{-1/2}\mathbf{W}_{s(h)}^{1/2}\mathbf{G}_{s(h)}\boldsymbol{\beta}_s)^{\mathrm{T}}(\hat{\boldsymbol{H}}_{s(h)}^{-1/2}\mathbf{W}_{s(h)}^{1/2}\mathbf{Y}_{s(h)} - \hat{\boldsymbol{H}}_{s(h)}^{-1/2}\mathbf{W}_{s(h)}^{1/2}\mathbf{G}_{s(h)}\boldsymbol{\beta}_s) \tag{3.2.14}$$

which gives the following second state generalised-least-square type estimator,

$$\tilde{\boldsymbol{\beta}}_h(s) = [\mathbb{I}_p, \mathbf{0}_{p,2p}](\mathbf{G}_{s(h)}^{\mathrm{T}}\mathbf{W}_{s(h)}^{1/2}\hat{\boldsymbol{H}}_{s(h)}^{-1}\mathbf{W}_{s(h)}^{1/2}\mathbf{G}_{s(h)})^{-1}\mathbf{G}_{s(h)}^{\mathrm{T}}\mathbf{W}_{s(h)}^{1/2}\hat{\boldsymbol{H}}_{s(h)}^{-1}\mathbf{W}_{s(h)}^{1/2}\mathbf{Y}_{s(h)}. \tag{3.2.15}$$

Compared with the first stage estimator, the second stage estimator incorporates the error covariance matrix along with the kernel weight matrix. It is worth noting that because we estimate $\boldsymbol{\beta}(s)$ using only local data $s(h)$ controlled by bandwidth h, implementation of the estimator does not require inverting an $n \times n$ covariance matrix, but rather depends on local sample size $n(s(h))$.

3.2.2 Spatially Non-Varying Coefficients

Now we consider an important extension of the model 3.2.1 to include coefficients that are non-varying. The model is specified as follows,

$$Y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}(\mathbf{s}_i) + \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\gamma} + \epsilon_i + \eta_i, \quad i = 1, \ldots, n, \tag{3.2.16}$$

where $\mathbf{z}_i$ is a vector of q additional explanatory variables associated with non-varying coefficient $\boldsymbol{\gamma}$.

To estimate $\boldsymbol{\gamma}$, we treat the coefficients as if they are spatially varying and estimate $\boldsymbol{\gamma}(\mathbf{s})$ using the methodology described above. The varying estimates are then aggregated to form non-varying estimates,

$$\hat{\boldsymbol{\gamma}}_h = \frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\gamma}}_h(\mathbf{s}_i). \tag{3.2.17}$$

Alternatively, we could apply the 'profile method' to estimate the non-varying coefficient (Fan and Tao, 2005). However, the approach would require inverting an $n \times n$ covariance matrix, which is not suitable in our context.

3.2.3 Spatial Prediction

In this section we consider the spatial prediction problem. Let $\mathbf{s}_0$ be a generic location on D, where $\boldsymbol{x}_0$ is observed. We are interested in predicting Y_0 at $\mathbf{s}_0$. This requires predicting $\boldsymbol{\beta}(\mathbf{s})$ at $\mathbf{s}_0$ and the error $\hat{e}_0$ at $\mathbf{s}_0$,

$$\hat{Y}_0 = \boldsymbol{x}_0^{\mathrm{T}} \tilde{\boldsymbol{\beta}}_h(\mathbf{s}_0) + \hat{e}_0 \tag{3.2.18}$$

While $\tilde{\boldsymbol{\beta}}(\mathbf{s}_0)$ can be evaluated directly, estimating $\hat{e}_0$ involves inverting an $n \times n$ covariance matrix (Cressie and Johannesson, 2008), which might be computationally intensive. To deal with the problem, we consider spatial prediction using local information. Specifically, let $s_0(h)$ be the the collection of sample locations within the distance of h to $\mathbf{s}_0$.

We define the local estimate of $\hat{e}_0$ as follows

$$\hat{e}_0 = \hat{\boldsymbol{H}}_{s_0(h)}^{\mathrm{T}}(s_0) \mathbf{W}_{s_0(h)}^{1/2} \hat{\boldsymbol{H}}_{s_0(h)}^{-1} \mathbf{W}_{s_0(h)}^{1/2} \boldsymbol{e}_{s_0(h)}, \tag{3.2.19}$$

where $\hat{\boldsymbol{H}}_{s_0(h)}(s_0) = [\hat{\Sigma}(\mathbf{s}_0, \mathbf{s}_i)]_{i \in s_0(h)}$, $\hat{\boldsymbol{H}}_{s_0(h)}$ is a $n(s_0(h)) \times n(s_0(h))$ with (i, j)th element being $\hat{\Sigma}(\mathbf{s}_i, \mathbf{s}_j)$, $i, j \in s_0(h)$, and $\mathbf{W}_{s_0(h)}^{1/2} = \mathrm{diag}([K_h^{1/2}(d(\mathbf{s}_0, \mathbf{s}_i)]_{\mathbf{s}_i \in s_0(h)})$.

3.2.4 Bandwidth Selection

In order to implement the estimator, the size of h should be determined. A relatively small h will lead to a smaller neighbourhood in the estimation of $\boldsymbol{\beta}$, therefore less approximation bias, but larger variance in $\hat{\boldsymbol{\beta}}$. A relatively large bandwidth, on the other hand, includes data farther away, hence increase the approximation bias. As more data are involved, the variance for each estimate is expected to be smaller. An appropriate bandwidth, therefore, strikes a balance between the bias and variance of the estimators. In this chapter, we use the cross-validation method to choose a bandwidth.

Let $\hat{Y}^{-(i)}$ be the prediction of Y_i without the i^{th} observation. Then we pick the bandwidth h by minimising the leave-one-out mean square error.

$$\mathrm{CV}(h) := \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{Y}^{-(i)} \right)^2 \tag{3.2.20}$$

Other cross-validation methods are also available. For example, instead of leaving one observation out, we could leave out k data points, k being a positive integer. Picking k out of N has potentially $\frac{N!}{k!(N-k)!}$ different combinations, which is not feasible when the sample size is large. Alternatively, non-exhaustive cross-validation methods could be applied, such as $k-$fold. The $k-$fold cross-validation randomly divides data into k groups. Each time, one group is reserved for prediction, whereas $k-1$ groups are used for estimation. The process is iterated for k times, and CV score is calculated by summing up k groups of prediction errors. The approach is computationally more appealing as it requires only k (e.g., 10) iterations.

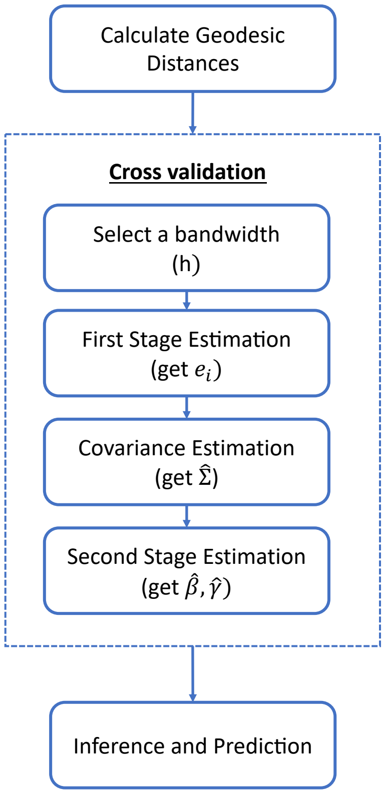

Figure 3.2: SVCR methodology road map

3.3 Monte Carlo Study

To study the finite-sample performance of our proposed methodology and compare it with other existing relatively simpler methods, we conduct a simulation study. We generate the simulated data from the following model

$$y_i = c + x_i\beta(\mathbf{s}_i) + \epsilon_i + \eta_i, \quad i = 1, \ldots, N. \tag{3.3.21}$$

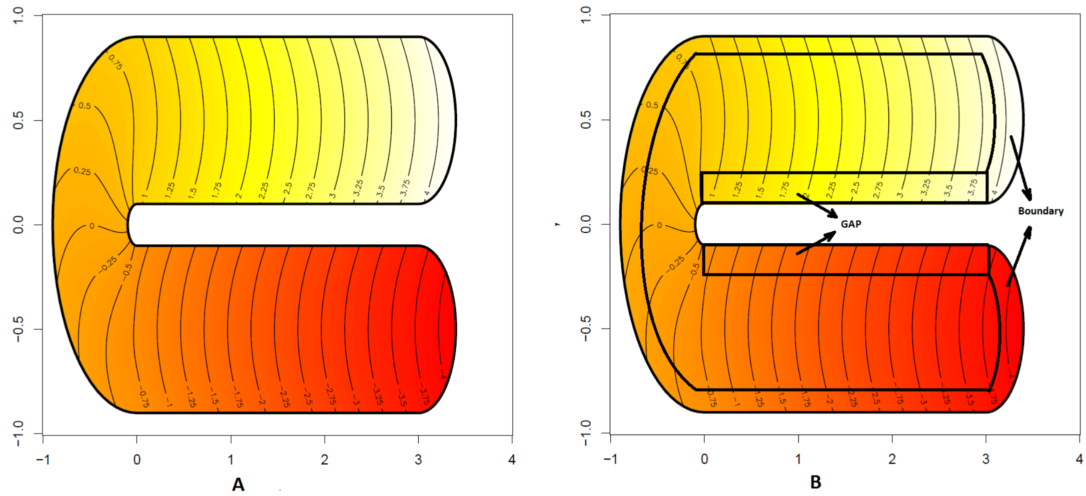

Figure 3.3: The distribution of coefficient in the horse-shoe shaped domain

We take $c = 5$, $x_i \sim N(3, 2^2)$ and $\beta(\cdot)$ is varying like Figure 3.3; ϵ_i is spatially distributed with correlation function $\rho(\epsilon_i, \epsilon_j) = \exp(-2d_g(\mathbf{s}_i, \mathbf{s}_j))$, $d_g(\mathbf{s}_i, \mathbf{s}_j)$ is the shortest path in the doamin between location $\mathbf{s}_i$ and $\mathbf{s}_j$; $\eta_i \sim N(0, 0.5^2)$. To test the robustness of our method against potential non-Gaussianity, we also consider a case where ϵ and η follow multivariate t distribution with the same covariance structure. The $\beta(\cdot)$ surface is a classic horseshoe-shaped test function originally proposed by Ramsay (2002) and later studied by Wang and Ranalli (2007); Wood et al. (2008). The values are smoothly defined over the horseshoe-shaped region, but the region itself is not regular. Specifically, it has a gap between the two arms, across which function values differ significantly. In this design, Euclidean distance across the gap does not indicate similarity.

Under the above design, we compare our proposed estimator against alternatives that differ in how the varying coefficients and non-varying coefficients are estimated, which distance measures are used and, whether or not a spatially weighted (by the inverse spatial covariance matrix) estimator is used.

For the varying coefficient, we consider two options: the covariance weighted local linear, the proposed approach, and the commonly used geographically weighted regression (GWR) (e.g., Lloyd, 2010; Bivand and Yu, 2015). For the non-varying part, both the profile method and the average-over-local-estimates method are considered. For the choice of the distance measure, both geodesic distance and Euclidean distance are examined. With the proposed estimator, the regression function is estimated with a weighting of the inverse spatial covariance matrix not shared by other approaches, including Zhu et al. (2014). The rows in each block of Tables 3.1 and 3.2 represent the Monte-Carlo results from various combination of estimation methods. The bandwidths in all methodologies are selected via 10-fold cross-validation.

We choose various sample sizes, $N = 50, 100, 200, 300$, each with 1000 repetitions. To reduce computational burden in this simulation study, βs are randomly sampled from the horseshoe-shaped region. To test how different methodologies perform at different parts of the domain, we divide the horse-shoe shape into three parts: boundary, gap and inner zone (see Figure 3.3). The boundary zone refers to the area close to the boundary of the shape. Data are not sample outside the boundary. The gap zone refers to the area where data from two sides of the tube are close in the Euclidean sense, but not close in terms of shortest path in the domain. The inner zone lies in the middle of the domain, away from boundary of the shape. Both β and c are estimated by different models mentioned above, and the performance is measured using the following quantities,

1. Bias and RMSE (root mean sum of squares) for the non-varying coefficient,

$$B(\hat{c}) = \frac{1}{Rep}\sum_{r=1}^{Rep}(\hat{c}_r - c) \text{ and } RMSE(\hat{c}) = \sqrt{\frac{1}{Rep}\sum_{r=1}^{Rep}(\hat{c}_r - c)^2}$$

2. Mean summed squared error (MSSE) at sampled locations for the varying coefficient,

$$MSSE(\beta) = \frac{1}{Rep \cdot N}\sum_{r=1}^{Rep}\sum_{i=1}^{N}\left[\hat{\beta}_r(\mathbf{s}_i) - \beta_r(\mathbf{s}_i)\right]^2$$

3. MSSE for the varying coefficient at boundary, gap and inner zones specified in Figure 3.3,

$$MSSE_b(\beta) = \frac{1}{Rep \cdot N}\sum_{r=1}^{Rep}\sum_{i=1}^{N}\left[\hat{\beta}_r(\mathbf{s}_i) - \beta(\mathbf{s}_i)\right]^2 \mathbb{I}(\mathbf{s}_i \in \text{Boundary Zone})$$

$$MSSE_g(\beta) = \frac{1}{Rep \cdot N}\sum_{r=1}^{Rep}\sum_{i=1}^{N}\left[\hat{\beta}_r(\mathbf{s}_i) - \beta(\mathbf{s}_i)\right]^2 \mathbb{I}(\mathbf{s}_i \in \text{Gap Zone})$$

$$MSSE_g(\beta) = \frac{1}{Rep \cdot N}\sum_{r=1}^{Rep}\sum_{i=1}^{N}\left[\hat{\beta}_r(\mathbf{s}_i) - \beta(\mathbf{s}_i)\right]^2 \mathbb{I}(\mathbf{s}_i \in \text{Inner Zone})$$

4. Estimated coverage probability for a 95% nominal level,

$$\text{Coverage Probability} = \frac{1}{N \cdot Rep} \sum_{r=1}^{Rep} \sum_{i=1}^{N} \mathbb{I}(Y_i \in 95\%\text{prediction interval})$$

The Monte Carlo output is presented in Tables 3.1 and 3.2. The simulation outcomes clearly indicate the superiority of the proposed procedure. The study also ensures that there is no loss in accuracy for our local estimates of the varying coefficients or the fixed coefficients. The maximum gain comes from the gap zone. Compared with the estimator which does not consider spatial dependence (third row and above), covariance weighted estimator have achieved higher efficiency, particularly for a large sample. For the varying coefficient β, our estimator has the best performance when the sample size is not too small. Both the bias and RMSE for all parameters reduce for increasing sample sizes, confirming the consistency of the estimator in the asymptotic sense. In terms of prediction and prediction intervals, the proposed method performs fairly well. There is slight under coverage, but the gap reduces with sample increasing sample size. When the true distribution is not normal, the proposed method enjoys some robustness, although not as good as in the case of normality.

Ceteris paribus, we find that (i) local linear estimators outperform local constant (GWR) estimators; (ii) estimators using geodesic distance have better performance compared with Euclidean counterparts; (iii) the non-varying coefficient obtained by averaging local estimates also performs better than the profiled method.

Sample Size	Error Distribution	Varying Term Estimation	Distance Measure	Constant Term Estimation	Covariance Weighting	β $MSSE$	β $MSSE_g$	β $MSSE_b$	Intercept BIAS	Intercept RMSE	Coverage (95% nominal)
$n = 50$	Normal	Local Linear	Geodesic	Average	Yes	1.1207	1.5576*	1.6604	-0.0065	0.8165*	91.65%
	Multivariate t	Local Linear	Geodesic	Average	Yes	1.1108	1.5755	1.5431	-0.002*	0.8612	86.16%
	Normal	Local Linear	Geodesic	Average	No	1.0475*	1.6380	1.2589	0.0219	0.9060	84.23%
	Normal	Local Linear	Euclidean	Average	No	1.0823	1.7028	1.2618	0.0233	0.9209	-
	Normal	Local Linear	Geodesic	Profile	No	1.1452	1.7154	1.2772	-0.1408	1.4607	-
	Normal	Local Linear	Euclidean	Profile	No	1.1747	1.7819	1.2737	-0.1418	1.4542	-
	Normal	GWR	Geodesic	Average	No	1.0852	1.7829	1.0911*	0.0272	0.9713	-
	Normal	GWR	Euclidean	Average	No	1.1621	1.8870	1.1214	0.0213	1.0284	-
	Normal	GWR	Geodesic	Profile	No	1.0936	1.7376	1.1195	-0.1319	1.2895	-
	Normal	GWR	Euclidean	Profile	No	1.1504	1.8295	1.1367	-0.1307	1.2682	-
$n = 100$	Normal	Local Linear	Geodesic	Average	Yes	0.7874*	1.4137	0.8914	-0.0478	0.4486*	92.13%
	Multivariate t	Local Linear	Geodesic	Average	Yes	0.8834	1.2583*	1.3024	0.0065*	0.7396	89.69%
	Normal	Local Linear	Geodesic	Average	No	0.8334	1.5050	0.7966*	-0.0187	0.5117	86.34%
	Normal	Local Linear	Euclidean	Average	No	0.9161	1.6642	0.8238	-0.0164	0.5319	-
	Normal	Local Linear	Geodesic	Profile	No	0.9088	1.5494	0.8682	-0.1200	1.0534	-
	Normal	Local Linear	Euclidean	Profile	No	0.9843	1.7095	0.8900	-0.1191	1.0480	-
	Normal	GWR	Geodesic	Average	No	0.8661	1.5021	0.8868	-0.0192	0.5650	-
	Normal	GWR	Euclidean	Average	No	0.9533	1.6546	0.9435	-0.0180	0.6035	-
	Normal	GWR	Geodesic	Profile	No	0.9080	1.5000	0.9319	-0.1298	0.9783	-
	Normal	GWR	Euclidean	Profile	No	0.9815	1.6437	0.9799	-0.1282	0.9604	-

Note: * indicates the smallest absolute value in each category for a given n.

Table 3.1: SVCR simulation results ($n = 50, 100$)

Sample Size	Varying Term Estimation	Distance Measure	Constant Term Estimation	Covariance Weighting	β $MSSE$	β $MSSE_g$	β $MSSE_b$	Intercept BIAS	Intercept RMSE	Coverage 95% nominal
$n = 200$	Local Linear	Geodesic	Average	Yes	0.6270*	1.1546	0.7216	-0.0119	0.3621*	96.24%
	Local Linear	Geodesic	Average	Yes	0.6643	1.0690*	0.8138	0.0177	0.4663	91.45%
	Local Linear	Geodesic	Average	No	0.6906	1.2226	0.6203*	-0.0047*	0.4171	87.35%
	Local Linear	Euclidean	Average	No	0.8091	1.5108	0.6929	-0.0050	0.4279	-
	Local Linear	Geodesic	Profile	No	0.7460	1.2598	0.6631	-0.1211	0.7677	-
	Local Linear	Euclidean	Profile	No	0.8585	1.5500	0.7311	-0.1229	0.7622	-
	GWR	Geodesic	Average	No	0.7185	1.1547	0.7299	-0.0060	0.4298	-
	GWR	Euclidean	Average	No	0.8119	1.4003	0.7926	-0.0084	0.4522	-
	GWR	Geodesic	Profile	No	0.7749	1.1814	0.7893	-0.1352	0.7389	-
	GWR	Euclidean	Profile	No	0.8588	1.4180	0.8460	-0.1366	0.7278	-
$n = 300$	Local Linear	Geodesic	Average	Yes	0.5667*	0.9318	0.6794	-0.0177	0.3394*	93.69%
	Local Linear	Geodesic	Average	Yes	0.6570	0.8884*	0.8091	-0.0141	0.4267	92.12%
	Local Linear	Geodesic	Average	No	0.6295	0.9688	0.5728*	-0.0030	0.3980	89.25%
	Local Linear	Euclidean	Average	No	0.7444	1.3447	0.6321	-0.0024*	0.4033	-
	Local Linear	Geodesic	Profile	No	0.6856	1.0083	0.6166	-0.1390	0.6697	-
	Local Linear	Euclidean	Profile	No	0.7955	1.3827	0.6724	-0.1398	0.6648	-
	GWR	Geodesic	Average	No	0.6628	0.9052	0.7033	-0.0015	0.4027	-
	GWR	Euclidean	Average	No	0.7430	1.1752	0.7500	-0.0005	0.4120	-
	GWR	Geodesic	Profile	No	0.7268	0.9516	0.7679	-0.1548	0.6543	-
	GWR	Euclidean	Profile	No	0.8008	1.2081	0.8119	-0.1539	0.6459	-

Note: * indicates the smallest absolute value in each category for a given n.

Table 3.2: SVCR simulation results ($n = 200, 300$)

3.4 Application

In this section, we apply the proposed methodology to a specific urban housing market in Portugal. The analysed area is located in the Centro Region of Portugal and includes two municipalities – Aveiro and Ílhavo (Figure 3.1). The municipality of Aveiro has a total area of $200km^2$, and Ílhavo an area of 75km; the total population in 2011 were 78,454 and 38,317 inhabitants respectively (Source: 2011 Census). Omitting the area of the lagoon (where a minimal number of people live on boats), the population density is 600 inhabitants per square kilometre, which is typical for an urban agglomeration in western Europe.

Following the early works of Lancaster (1966) and Rosen (1974), hedonic pricing models are widely used to decompose housing values. A general form of a hedonic pricing model can be represented as follows,

$$P = f(S, N, L, C, T) \tag{3.4.22}$$

where P denotes the value of a house (price), S, N, L, C, T denote, respectively, structural characteristics, neighbourhood characteristics, location within the market, other characteristics and the time when the value is recorded. f is a function that links these characteristics to the value. Its partial effect with respect to one variable represents people's willingness to pay an additional unit of that hedonic feature.

Three issues surround the estimation of a hedonic model. The first is concerned with the choice of $f(\cdot)$, which is not very precisely informed by theories such as Lancaster (1966) or Rosen (1974). A common practice is to adopt a simple parametric form, such as the semi-log form where the dependent variable takes a logarithm transformation, or the log-log form where both the dependent and independent variables take logarithm transformations. The log-log form also offers a simple interpretation for the coefficients in terms of price elasticities. Follain and Malpezzi (1980) tested the log-linear as well as the linear specifications. They found that the former enjoys some advantages such as the ability to capture nonlinear relations and the convenience to interpret the coefficients as elasticities. Moreover, the nonlinear hedonic function makes it possible to recover the parameters of structural demand curve (Malpezzi, 2002). In this chapter, we consider a hedonic house price regression model where the dependent variable is the logarithm of house prices per square meter and a collection of housing features are regressors. Our specific interest lies in the house price elasticity of living space, which is allowed to vary over the spatial domain.

The second issue involves the selection of relevant regressors. Due to the complex and heterogeneous nature of a house, literally hundreds of potential regressors could be included in a hedonic regression model. To reduce the dimension, factor analysis has been actively used in the literature; see Wilkinson and Archer (1973) for the pioneering work in this area and Malpezzi (2002) for an excellent review. Factor analysis can aggregate numerous housing characteristics into a small number of interpretable categories, and therefore greatly reduces the dimension and simplifies further analysis. For this particular housing market, Bhattacharjee et al. (2012) combine factor analysis with regression and obtained a parsimonious model. They extract five leading orthogonal factors from a large number of characteristics and then treat the factors as the regressors in the estimation of a hedonic model.

Third, the hedonic relationship is expected to vary over the spatial domain. Therefore, in practice, a hedonic regression is not applied directly to a whole housing market but separately to several smaller homogeneous areas, often called sub-markets. If sub-markets are not specified properly, estimates of heterogeneous coefficients would be misleading. Recently, it has been argued that the heterogeneous coefficients should be estimated prior to the segmentation of sub-markets, and the delineation itself should be 'data-driven'–building upon the similarity of hedonic prices and characteristics, rather than the other way around (Bhattacharjee et al., 2016a).

To understand the spatial heterogeneity in the hedonic regression, we combine the multi-factor structure (Bhattacharjee et al., 2012) and the varying coefficient component into a single regression model,

$$\ln P_{i,j} = \beta(\mathbf{s}_i) \ln M_{i,j} + \eta \ln T_{i,j} + \gamma_0 + \sum_{k=1}^{5} \gamma_k F_{ki,j} + \varepsilon_i + \eta_{i,j}, \quad j = 1, \ldots, n_i \quad i = 1, \ldots, N \tag{3.4.23}$$

where $\ln P_{i,j}$ is the logarithm of house price per square meter of living area (P/S^2); $\ln M_{i,j}$ denotes the logarithm of square meters of living area (S^2); $\ln T_{i,j}$ denotes the logarithm of time on the markets (in days); $F_{ki,j}, k = 1, \ldots, 5$ are five orthogonal factors which can be interpreted as access to the centre amenities, access to local amenities, access to beaches, physical attributes such as house type, number of rooms and additional house facilities (balcony, garage, etc.). The double subscript (i, j) is used as each location (i) might be shared by more than one data point (j), in which case there are replications. Here i denotes a distinct location and j indexes a specific house (observation) at that location. N is the total number of distinct locations and n_j is the number of houses at location i.

We assume a random effect structure for the error terms similar to that in the panel data setting (Baltagi, 2008). ε_i is the location-specific random effect, with $\mathbb{E}(\varepsilon_i) = 0$, $var(\varepsilon_i) = \sigma_\varepsilon^2(\mathbf{s}_i)$, and $\rho_\varepsilon(\mathbf{s}_i, \mathbf{s}_j) = \rho_\varepsilon(d(\mathbf{s}_i, \mathbf{s}_j); \boldsymbol{\theta})$. η_{ij} is the idiosyncratic error such that $\mathbb{E}(\eta_i) = 0$, $var(\eta_{i,j}) = \sigma_\eta^2(\mathbf{s}_i)$, $\rho_\eta(\mathbf{s}_i, \mathbf{s}_j) = 0$ for all $\mathbf{s}_i \neq \mathbf{s}_j$. This implies spatial autocorrelation is captured only through ε.

3.4.1 Data

The database used for this empirical work is provided by the firm Janela Digital S.A., which owns and manages the real estate portal CASA SAPO – the largest site of real estate advertisement in Portugal. For a detailed discussion with the database, see the online supplement and the data archive of Bhattacharjee et al. (2016a). The database covers the time interval between October 2000 and March 2010 and includes around 4 million properties available for transactions in Portugal. For the municipalities of Aveiro and Ílhavo, there are 47,188 properties recorded. This empirical work uses $12,467$ observations, after cleaning the data and removing all cases where data were incomplete.[2]

Descriptive statistics are reported in Bhattacharjee et al. (2012). For most of the properties, the coordinates are not precisely observed. Rather, each property is assigned to a relatively homogeneous area, which we call a zone, whose centroid is geo-referenced. We use the zone-centroid coordinates as the measurement of house locations. The entire territory is divided into 76 such zones (Figure 3.4), each of which contains at least two data points.

3.4.2 Complex Domain and Distance Measure

The spatial domain under study is a coastal region with the Atlantic Ocean flanking the western boundary of the area. The north of the region is dominated by the Aveiro lagoon (Ria de Aveiro) which has no houses and is therefore excluded from our analysis. Two waterways extend from the lagoon southwards (Figure 3.4). The western branch is a river (also called Ria de Aveiro) that flows southwards and westwards into the Atlantic. The eastern waterway is also a river (Rio Boco) extending southwards and inland. The central business district (CBD) of Aveiro lies on the bank of the lagoon slightly eastwards from Rio Boco. The CBD of the twin municipality of Ílhavo lies southwards on the eastern bank of Rio Boco. These central urban areas are surrounded by suburban housing combining traditional and newly built areas; see Bhattacharjee et al. (2012) for a detailed

[2]For a detailed discussion of data cleaning, see Marques (2012).

Figure 3.4: Complex domain of Aveiro housing market, marking the centroids of the 76 zones

description of the suburban areas. Beyond these, there are rural areas with mixed agricultural and (small scale) industrial land use and sparse traditional housing.

Unlike Rio Boco, which is traversed by many bridges providing continuous connection between its two banks, there are only limited connections across Ria de Aveiro. Thus, the domain includes peninsulas, interior holes, gaps and concavities along the boundaries. To the west, the lagoon and western branch of water trisect the spatial domain into three parts – southern beach (Barra, Costa Nova) and northern beach (Sao Jacinto) and a peninsula (Gafanha da Nazaré) which is the centre of the provincial town Gafanhas. Connections between beach areas and the mainland are limited. The southern beach is connected via a bridge from Gafanhas across the river which across the peninsula. The northern beach is connected to the mainland by a ferry, from Gafanhas as well.

When, like the study region, a spatial domain is fragmented with only limited connections between its parts, the Euclidean metric does not always measure the true proximity between spatial points.[3] Consider two sites, Costa Nova and Gafanha da Encarnacao situated on opposite banks of the western river (Ria de Aveiro). These two locations are quite close in the Euclidean sense – just across the river. However, in order to travel from one to the other, a detour is needed northwards along the river, across the bridge, and southwards again; this is because there is no direct connection between these two places.

Intuitively, a suitable distance measure should be the one that is based on the shortest path bypassing inadmissible areas. The recent spatial smoothing literature provides several alternative distance measures in this context. One option is developed by Feng et al. (2012). Their approach relies on *a priori* known boundaries of regions, or at least the adjacency matrix. Distances are computed by either the minimal number of boundaries to cross from one region to another, or the Euclidean distance between centroid coordinates of two regions. In our application, this approach is not feasible because we have point data and an adjacency matrix is not known *a priori*. Furthermore, it cannot be used to deal with complicated and irregular domain problem since the Euclidean distance between regions can cut across boundary features.

Here we adopt the idea of the geodesic distance (Wang and Ranalli, 2007) to measure the intrinsic proximity (similarity/ dissimilarity) between two locations. To compute geodesic distances, we first treat all the distinct locations as a set of nodes and impose a weighted graph structure, say G, on them. All the locations are fully connected, which means every pair of nodes (locations) has an edge between them. Edge weights are measured by pairwise Euclidean distances; the weight matrix is denoted by $D = (d_{i,j})$ for $i, j = 1, \ldots, n$, whose (i, j)-th entry represents the weight between location i and j. Next, we consider a restricted graph of G, called G_k, where each node is only connected to its k nearest neighbours. We denote G_k and D_k the restricted graph and restricted weight matrix, respectively. The (i, j)-th entry of D_k matrix is equal to the same entry in D if there is an edge between them and infinity otherwise. Wang and Ranalli (2007) show that with a properly chosen k, the pairwise shortest path[4] in the restricted graph G_k can serve as a suitable distance metric in the irregular and complicated

[3]For another example, see the discussion on the spatial domain of the island of Montreal in Ramsay (2002).

[4]Computation involves the use of Floyd's algorithm (Floyd, 1962).

domain. The method, however, does not address the problem of holes and gaps directly. When a location's k nearest neighbours include those lying at the opposite side of the gap, the geodesic distance will be wrongly calculated. Choosing a small k could reduce potential bias, but the restricted graph might not be connected. The problem is conspicuous in the Aveiro housing market because the problematic region (across the river) is sparsely sampled.

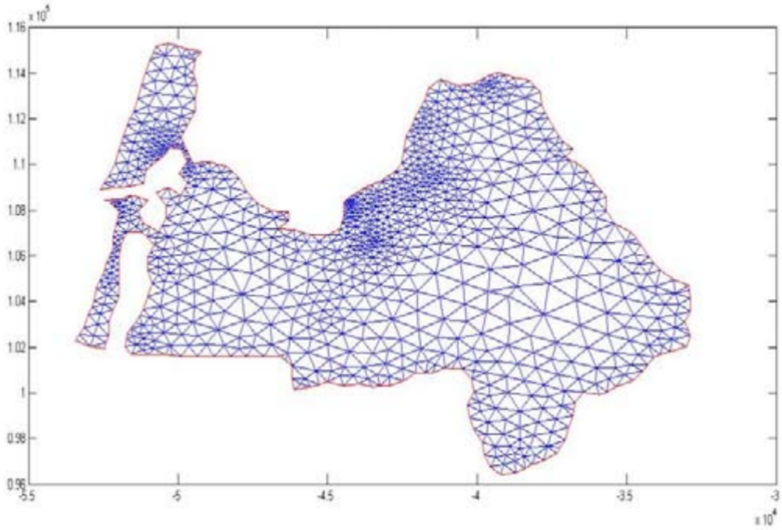

Figure 3.5: Delaunay triangulation of Aveiro housing market

To address this issue, we apply Delaunay triangulation to the set of data locations within the domain (Ramsay, 2002; Sangalli et al., 2013). Delaunay triangulation is a method that approximate the original domain with the union of disjoint triangles, where no point is inside the circumcircle of any triangle (Hjelle and Dæhlen, 2006). Figure 3.5 shows the result of the Delaunay triangulation of the Aveiro housing market. It is clear that triangulation delineates both the boundary of the domain and inadmissible areas. As expected, triangulation in the city centre is finer because of more available data points, whereas in the periphery, triangulation is less dense. Two connections from the mainland to the western beach areas are clearly visible. With the triangulated map, we then treat the vertices of triangles as the nodes of a graph and apply Wang and Ranalli (2007)'s method. The pairwise geodesic distances of 76 regions are thus obtained.

The distance between two data points sharing the same location also needs to be appropriately addressed. We assume a house is randomly located within its zone. If this zone were a circle, then the expected distance between two randomly scattered points is $\frac{128}{45\pi}r\sum \approx 0.91r$, where r is the radius of the circle. Although in our application, we may view each zone as a polygon rather than a circle, the above rule is approximately valid. We let r denote the Euclidean distance from the centroid of a zone to the nearest border and use this to measure the distance of points sharing the same location. We also verify that our estimates are not sensitive to this modelling assumption.

3.4.3 Estimates and Discussion

With a proper geodesic distance defined on the domain, we proceed to estimate the varying coefficient hedonic regression model (3.4.23) based on the proposed methodology. Figure 3.7 shows the variogram and estimated variogram function. For the parametric form of the correlation function, we choose a mixture exponential function $\rho(\mathbf{s}_i, \mathbf{s}_j; \theta_1, \theta_2, \theta_3) = \theta_3 e^{-\theta_1 d(\mathbf{s}_i, \mathbf{s}_j)} + (1 - \theta_3) e^{-\theta_2 d(\mathbf{s}_i, \mathbf{s}_j)}$, with $\theta_1, \theta_2 > 0$ and $0 \leq \theta_3 \leq 1$. As mentioned in Opsomer et al. (1999), the mixture exponential function is generally much more flexible compared with alternatives and guarantees the positive definiteness of the estimated covariance matrix. Using the weighted least square method outlined in Cressie (1993), we obtained the parameter estimates and hence the spatial covariance matrix. After fitting the proposed model, the histogram from the fitted residuals is presented in Figure 3.6. No clear model violation or normality assumption is indicated.

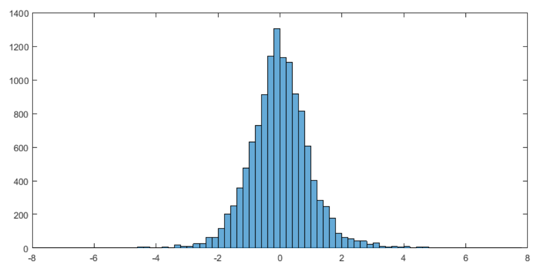

Figure 3.6: Histogram for fitted residuals

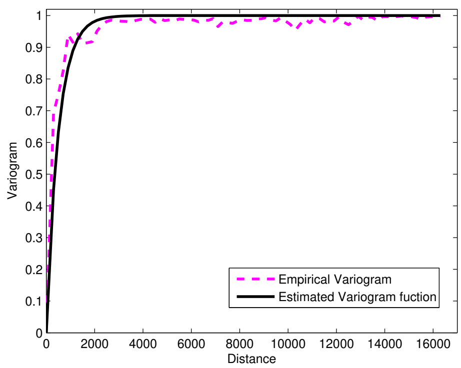

Figure 3.7: Variogram with fitted curve

Figure 3.8 and Table 3.3 report the varying coefficient β estimates at the 76 locations. The value of $\widehat{\beta}$ informs the percentage change in house price per square meters in response to 1 percentage change in living space, whereas $1 + \widehat{\beta}$ is more conveniently interpreted as the price elasticity of space or the implicit price of living space.

Full estimates reveal that the coefficient displays substantial variability over space. The highest estimated elasticity $(1 + \widehat{\beta})$ is observed for Barra (0.6158), which is in the beach area. For an average property in Barra, 1 percent increase in the living space will bring an estimated 0.6158 percent rise in housing price. Interestingly, Barra is far from the city centre and only have limited access to the mainland. This suggests that the beach area is valued for vacation use and tourist rental which attracts the high premium. Apart from the beach, high implicit prices are found in the city centres of Aveiro and Ílhavo (yellow and green dots in Figure 3.8), where the population is dense and location there brings easy access to central amenities.

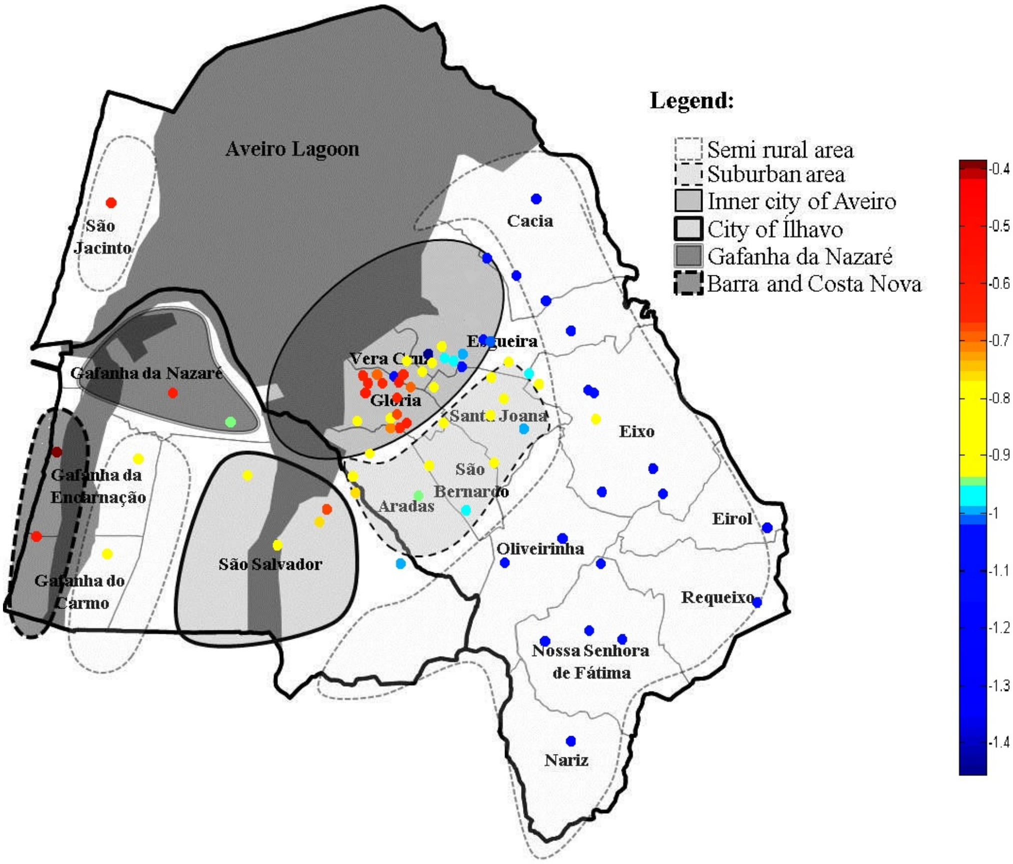

Figure 3.8: Spatially varying coefficient estimates over the 76 zones

Around the eastern and southern boundary of the map, most implicit price estimates are almost negligible (blue dots). The lowest estimate is observed at Eirol (−0.2865), followed by Sarrazola (−0.2168), Nariz (−0.1963), all of which are rural areas. Note that a few of estimates are slightly below the theoretical range of price elasticity for normal goods (0, 1). Table 3.3 shows that all the out-of-range zones, where a larger house brings added costs of maintenance and low rental

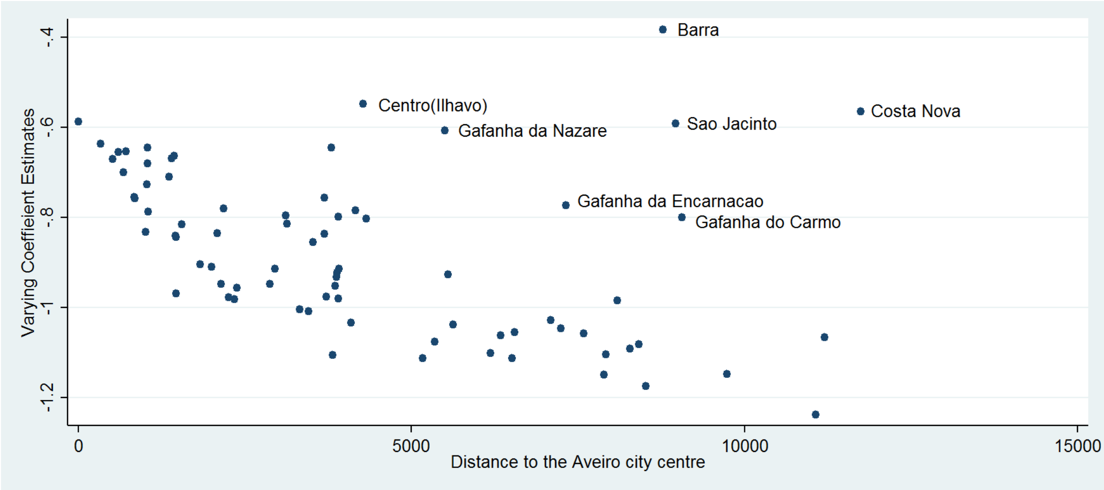

Figure 3.9: Distance decay of implicit price by distance from Aveiro CBD

potential, are either in rural or suburban areas. In those districts, the population is relatively sparse and facilities are less accessible, which makes extra living space undesirable. Figure 3.9 plots varying coefficient estimates against distance to the Aveiro city centre. A clear decreasing trend shows that the elasticity decreases gradually as we move from centre to periphery, which agrees with both intuition and economic theory. A few exceptions are the beach areas, and areas close to Ílhavo.

Table 3.3 also indicates a large variation in the number of data points used for local estimation. In the city centre, each location generally has one or more neighbours, while in the periphery, many do not have any neighbour at all. The variation is caused by the non-uniform design density across space, which leads to uneven sampling frequencies. Varying local sample size suggests that the accuracy of prediction could be different at different locations. Table 3.3 also reports local MSE at 76 zones and regression analysis indicates negative relations between local MSE and both number of houses in locality and in the neighbourhood.

Zone	Region-type	Price Elasticity	No. Obs (Within zone)	No. of Neighbours	Obs in the neighbour	MSE (Within zone)
Barra	Beaches	0.6158	705	0	0	0.0641
Centro (Ilhavo)	CBD Ilhavo	0.4639	1158	1	29	0.0503
Gulbenkian	CBD Aveiro	0.4428	2	5	430	0.6407
Costa Nova	Beaches	0.4182	100	0	0	0.0905
Glicinias	CBD Aveiro	0.4091	151	4	355	0.0344
Gafanha da Nazare	Gafanhas	0.3962	1444	0	0	0.0607
Alboi	CBD Aveiro	0.3878	89	6	331	0.0465
Bairro do Liceu	CBD Aveiro	0.3865	169	7	312	0.0464
Rossio	CBD Aveiro	0.3846	100	5	199	0.0572
Sao Jacinto	Rural	0.3844	28	0	0	0.0959
Mario Sacramento	CBD Aveiro	0.3751	102	5	573	0.0583
Forum	CBD Aveiro	0.3555	2	9	600	0.4284
Eucalipto	CBD Aveiro	0.3522	189	4	317	0.0438
Cancela	CBD Ilhavo	0.3517	29	1	1158	0.0325
Bairro de Santiago	CBD Aveiro	0.3366	43	6	667	0.1346
Beira Mar	CBD Aveiro	0.3243	70	7	357	0.0810
Vila Jovem / Santiago	CBD Aveiro	0.3022	21	5	520	0.0326
Centro de Congressos	Gafanhas	0.2955	6	7	1005	0.0019
Feira de Marco	Suburban	0.2871	121	10	1415	0.0251
Oita	CBD Aveiro	0.2627	7	8	1247	0.0091
Avenida Dr Lourenco Peixinho	Suburban	0.2563	36	11	1600	0.0786
Ribas	Suburban	0.2390	44	1	37	0.0332
Verdemilho	CBD Aveiro	0.2147	233	0	0	0.0287
Gafanha da Encarnacao	CBD Aveiro	0.2141	258	0	0	0.0349
Gafanha D"aquem	Rural	0.2042	82	0	0	0.0240
Alagoas	Suburban	0.2007	526	0	0	0.0174
Coutada/Medela	Suburban	0.1955	37	1	44	0.0294
Forca	CBD Aveiro	0.1952	635	2	41	0.0256
Patela	Suburban	0.1834	223	2	10	0.0182
Santiago	CBD Aveiro	0.1812	35	2	64	0.0576
Vista Alegre	CBD Aveiro	0.1812	36	0	0	0.0171
Gafanha do Carmo	Rural	0.1770	40	0	0	0.0367
Estacao	CBD Aveiro	0.1678	35	9	2066	0.0202
Barrocas	CBD Aveiro	0.1658	975	6	203	0.0388
Vilar	Suburban	0.1589	217	1	2	0.0177
Agras do Norte	Suburban	0.1452	11	1	2	0.0974
Cabo Luis/Quinta das Acacias	Suburban	0.1359	3	2	136	0.1819
Viaduto	Suburban	0.0931	2	7	1631	0.0028
Sol Posto/Presa	Suburban	0.0904	3	1	223	0.0009
Viso/Caiao	Suburban	0.0800	17	2	116	0.0747
Aradas	Suburban	0.0706	32	0	0	0.0547
Sao Bernardo	Suburban	0.0694	1143	0	0	0.0297
Agras	CBD Aveiro	0.0661	2	2	986	0.0144
Azenha de Baixo	Suburban	0.0652	102	0	0	0.0459
Bonsucesso	Suburban	0.0604	193	0	0	0.0286
Quinta do Cruzeiro	CBD Aveiro	0.0524	233	4	201	0.0249
Cale da Vila	Gafanhas	0.0493	63	0	0	0.0347
Quinta do Picado	Rural	0.0250	139	0	0	0.0411
Esgueira	CBD Aveiro	0.0191	51	4	383	0.0368
Quinta da Bela Vista	Suburban	0.0147	119	1	3	0.0119
Carramona	CBD Aveiro	0.0117	109	0	0	0.0152
Quinta do Gato	Suburban	0.0034	4	0	0	0.0081
Moitinhos	Rural	0.0029	26	2	157	0.0391
Escolas	CBD Ilhavo	-0.0152	113	4	303	0.0373
Cidadela/Quinta de Santo Antonio	Suburban	-0.0173	7	2	232	0.0108
Olho d"agua	Rural	-0.0396	9	2	159	0.1001
Mataducos	Suburban	-0.0561	152	1	9	0.0446
Quinta do Loureiro	Suburban	-0.0591	129	0	0	0.0390
Paco	Rural	-0.0676	92	1	97	0.0249
Nossa Senhora	Suburban	-0.0863	94	0	0	0.0730

de Fatima						
Oliveirinha	Rural	-0.0865	22	0	0	0.0195
Granja de Baixo	Rural	-0.0925	13	0	0	0.0421
Povoa do Paco	Rural	-0.0937	97	1	92	0.0524
Costa do Valado	Rural	-0.1063	118	0	0	0.0199
Requeixo	Suburban	-0.1081	41	0	0	0.0866
Cilhas	Rural	-0.1223	17	0	0	0.0248
Cacia	Rural	-0.1245	13	0	0	0.0237
Quintas	Rural	-0.1362	127	0	0	0.0499
Azurva	Rural	-0.1407	563	0	0	0.0464
Povoa do Valado	Rural	-0.1419	3	0	0	0.0037
Taboeira	Suburban	-0.1462	69	0	0	0.0707
Mamodeiro	Rural	-0.1464	29	0	0	0.0762
Eixo	Rural	-0.1891	281	0	0	0.0711
Nariz	Rural	-0.1963	37	0	0	0.0963
Sarrazola	Rural	-0.2168	235	0	0	0.0389
Eirol	Rural	-0.2865	6	0	0	0.0881

Table 3.3: SVCR hedonic estimates by zones

3.4.4 Prediction and prediction intervals

The proposed methods are BLUP-based, and therefore an important advantage is the ease with which predictions and prediction intervals can be obtained. In the context of the current application, such predictions have immense importance. For real estate agencies, the proposed methods provide an easy way to value properties for their clients, paying adequate attention to different preferences, and social and neighbourhood conditions in different places. Such differences then lead to different implicit prices, for example, for living space. For homebuyers, such predictions provide easy ways to balance different priorities across a range of heterogeneous characteristics of alternate houses and neighbourhoods. Finally, for local and national governments, predictions provide useful ways to evaluate the affordability for housing and the potential need for state support. Importantly, our method provides simple-to-compute out-of-sample predictions not only in terms of regressors with fixed and varying coefficients but also potential new locations that have not been sampled. This is also useful for developers planning new build of houses. In other contexts, there is substantial importance attached to the predictions at out-of-sample locations where data could not be collected, typically because of cost considerations; see, for example, Goulard et al. (2017).

To explore predicted prices, we choose three regions within our study domain with a high price elasticity of living space: CBD Aveiro, CBD Ílhavo and Beaches. In each case, we consider the centroid of the sampled locations from the specific region, which is not a sampled location in itself. For the central area of Aveiro, this location turns out to be in Centro de Congressos, which is the seat of the local government but contains only six sampled houses. This small local sample size does not matter because the method draws adequately upon information from other locations in the neighbourhood. For the CBD of Ílhavo, this centroid lies in the zone Centro (Ílhavo) which has high design density. For Beaches, the chosen location lies in Barra, which is right on the edge of the study region. Then, estimates for the location will also provide some intuitive insights as to boundary spillages.

Predictions are obtained under two different scenarios: (a) local median and (b) aggregate average. For the first, we set covariates, including living space, at the median value for the respective area: CBD Aveiro, CBD Ílhavo or Beaches. For the second, we place the covariates at the average value for the entire housing market, including Gafanhas, suburban and rural areas. This design will allow for evaluation of prices locally within the region but also for houses with comparable characteristics placed at different locations. In effect, this enables us to evaluate how important living space is to homebuyers, as compared with other characteristics of the lived environment, including access to local and central services and facilities within the house.

Region	Scenario	Pred. $\ln(P/S^2)$	Pred. P (Euros)
CBD Aveiro	Local	7.6049 (7.2683,7.9413)	EUR 209,065 (200780,292708)
CBD Ílhavo	Local	6.7809 (6.3420,7.2197)	EUR 123,317 (79512,191254)
Beaches	Local	7.3725 (6.8758,7.8692)	EUR 162,344 (98796,266770)
CBD Aveiro	Average	7.4650 (7.1285,7.8015)	EUR 230,892 (164913,323268)
CBD Ílhavo	Average	8.2876 (7.8487,8.7264)	EUR 525,597 (338895,815156)
Beaches	Average	9.0296 (8.5329,9.5262)	EUR 1,103,809 (671731,1813812)

Table 3.4: Spatial prediction using SVCR

Table 3.4 reports the predicted prices. We first predict the logarithm of the price per square meter and its 95% prediction interval. Then, we translate this into absolute house prices in Euros. Some interesting observations follow. As one would expect, the price of a local representative house is highest in the CBD of Aveiro, at Euros 209,065 (95% prediction interval: 200,780 to 292,708). The price of a representative house on the beaches is about 22 percent lower, and one in the CBD of Ílhavo is lower by another 24 percent. However, this is only a representative house in that area. There is important variation over the spatial domain in living space and housing quality, neighbourhood quality and perhaps most importantly in access to services. Now, if a house with average (representative) characteristics from the entire domain are placed in each of these three areas, the situation dramatically changes. This index house has a predicted price of a whopping Euros $1,103,809$ (95% prediction interval: $671,731,181,3812$) if it is located on the beaches. The same house would be valued at one-fifth of the price (Euros $230,892$) if it were placed in the centre of Aveiro. Even the 95% prediction interval (Euros $164,913$ to $323,268$) does not intersect. Thus, the high predicted price of a local representative house in the CBD of Aveiro was not so much due to living space, but other covariates. In fact, residents value closeness to the centre and therefore the labour market and central facilities far more than the size of the house itself. The same house placed in the CBD of Ílhavo would fetch a value 47 percent of that on the beach, but the 95% prediction interval is still disjoint from that in the CBD of Aveiro. Houses in Ílhavo are on average larger compared to both the CBD of Aveiro and the beach area. However, prices in the centre of Aveiro are higher because of proximity to the CBD, and prices in the beaches are higher because of investment and leisure value. Such insights are crucial for the participants in the housing market, but also the local government.

3.5 Conclusion

Spatial applications in housing markets, as well as many other areas, have several distinguishing features: irregular and complex spatial domain, spatially varying coefficients, spatially autocorrelated errors, large spatial dimension, and measurement error. The current literature in spatial statistics addresses these issues in a piecemeal manner. There is substantial existing research in spatial smoothing over irregular domains, but these methods are not appropriately extended to a regression context. The literature on spatially varying coefficients pays little attention to spatially autocorrelated errors. Methods for modelling spatial autocorrelation by variogram fitting also include measurement error, but not spatially varying slopes.

We consider an application in housing markets where all of the above issues are concurrently present. Our objects of inference are both the spatially varying coefficients, as well as prediction at new locations where there are no observed data. We propose an innovative methodology to estimate the spatially varying surface of a regression slope parameter, in the presence of potentially other covariates with fixed slopes, measurement error and spatially autocorrelated errors. The method is easy to apply and computationally efficient when the spatial domain is irregular, and when the dataset is large. Applied to data on the Aveiro-Ílhavo urban housing market in Portugal, the method takes adequate account of the complex nature of the spatial domain and recovers intricate spatial patterns in the implicit price of living space. The estimates obtained offer useful interpretation within the context of the application. The predictions derived from our model provide exciting new insights into housing preferences and spatial variation in prices.

Chapter 4

The Geography of Human Activity and Urban Consumption in Beijing: A Spatiotemporally Varying Coefficient Regression Approach

4.1 Introduction

The rise of Chinese megacities is marvellous. Since the economic reform in 1978, the ratio of the urban population in China has surged from 19% in 1980 to 58% in 2017 (The World Bank, 2018). In Beijing, the capital city, the population has more than doubled from 8.7 million to 21.7 million in the same period (Beijing Municipal Bureau of Statistics, 2018). Despite the adverse effects such as congestion and pollution, Chinese megacities play a critical role in boosting productivity and economic growth by lowering the cost of transport for goods, services, people and ideas. As urban household income continues to rise with the booming economy, the role of a city in facilitating consumption has been highlighted in the recent literature (Glaeser et al., 2001; Glaeser and Gottlieb, 2006; Rappaport, 2008). Like its western counterparts, modern-day Chinese megacities attract people by providing a wide range of consumer goods such as restaurants, pubs, theatres, museums, and grocery stores. The growth of the urban population, on the other hand, makes it possible for cities to take advantage of the economy of scale. Higher urban population density brings better and more diverse

[0]Earlier version of this chapter was co-authored with Wenjie Wu, Jianghao Wang and Chengyu Li under the title '*The Geography of City Liveliness and Consumption: Evidence from Location Based Big Data*' and is archived as SERC Discussion Paper (see Wu et al., 2016b)

consumption amenities such as concerts, basketball matches, exotic restaurants and transportation infrastructure. As Jane Jacobs put it, *'Liveliness and variety attract more liveliness; deadness and monotony repel life'* (Jacobs, 1962).

In this chapter, we study the relationship between human activity and consumption amenities in Beijing, China. We ask whether the liveliness of Beijing benefits from the configuration of consumption amenities, and what insight can be gained about the spatiotemporal heterogeneity of this relationship. To facilitate our investigation, we collect two unique datasets, (1) high frequency mobile phone positioning (MPP) data from Chinese social network application *WeChat*, and (2) large-scale point-of-interest (POI) data from local consumption website *Meituan-Dianping*[1], the Chinese version of Yelp. The first dataset records locations of all the *WeChat* users with mobile phone GPS enabled in metropolitan Beijing each hour during a two-week window in June 2015. Due to the popularity of *WeChat* in China, the spatiotemporal distribution of mobile phone positioning serves as an excellent proxy for the real-time human activity distribution in Beijing. The second dataset gathers all the businesses and amenities with physical addresses registered at *Meituan-Dianping*. The comprehensive geo-coded datasets give us detailed information about the spatial distribution of consumption amenities.

Our study is of interest in several aspects. First of all, the future of cities depends on people's continued willingness to live in high density and take advantage of agglomeration effects (Glaeser, 2010). With Chinese urban residents growing richer, the attractiveness of a city not only lies in salaries it can offer but also in the quality of life it provides. As urban consumption is most relevant to the quality of life a resident would enjoy, understanding the linkage between consumption amenities and human activity carries significant importance in urban planning and urban policy-making. In the 'Consumer City' paper, Glaeser et al. (2001) sheds light on the role of urban density in facilitating consumption activities. They argue that the existence of a wide variety of consumer goods is one of the most important urban amenities. Empirically, they find a strong positive relationship between urban amenities and population growth. Following Glaeser et al. (2001), an increasing number of theoretical and empirical studies have emerged to discuss the role of urban consumption and its interaction with the urban population. Using US restaurant data, Schiff (2014) finds that the population density has a substantial impact on the consumer product variety through demand aggregation, even if cities have otherwise identical characteristics. Couture (2013)

[1] *Meituan* and *Dianping* were previously independent websites followed by a merger in 2015.

estimates the consumption value of density using travel and local business data, in which he finds higher density areas benefits consumers through gains in variety. In the context of China, Zheng and Kahn (2008) find that local urban amenities, such as public infrastructure, good schools, pleasant environment are significantly capitalised into households' willingness to pay for proprieties. Wu et al. (2017) confirm that substantial value is attached to the accessibility of various amenities in Beijing, with significant spatial heterogeneity. This allows us to assume that the configuration of local amenities influences residents' quality of life. Understanding the spatial and temporal connection between consumption amenities and human activity at the local level enables policymakers and developers to make better decisions regarding how land resources and capitals should be allocated. Our study about the relationship between the distribution of consumption amenities and human activity contributes directly to this line of research.

Second, since 1990s China has started reinstating the urban land market as well as removing migration restrictions in an effort to transform the centrally planned economy into a market-oriented economy. Economic liberalisation catalyses the explosive growth of construction and urban development which continually changing the face of Beijing (Zheng and Kahn, 2008). On the other hand, with better public transportation and increasing car ownership (Zheng et al., 2016), Beijing residents are able to move swiftly from one place to another, constantly shifting the spatiotemporal distribution of human activity. As the texture of the city evolves rapidly, current county-level census data published after relatively long periods often fail to provide up-to-date information at granular geographical scale (Glaeser et al., 2017). To gain insight into the contemporaneous local economic activity, many researchers have turned to alternative location-based data sources, such as those from social media, traffic cameras, smart cards, crowdsourced websites and mobile phones (Long et al., 2012; Li et al., 2013; Liu et al., 2014; Wu et al., 2016a; Glaeser et al., 2017). Our study adds to a growing body of urban economics literature using location-based big data from online platforms. In particular, using the MPP data allows us to track the quantity and distribution of within-city human activity with high spatiotemporal precision, therefore providing an opportunity to dive into the administrative boundary and investigate the dynamics of the city liveliness at fine scales.

Third, whereas the distribution of human activity can display significant variation within a short window (e.g., 24 hours), the distribution of consumption amenities typically remains stable overnight. This intuition suggests that the relationship between human activity and configuration of amenities is far from static within a short period. For example, shopping malls are likely to attract more customers during the day when they are open, but fewer people when they are closed at night. However, few empirical studies investigate the heterogeneous relationship between consumption and human activity.

Our study fills the gap. To identify the spatial and temporal heterogeneous effects between human activity and local consumption amenities, we propose a semiparametric regression model called spatially and temporally varying coefficient regression (STVCR), adapted from the statistical varying coefficient models (Fan and Zhang, 2008). We relax the constant-coefficient regression assumption in favour of coefficients that vary with both space and time. This setup provides a framework to investigate spatial-temporal heterogeneous effects. Methodologically, the proposed STVCR model contributes to the recent development in semi-nonparametric spatial and temporal models. Fotheringham et al. (2015) proposed the Geographical and Temporal Weighted Regression (GTWR), which extends the Geographical Weighted Regression (GWR) into the temporal dimension. Our approach is similar to GTWR but also differs in two ways. On the one hand, the STVCR allows some of the regression coefficients to be non-varying, rather than fully varying in the GTWR, therefore provides a more flexible regression framework with different types of coefficients. On the other hand, we apply a local linear estimator to estimate the model, as opposed to the GTWR, which is a type of local constant estimator. Using the first-order Taylor approximation, the local linear estimator controls the size of bias in semi-parametric estimation. We conduct a Monte Carlo study to compare the performances of the STVCR and GTWR. The results back the theory that the local linear estimator outperforms the local constant estimator.

The remainder of this paper is organised as follows. Section 4.2 details the context and the data used in this study. Section 4.3 proposes the methodology. Section 4.4 presents the empirical results and Section 4.5 concludes.

4.2 Context and Data

4.2.1 Beijing Metropolitan Area

Beijing (or Peking) is located in northern China, south of Yanshan mountain and surrounded by Hebei province. The city displays a monocentric structure with Tiananmen square being the geographic centre and the surrounding CBD ('*JianGuoMenWai*') being the centre of economic activity and employment (Zheng and

Kahn, 2008). The monocentric city structure has been a significant feature that differentiates Chinese cities from cities in developed economies, where urban employment has been decentralised to suburbs (Glaeser and Kahn, 2001). From the city centre, Beijing extends along with all directions with eight rings roads nested on each other. Under Beijing municipality there are 16 districts, often categorised into three types: (1) Inner city–districts within 5th ring road, including *Dongcheng*, *Xicheng*, *Chaoyang*, *Fengtai*, *Shijingshan*, *Haidian*; (2) Inner suburb–districts connected by 6th ring road including *Mentougou*, *Fangshan*, *Tongzhou*, *Shunyi*, *Changping*, *Daxing*; (3) Outer suburb – districts outside 6th ring roads, including *Huairou*, *Pinggu*, *Miyun*, *Yanqing*. Each district is further split into smaller administrative areas, called neighbourhoods, or '*Jiedao*'[2]. There are altogether 149 Jiedao.

4.2.2 Mobile Phone Positioning (MPP) Data

To measure the exact real-time human activity pattern in Beijing, we extract mobile phone positioning (MPP) data from a Chinese mobile application '*WeChat*'[3]. *WeChat* is an instant messaging and social networking application developed by *Tencent*[4], a technology conglomerate based in Shenzhen, China. Since its first release in 2011, it has become one of the largest mobile applications in the world with monthly active users rising from 14 million in 2011 Q3 to 1.08 billion in 2018 Q4 according to *Tencent*'s quarterly report. As the popularity of *WeChat* soars in China, *Tencent* starts assembling large quantities of user and usage information for market research and personalisation purposes. A critical piece of the information *Tencent* has been compiling is 'user location', the latitude and longitude tracked by global positioning system (GPS). These records can be accessed on *Tencent*'s platform '*Easygo*' (https://heat.qq.com/), via the Institute of Geographic Science and Natural Resources Research Centre (NRRC), Chinese Academy of Sciences (CAS). The platform provides an application programming interface (API) through which mobile position numbers at a spatial unit, during a time interval can be retrieved. From the platform, we extracted the MPP data from 15th June 2015 to 28th June 2015 in Beijing. The dataset is presented as spatiotemporal count data. During each hour, the MPP number is counted over $1km^2$ grids in Beijing. A larger number indicates that more WeChat users have been visiting this area for that hour and vice versa. To avoid large sampling errors

[2] 'Jiedao' literally means 'street' in Mandarin.
[3] '*WeChat*', also known as '*WeiXin*', means 'micro-messaging' in Mandarin Chinese.
[4] https://www.tencent.com/en-us/

from specific days, we further aggregate the original dataset which spans over two weeks into two sub-datasets, a weekday dataset and a weekend dataset, each of which spans 24 hours.

The weekday data ($Y^{(wd)}$) are aggregated as follows,

$$Y_{i,t}^{(wd)} = \frac{1}{10} \sum_{\tau} Y_{i,t}^{\tau}, \quad t \in 0, 1, 2, \ldots, 23, i \in 1, 2, \ldots, N \tag{4.2.1}$$

where $Y_{i,t}^{\tau}$ is the mobile phone density at hour t and location i on day τ between 15/06/2015 to 19/06/2015 and 22/06/2015 to 26/06/2015.

Likewise, we aggregate the weekend data ($Y^{(we)}$) as

$$Y_{i,t}^{(we)} = \frac{1}{4} \sum_{\omega} Y_{i,t}^{\omega}, \quad t \in 0, 1, 2, \ldots, 23, i \in 1, 2, \ldots, N \tag{4.2.2}$$

where $Y_{i,t}^{\tau}$ is the mobile phone density at hour t and location i on day τ between 20/06/2015 to 21/06/2015 and 27/06/2015 to 28/06/2015.

Figure 4.1 plots the heatmap of the spatial distribution of the weekday data over 6 periods: midnight (0:00-1:00), early morning (4:00-5:00), morning peak hours (8:00-9:00), midday (12:00-13:00), afternoon (16:00-17:00), and evening (20:00-21:00). In each plot, we find that the spatial distribution of the densities displays a monocentric pattern that stretches from the CBD to the outer suburbs in all directions. The colour is darker in the centre of Beijing, indicating a higher positioning density. The density decays with distance from the centre, consistent with Beijing's economic pattern found in the literature (Zheng and Kahn, 2008; Wang, 2009). Comparing different plots, we find that the temporal distribution of the density follows humans' periodic living pattern in a day. The overall density starts lowering in the evening as people turn off the phones and go to sleep. It reaches the trough in the early morning and bounces back. The positioning density maintains high during the day until the next evening starts. The dataset reports on average 450 million mobile phone positioning in Beijing per day, with 22 million unique users. Given the official population data is 16.44 million in 2015 (Beijing Municipal Bureau of Statistics, 2018), the positioning data suggest a high proportion of residents not captured by the current census data, such as non-local commuters, visitors and tourists.

Though the sheer size of the positioning number gives us confidence about the repressiveness of the *WeChat* users in Beijing, there are several limitations associated with the MPP data. First, despite its popularity, there is still no guarantee that the MPP data cover all the demographics and areas equally. For example,

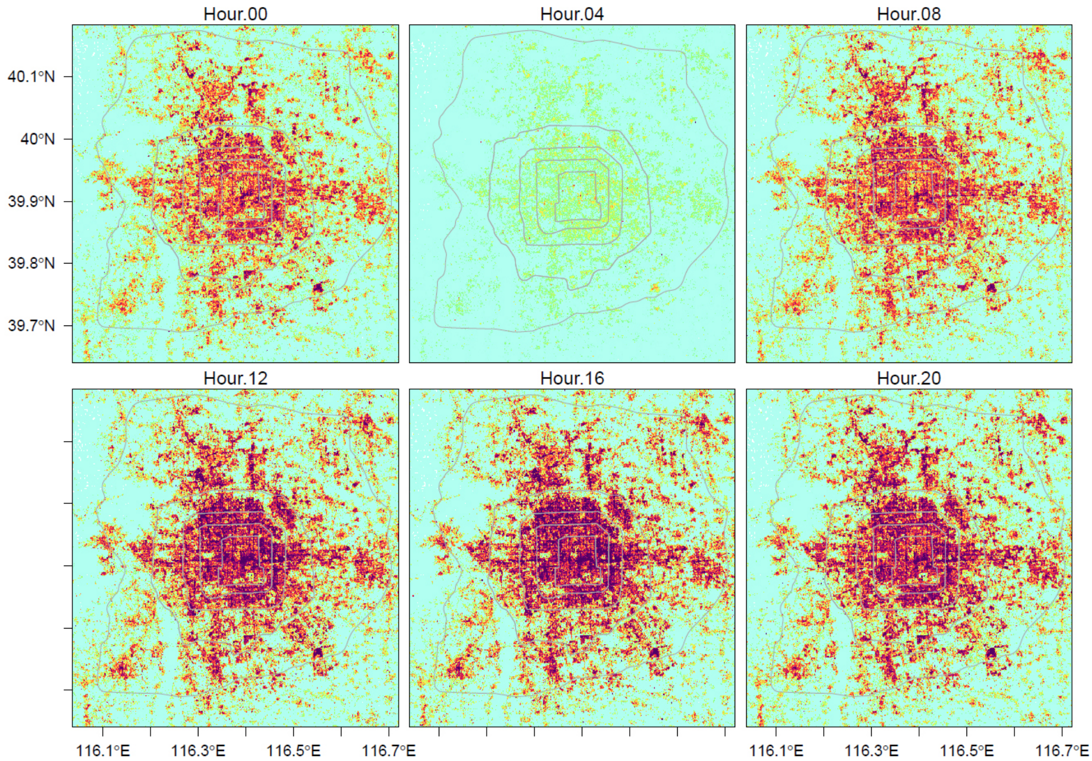

Figure 4.1: Spatiotemporal distribution of the MPP data in Beijing

it is likely that millennials have better access to smartphones required to install *WeChat* compared with the elderly who are less technologically sophisticated. Income difference may also generate a gap in mobile phone and *WeChat* usage among different demographics. Addressing this issue requires the socioeconomic data associated with each positioning point, which are not available due to privacy protection reasons. We are also not able to pinpoint the address of each mobile phone positioning, limiting the possibility to infer user activities, nor are we able to know the motivation behind each movement across the city. This is not a grave issue in our study, as we are investigating the urban human activity pattern at an aggregated time-space level, rather than mobility at an individual level. A similar study in Amsterdam by Jacobs-Crisioni et al. (2014) suggests that mobile positioning aggregated at collective levels are useful to analyse urban activity. We do acknowledge that using the mobile phone position data to understand individual behaviour would be an exciting field that deserves more future studies. The second issue with the MPP data is related to the nature of the *WeChat* data collection mechanism. The location of a user is logged every time he (or she) opens a *WeChat* application on a GPS-enabled phone. The frequency of recording, therefore, varies significantly depending on user habit and immediate environment. For *WeChat* users without functioning GPS, no positioning will be reported. Some workplaces and schools have mobile phone usage restrictions which disguises the population density in the area. On the other hand, users with a high dependency on the application will generate more records than those who use *WeChat* less frequently. To reduce the bias from a single user's duplicated records, we only take the first positioning in any given hour for each unique user. Repeated locations generated by the same user in each hour is therefore dropped.

To verify the appropriateness of the *WeChat* positioning as a proxy for human activity level, as well as its representativeness, we conduct two analyses to compare the MPP data with other data sources. In the first analysis, we collect the Beijing population density estimates from LandScan$^{\text{TM}}$, a community standard for global population distribution. Using various sources such as census, geographic and satellite data, LandScan$^{\text{TM}}$ employs multivariate dasymetric modelling to estimate the ambient population distribution (Dobson et al., 2000; Bhaduri et al., 2002). We compare the LandScan$^{\text{TM}}$ population density estimates in 2015 against the temporal average of the *WeChat* MPP at the same $1km^2$ spatial scale. Figure 4.2 shows the scatter plot between the LandScan population and the MPP data, with the correlation calculated as 0.76 (p-value< 0.001). This suggests that in the spatial dimension, these two estimates exhibit a high correlation.

Second, we extend our investigation into the temporal dimension. From Visible Infrared Imaging Radiometer Suite (VIIRS)[5], we collect the nightlight satellite data in Beijing from 15th June 2015 to 28th June 2015, the same period as the *WeChat* data. Since Henderson et al. (2012), luminosity has been widely used in many empirical studies as a robust proxy for GDP, in the absence of reliable data sources (see, e.g., Michalopoulos and Papaioannou, 2013; Ahrens, 2015). Here, we consider the night light in Beijing as a proxy for economic activities carried out by residents at night time. In Figure 4.3, we plot the $1km^2$-aggregated luminosity in Beijing against the MPP between 8 to 10 pm weekday. The scatter plot indicates a significant correlation, 0.71 (p-value< 0.001), between the two variables. These two exercises provide evidence that the *WeChat* MPP are indeed good proxies for human activities at the local aggregate level. The finding is consistent with recent studies using similar data sources (Mayer-Schönberger and Cukier, 2013; Jacobs-Crisioni et al., 2014; Wu et al., 2016a).

[5]https://earthdata.nasa.gov/earth-observation-data/near-real-time/download-nrt-data/viirs-nrt

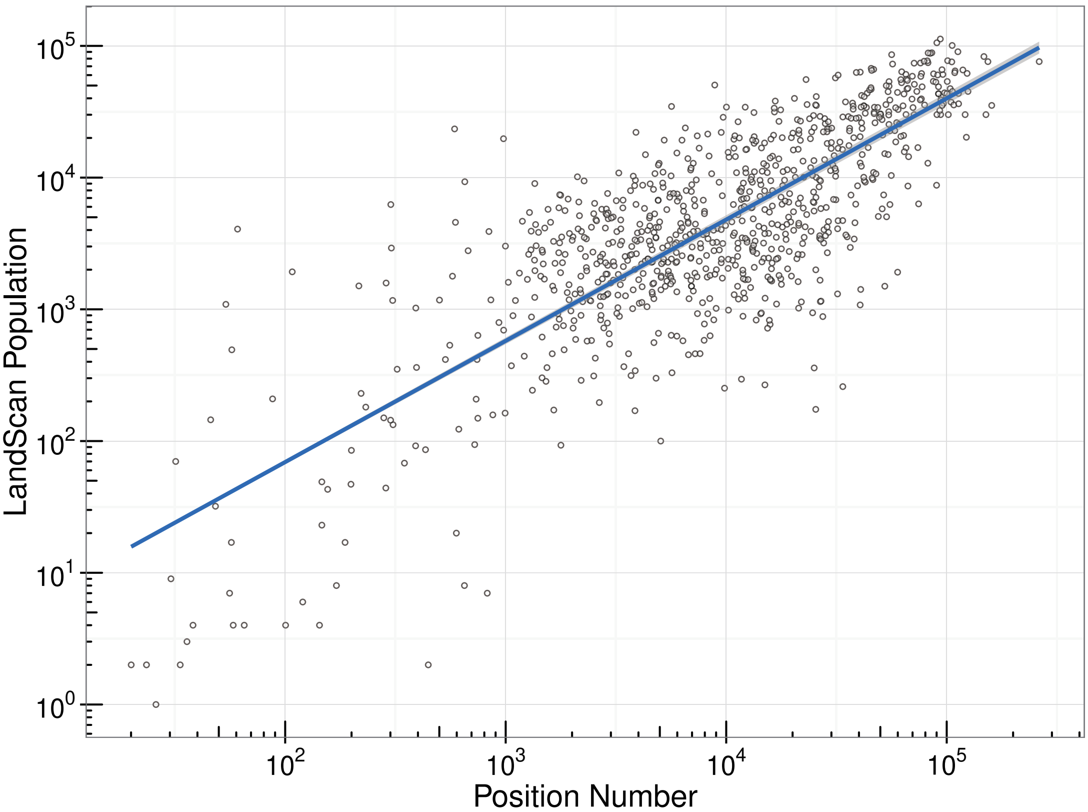

Figure 4.2: Correlation between LandScan™ population density and *WeChat* mobile phone positioning density. (x, y axes in log scales)

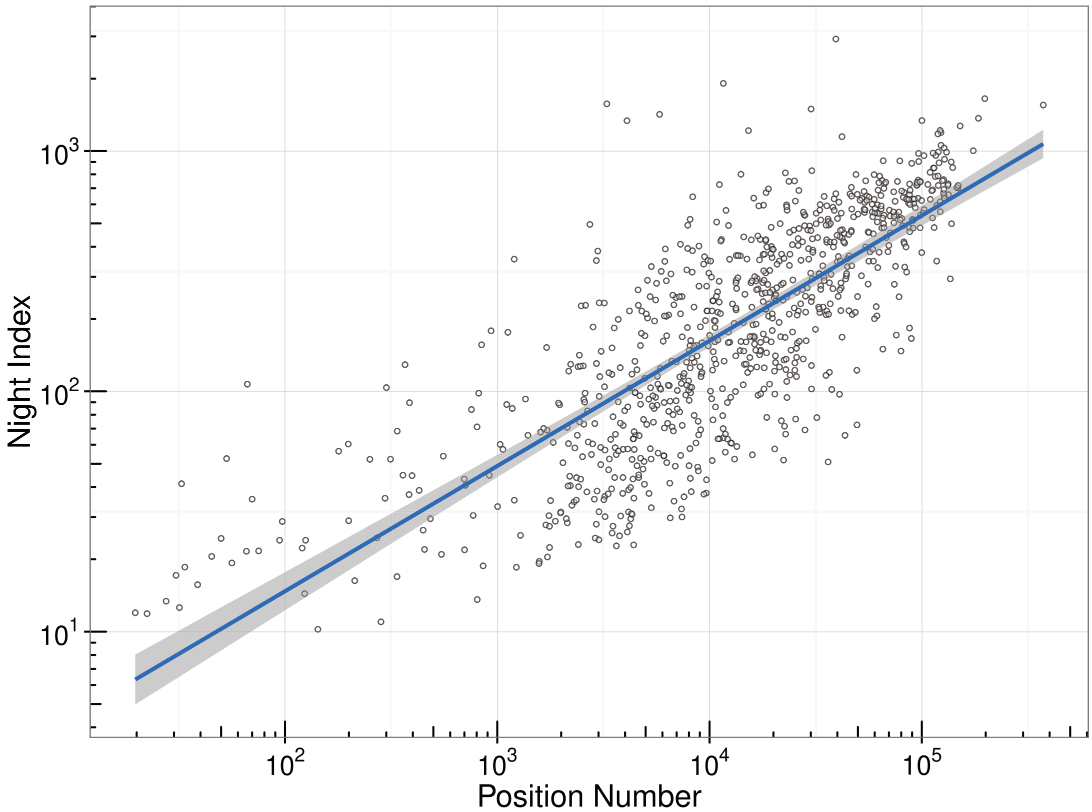

Figure 4.3: Correlation between night light intensity distribution and *WeChat* mobile phone positioning density. (x, y axes in log scales)

4.2.3 Point of Interest (POI) Data

To measure the spatial configurations of the consumption amenities in Beijing, we collect the geo-coded point of interest (POI) data from China's group buying and crowdsourced review website *Meituan-Dianping*[6]. *Meituan*[7] and *Dianping*[8] were previously two independent online platforms providing local consumption services. While *Meituan* specialised in group buying like *Groupon*[9], *Dianping* focused on crowdsourced reviewing like *Yelp*[10] or *Tripadvisor*[11]. Both websites had assembled large numbers of detailed points of interest in China.

From the merged website cached on 28th June 2015, we extract consumption related POIs located in the Beijing metropolitan area. Circa 345,000 consumption related POIs are identified, which include a wide range of amenities such as restaurants, coffee shops, shopping malls, grocery stores, movie theatres, gyms, museums. For each POI, we collect relevant details such as business name, business scope, address, and Meituan-Dianping rating. Coordinates are mapped and retrieved from Google Map. Features related to the heterogeneity of consumption amenities, such as service, quality, size, decoration, are not available. To control for the heterogeneity, we collect 'hits' number reported by *Meituan-Dianping* which reflects the popularity of each POI. Both consumption POIs and hits are counted and aggregated at the same $1km^2$ areal level to match *WeChat* positioning density.

Apart from the consumption amenities, we also collect housing POIs such as residential complex and other POIs which cannot be classified as consumption nor housing, such as hospitals and government buildings. The comprehensiveness of these POIs allows us to visualise the spatial distributions of the amenities in Beijing (Figure 4.4). The left panel of Figure 4.4 displays the overall distribution of the POIs in Beijing, with each blue pixel representing a POI. The right panel zooms in two areas, Wangjing in the upper right and Guomao in the lower right. Like the positioning density, the distribution of the POIs displays a similar monocentric pattern, with higher density in the city centre diminishing density outwards.

[6] http://www.meituan.com/
[7] *Meituan* literally means 'happy group buying'.
[8] *Dianping* literally means 'review'.
[9] https://www.groupon.com/
[10] https://www.yelp.co.uk
[11] https://www.tripadvisor.co.uk/

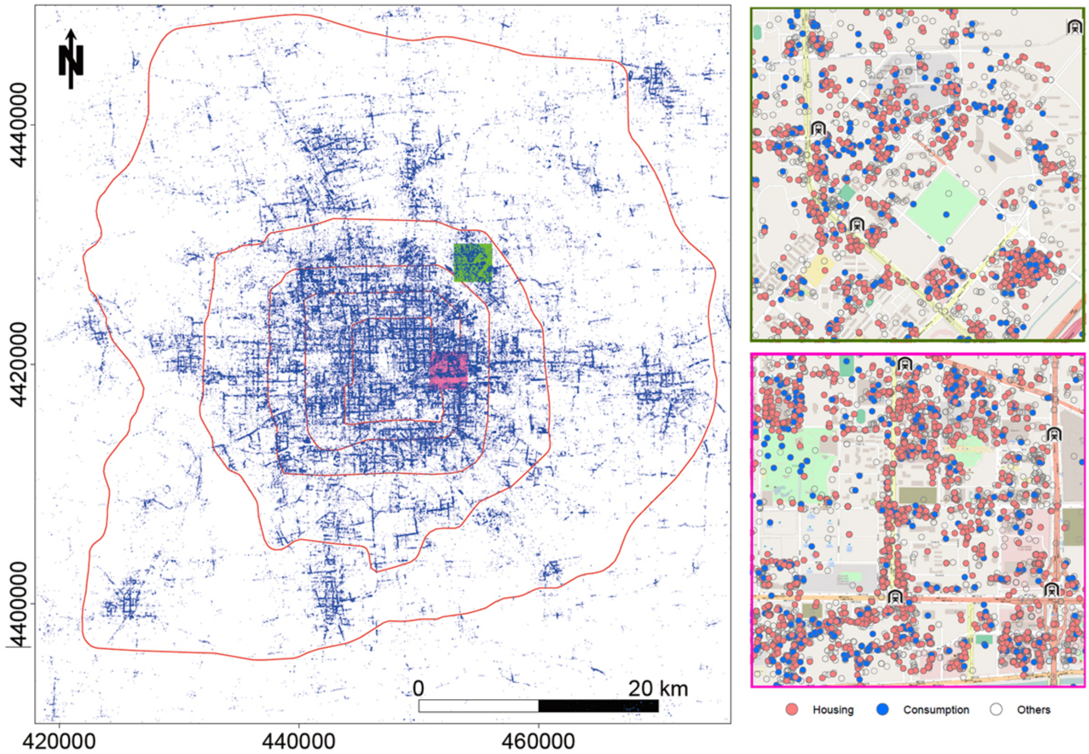

Figure 4.4: Spatial configurations of amenities in Beijing

4.3 Methodology

During 24 hours, the distribution of the human activities shows significant variation at both spatial and temporal dimension. The distribution of the urban amenity, on the other hand, remains stable and static. This fact implies the short-run spatiotemporal relationship between the two variables must be highly non-stationary and heterogeneous. In this section, we develop a spatiotemporally varying coefficient regression (STVCR) model to estimate the heterogeneous effects.

4.3.1 Spatiotemporally Varying Coefficient Regression

A regression model attempts to explain a dependent variable y using a set of explanatory variables. We differentiate two types of explanatory variables, the variables associated with coefficients that are allowed to change with location and time, denoted by $\mathbf{x}^{\mathrm{T}} = (x_1, \ldots, x_k)$ and the variables with non-varying coefficients, denoted by $\boldsymbol{z}^{\mathrm{T}} = (z_1, \ldots, z_l)$. The STVCR model is specified as follows,

$$y_{it} = \mathbf{x}_{it}^{\mathrm{T}} \boldsymbol{\beta}(\mathbf{s}_i, \tau_t) + \mathbf{z}_{it}^{\mathrm{T}} \boldsymbol{\gamma} + \epsilon_{it}, \tag{4.3.3}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^{\mathrm{T}}$, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_l)^{\mathrm{T}}$, $l, k \in \mathbb{Z}^+$. $\mathbf{s}_i = (a_i, b_i) \in S \subset \mathbb{R}^2$, $\tau_t \in I \subset \mathbb{R}$, where both S and I and compact sets. $\mathbf{x}_{it}^{\mathrm{T}}$ and $\mathbf{z}_{it}^{\mathrm{T}}$ are the sample analogue of $\mathbf{x}$ and $\mathbf{z}$. To make our model generic and flexible enough, we label both dependent variables with subscript i and t, even if all the independent variables in our applications are not time-varying.

The STVCR nests many classes of regression models as special cases when restrictions are imposed on the coefficients,

1. (Multiple Linear Regression) $\boldsymbol{\beta} = \mathbf{0}$,
2. (Geographical and Temporal Weighted Regression, GTWR) $\boldsymbol{\gamma} = \mathbf{0}$,
3. (Geographical Weighted Regression, GWR), $\boldsymbol{\beta}(\mathbf{s}_i, \tau_t) = \tilde{\boldsymbol{\beta}}(\mathbf{s}_i)$, $\boldsymbol{\gamma} = \mathbf{0}$,
4. (Spatially Semi-Varying Coefficient Regression) $\boldsymbol{\beta}(\mathbf{s}_i, \tau_t) = \tilde{\boldsymbol{\beta}}(\mathbf{s}_i)$.
5. (Temporally Varying Coefficient Regression) $\boldsymbol{\beta}(\mathbf{s}_i, \tau_t) = \check{\boldsymbol{\beta}}(\tau_t)$.

4.3.2 Local Linear Estimator

Given a sample $(y_{it}, \mathbf{x}_{it}, \mathbf{z}_{it}), i = 1, \ldots, N, t = 1, \ldots, T$, we are particularly interested in estimating the varying coefficients, $\boldsymbol{\beta}(\mathbf{s}, \tau), \mathbf{s} \in S, \tau \in I$, and the non-varying components $\boldsymbol{\gamma}$. Whereas $\boldsymbol{\gamma}$ lies in a finite l dimensional Euclidean space, $\boldsymbol{\beta}(\cdot)$ is a vector of functions that are defined on an infinite dimensional parameter space $S \times I$. To estimate the functional coefficients using a finite sample, we need to apply smoothing or regularisation. In this chapter, we consider the local linear estimator that uses a weighting function and first-order Taylor expansion to achieve estimation.

To illustrate the idea, let $(\mathbf{s}, \tau) \in S \times I$ be a generic location/time pair. For observations that are close to the location/time, the following approximation holds,

$$\begin{aligned} \boldsymbol{\beta}(\mathbf{s}_i, \tau_t) &\approx \boldsymbol{\beta}(\mathbf{s}, \tau) + \frac{\partial \boldsymbol{\beta}(\mathbf{s})}{\partial \mathbf{s}}(\mathbf{s}_i - \mathbf{s}) + \frac{\partial \boldsymbol{\beta}(\mathbf{s})}{\partial \tau}(\tau_t - \tau) \\ &:= \boldsymbol{\beta}(\mathbf{s}, \tau) + \boldsymbol{\beta}^{(\mathbf{s})}(\mathbf{s}_i - \mathbf{s}) + \boldsymbol{\beta}^{(\tau)}(\tau_t - \tau) \end{aligned} \quad (4.3.4)$$

where $\frac{\partial \boldsymbol{\beta}(\mathbf{s})}{\partial \mathbf{s}} = \left(\left(\frac{\partial \beta_1}{\partial a}, \ldots, \frac{\partial \beta_k}{\partial a} \right)^{\mathrm{T}}, \left(\frac{\partial \beta_1}{\partial b}, \ldots, \frac{\partial \beta_k}{\partial b} \right)^{\mathrm{T}} \right)$ and $\frac{\partial \boldsymbol{\beta}(\mathbf{s})}{\partial \tau} = \left(\frac{\partial \beta_1}{\partial \tau}, \ldots, \frac{\partial \beta_k}{\partial \tau} \right)^{\mathrm{T}}$. The Taylor approximation quantifies the relationship between $\boldsymbol{\beta}$ at a generic location/time and that where data are available. Figure 4.5 visualises the approximation. The blue curve represents a generic varying coefficient function β mapped into one dimension, with the point $A(a, b, \tau)$ being a generic location/time where we are interested in estimating the coefficient value, and $B(a_i, b_i, \tau_t)$ a generic

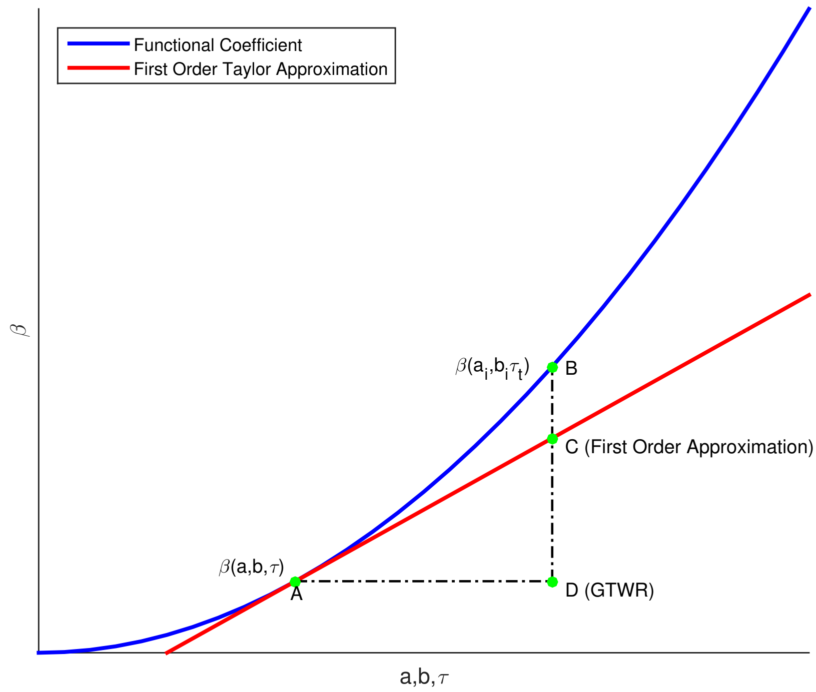

Figure 4.5: Geometry of GTWR and first-order Taylor approximation

location/time where we have data. Using the Taylor approximation, we are able to use the neighbourhood data points to estimate the coefficient at an unsampled location/time. The segment BC measures the approximation error. The GTWR is actually a special case of the Taylor approximation. With $(\mathbf{s}_i, \tau_t)$ close to $(\mathbf{s}, \tau)$, the following approximation holds,

$$\boldsymbol{\beta}(\mathbf{s}_i, \tau_t) \approx \boldsymbol{\beta}(\mathbf{s}, \tau). \tag{4.3.5}$$

The approximation error from the GTWR is reflected in the segment BD. Note that the error of the first-order Taylor approximation is less than that of the GTWR. By including first-order derivatives in the regression, the local linear estimator improves the accuracy of the GTWR.

The quality of Taylor approximation depends critically on two factors (1) the smoothness of the function, (2) the distance between the target location time and the sample location time. For the former, smoother functions are better approximated by a linear function. However, the smoothness of the function is treated as given and determined by the data generating process. For the latter, a sample location/time that is closer to the target location/time incurs less bias.

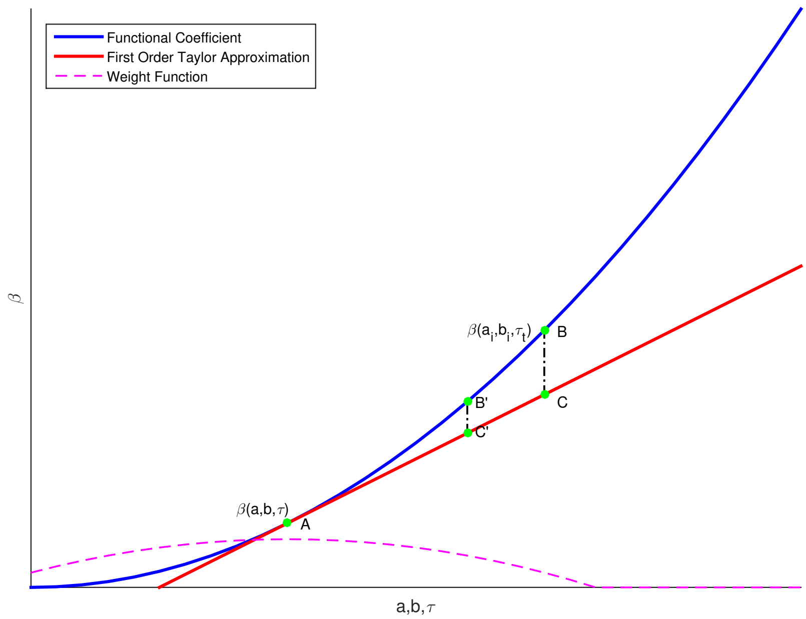

Figure 4.6: STVCR weighting scheme

In Figure 4.6, the location/time B' is closer to A and the bias $B'C'$ is less than BC as a result. To take into account this effect, we define a weight function as the product of a spatial and a temporal kernel,

$$K_{h_s,h_\tau,\mathbf{s},\tau}(\mathbf{s}_i,\tau_t) = K_s(\|(a_i,b_i)-(a,b)\|/h_s) \times K_\tau(|\tau_t-\tau|/h_\tau) \tag{4.3.6}$$

where h_s and h_τ control the spatial bandwidth and temporal bandwidth, respectively. The weight function assigns a high weight if (a_i, b_i, τ_t) is close to (a, b, τ) and a low weight if they are further apart. The dashed line in Figure 4.6 illustrates a weighting scheme. With A being assigned a highest weight, the weight function declines symmetrically at both directions and drops to 0 outside a bound.

To estimate β, we first assume that γ is known and define

$$y^*_{it} := y_{it} - \mathbf{z}_{it}^{\mathrm{T}}\boldsymbol{\gamma}. \tag{4.3.7}$$

Combine equation (4.3.7), (4.3.3) and (4.3.4), we get

$$y^*_{it} \approx \mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}(\mathbf{s}) + \mathbf{x}_{it}^{\mathrm{T}}(\mathbf{s}_i-\mathbf{s})\boldsymbol{\beta}^{(\mathbf{s})} + \mathbf{x}_{it}^{\mathrm{T}}(\tau_t-\tau)\boldsymbol{\beta}^{(\tau)} + \epsilon_{it} \tag{4.3.8}$$

Next we define an objective function as a weighted sum of errors,

$$\sum_{i,t} \left(y_{it}^* - \mathbf{x}_{it}^{\mathrm{T}} \boldsymbol{\beta}(\mathbf{s}) - \mathbf{x}_{it}^{\mathrm{T}} (\mathbf{s}_i - \mathbf{s}) \boldsymbol{\beta}^{(\mathbf{s})} - \mathbf{x}_{it}^{\mathrm{T}} (\tau_t - \tau) \boldsymbol{\beta}^{(\tau)} \right)^2 \times K_{h_s, h_\tau, \mathbf{s}, \tau}(\mathbf{s}_i, \tau_t) \quad (4.3.9)$$

Optimising the objective function with respect to the unknown parameters leads to a classic Weighted Least Square (WLS) type estimator and can be written compactly as follows,

$$\begin{aligned} \tilde{\boldsymbol{\beta}}_{h_s, h_\tau}(\mathbf{s}, \tau) &= a_{k,k+2l} (\mathcal{Q}_{\mathbf{s},\tau}^{\mathrm{T}} \mathcal{W}_{h_s, h_\tau, \mathbf{s}, \tau} \mathcal{Q}_{\mathbf{s},\tau})^{-1} \mathcal{Q}_{\mathbf{s},\tau}^{\mathrm{T}} \mathcal{W}_{h_s, h_\tau, \mathbf{s}, \tau} \mathcal{Y}^* \\ &= a_{k,k+2l} (\mathcal{Q}_{\mathbf{s},\tau}^{\mathrm{T}} \mathcal{W}_{h_s, h_\tau, \mathbf{s}, \tau} \mathcal{Q}_{\mathbf{s},\tau})^{-1} \mathcal{Q}_{\mathbf{s},\tau}^{\mathrm{T}} \mathcal{W}_{h_s, h_\tau, \mathbf{s}, \tau} (\mathcal{Y} - \mathcal{Z}^{\mathrm{T}} \boldsymbol{\gamma}) \quad (4.3.10) \end{aligned}$$

where $a_{k,k+2l} = (\mathbb{I}_k, \mathbf{0}_{k,2l})$, $\mathbb{I}_k$ being a $k \times k$ identity matrix and $\mathbf{0}_{k,2l}$ a $k \times 2l$ matrix with zero entries. The rest is defined as follow, $\mathcal{W}_{a,b,\tau} = \operatorname{diag}\left((K_{h_s, h_\tau, \mathbf{s}, \tau}(\mathbf{s}_i, \tau_t))_{i,t} \right)$, $\mathcal{Y}^* = (y_{it}^*)_{i,t}$, $\mathcal{Y} = (y_{it})_{i,t}$, $\mathcal{Q}_{\mathbf{s},\tau} = (\mathbf{x}_{it}^{\mathrm{T}}, \mathbf{x}_{it}^{\mathrm{T}}(\mathbf{s}_i - \mathbf{s}), \mathbf{x}_{it}^{\mathrm{T}}(\tau_t - \tau))_{i,t}$, $\mathcal{Z} = (\mathbf{z}_{it})_{it}$.

$\tilde{\boldsymbol{\beta}}$ is not a feasible estimator as $\boldsymbol{\gamma}$ is unknown at the moment. To estimate $\boldsymbol{\gamma}$, we substitute equation (4.3.10) into regression (4.3.3),

$$y_{it} = \mathbf{x}_{it}^{\mathrm{T}} \tilde{\boldsymbol{\beta}}_{h_s, h_\tau}(\mathbf{s}_i, \tau_t) + \mathbf{z}_{it}^{\mathrm{T}} \boldsymbol{\gamma} + \epsilon_{it} \quad (4.3.11)$$

Equation (4.3.11) can then be organised into the following matrix format

$$(\mathbb{I}_N - \mathcal{D}_{h_s, h_\tau}) \mathcal{Z} \mathcal{Y} = (\mathbb{I}_N - \mathcal{D}_{h_s, h_\tau}) \mathcal{Z} \boldsymbol{\gamma} + \varepsilon \quad (4.3.12)$$

where $\mathbb{I}_N$ is an $N \times N$ identity matrix, N being the number of observations,

$$\mathcal{D}_{h_s, h_\tau} = \left([\mathbf{z}_{it}^{\mathrm{T}}, \mathbf{0}_{1,2l}] (\mathcal{Q}_{\mathbf{s}_i,\tau_t}^{\mathrm{T}} \mathcal{W}_{h_s, h_\tau, \mathbf{s}_i, \tau_t} \mathcal{Q}_{\mathbf{s}_i,\tau_t})^{-1} \mathcal{Q}_{\mathbf{s}_i,\tau_t}^{\mathrm{T}} \mathcal{W}_{h_s, h_\tau, \mathbf{s}_i, \tau_t} \right)_{i,t} \quad (4.3.13)$$

$\boldsymbol{\gamma}$ is estimated via the least square method,

$$\hat{\boldsymbol{\gamma}}_{h_s, h_\tau} = \left(\mathcal{Z}^{\mathrm{T}} (\mathbb{I}_N - \mathcal{D}_{h_s, h_\tau})^{\mathrm{T}} (\mathbb{I}_N - \mathcal{D}_{h_s, h_\tau}) \mathcal{Z} \right)^{-1} \left(\mathcal{Z}^{\mathrm{T}} (\mathbb{I}_N - \mathcal{D}_{h_s, h_\tau})^{\mathrm{T}} (\mathbb{I}_N - \mathcal{D}_{h_s, h_\tau}) \mathcal{Z} \mathcal{Y} \quad (4.3.14)$$

Finally if we substitute the γ estimator into (4.3.10), a feasible $\boldsymbol{\beta}$ estimator is obtained,

$$\hat{\boldsymbol{\beta}}_{h_s, h_\tau}(\mathbf{s}, \tau) = a_{k,k+2l} (\mathcal{Q}_{\mathbf{s},\tau}^{\mathrm{T}} \mathcal{W}_{h_s, h_\tau, \mathbf{s}, \tau} \mathcal{Q}_{\mathbf{s},\tau})^{-1} \mathcal{Q}_{\mathbf{s},\tau}^{\mathrm{T}} \mathcal{W}_{h_s, h_\tau, \mathbf{s}, \tau} (\mathcal{Y} - \mathcal{Z}^{\mathrm{T}} \hat{\boldsymbol{\gamma}}_{h_s, h_\tau}) \quad (4.3.15)$$

In some cases, we are only interested in estimating the spatially or temporally varying coefficients. We propose the following two estimators aggregated from

(4.3.15),

$$\hat{\boldsymbol{\beta}}_{h_s,h_\tau}(\mathbf{s}) = \frac{1}{T}\sum_{t=1}^{T}\hat{\boldsymbol{\beta}}_{h_s,h_\tau}(\mathbf{s},\tau_t) \qquad \text{(Spatially Varying Coefficient)}$$

$$\hat{\boldsymbol{\beta}}_{h_s,h_\tau}(\tau) = \frac{1}{N}\sum_{i=1}^{N}\hat{\boldsymbol{\beta}}_{h_s,h_\tau}(\mathbf{s}_i,\tau) \qquad \text{(Temporally Varying Coefficient)}$$

4.3.3 Bandwidth Selection

Both $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ estimators rely on two tuning parameters h_s and h_τ, which control the level of smoothing over space and time, respectively. A smaller h_s and h_τ would produce a smoother estimated coefficient surface and hence smaller variance but lead to a higher level of bias. A larger h_s and h_τ, on the other hand, yields less bias but could result in unstable estimates and over-fitting. Suitable h_s and h_τ should strike a balance between the variance and bias.

In this chapter, we use cross-validation (CV) to select the most suitable tuning parameters. The idea of CV is straightforward–the validity of a regression model lies in its ability to fit and predict out-of-sample data, as in-sample fitting statistics can be highly misleading with semi-parametric models. For each tuning parameter value on a predefined parameter grid, CV evaluates the out-of-sample performance (e.g. MSE) of the estimator corresponding to this value. The value with the best out-of-sample performance is chosen to fit the model.

In particular, we will use the $K-$fold cross-validation to split in-sample and out-of-sample. To implement the $K-$fold CV, we first randomly assign observations into K groups, $(G_1,\ldots,G_K)$. For each $(h_{\mathbf{s}}, h_\tau)$ combination, $K-1$ groups of data are used to estimate the model, whereas the remaining group is reserved to evaluate the out-of-sample performance of the estimator given the tuning parameters.

$$\hat{h}_s, \hat{h}_\tau = \frac{1}{N}\operatorname*{arg\,min}_{h_s,h_\tau}\sum_{m=1}^{K}\sum_{it\in G_m}\left(y_{it} - \hat{y}_{it,h_s,h_\tau}^{(-G_m)}\right)^2 \tag{4.3.16}$$

where $\hat{y}_{it,h_s,h_\tau}^{(-G_m)}$ is the fitted value for it in group G_m using model estimated from data excluding G_m.

4.3.4 Simulation

To test the finite sample performance of the proposed estimator against the GTWR in the presence of varying-coefficient only regression, we conduct the following Monte Carlo simulation studies.

We define the space $\mathbf{S}$ as a $[0,1]\times[0,1]$ rectangle and time interval I as $[0,2]$. Three different simulation designs are considered,

$$y_{it} = x_{it}\cdot\beta_1(a_i,b_i,\tau_t)+\varepsilon_{it} \qquad \text{(Design 1)}$$
$$y_{it} = x_{it}\cdot\beta_2(a_i,b_i,\tau_t)+\varepsilon_{it} \qquad \text{(Design 2)}$$
$$y_{it} = x_{it}\cdot\beta_3(a_i,b_i,\tau_t)+\varepsilon_{it} \qquad \text{(Design 3)}$$

where $\beta_1(a_i,b_i,\tau_t)=a_i^2+b_i^2+\cos\tau_t$, $\beta_2(a_i,b_i,\tau_t)=a_i^2+b_i^2$, $\beta_3(a_i,b_i,\tau_t)=a_i^2+b_i^2+\cos\tau_t$. x_{it} is independently drawn from $N(3,1)$. ε_{it} is independently drawn from $N(0,0.5^2)$. Coordinates and time (a_i,b_i,τ_t) are randomly drawn from $\mathbf{S}\times I$.

We consider the following sample size $N=50, T=10$; $N=100, T=20$; $N=200, T=50$, which correspond to small, medium and large sample designs. For each sample size, the coefficients are estimated by both GTWR (outlined in Fotheringham et al., 2015) and our estimator for $R=1000$ replications. To evaluate the performance of each estimator, we compute the mean summed squared error (MSSE) for each case (Ruppert et al., 2003), defined as

$$MSSE(\hat{\beta}_k)=\frac{1}{R\cdot N\cdot T}\sum_{j=1}^{R}\sum_{i=1}^{N}\sum_{t=1}^{T}[\hat{\beta}_k^{(j)}(a_i,b_i,\tau_t)-\beta_k(a_i,b_i,\tau_t)]^2,\quad k=1,2,3 \tag{4.3.17}$$

Note that when both the replication number and sample size is large, MSSE will be made arbitrarily close to the integrated mean square error (IMSE),

$$IMSE(\hat{\beta}_k)=\mathbb{E}\int_a\int_b\int_I[\hat{\beta}_k(a,b,\tau)-\beta_k(a,b,\tau)]^2\mathrm{d}a\mathrm{d}b\mathrm{d}\tau \tag{4.3.18}$$

which is the L_2 measure of the squared error between an functional estimate and its true value.

The Monte Carlo results in Table 4.1 show that all the estimates from the local linear estimator procedure have lower MSSE compared with the GTWR for each simulation design. It suggests that the local linear estimator significantly outperforms the GTWR. This gives us more confidence in the reliability of our estimation strategy.

Estimator	N	T	(Design 1) MSSE(β_1)	(Desgin 2) MSSE(β_2)	(Desgin 3) MSSE(β_3)
Local Linear Estimator	50	10	0.02038*	0.0185*	0.00049*
	100	20	0.0076*	0.0053*	0.000101*
	200	50	0.0028*	0.002*	0.00006*
GTWR	50	10	0.0259	0.0224	0.0034
	100	20	0.0096	0.0067	0.00034
	200	50	0.0038	0.0025	0.000074

Replication:1000. $*$ indicates smaller MSSE between Local Linear Estimator and GTWR for each design.

Table 4.1: STVCR simulation results

4.4 Empirical Application

In this section, we apply the proposed methodology to study the relationship between human activity and the configuration of consumption amenities. Following Glaeser et al. (2001), we argue that a location attracts more footprints if it offers more consumption amenities, has better public transport, road density and so forth. We specify two linear regression models,

$$\ln y_{it} = \ln C_i \beta + \mathbf{z}_i^{\mathrm{T}} \boldsymbol{\theta} + \varepsilon_{it} \qquad \text{(Non-varying coefficient regression)}$$

$$\ln y_{it} = \ln C_i \beta(a_i, b_i, \tau_t) + \mathbf{z}_i^{\mathrm{T}} \boldsymbol{\theta} + \varepsilon_{it} \qquad \text{(Varying coefficient regression)}$$

In both specifications, the dependent variable is human activity level $\ln y_{it}$, measured as the logarithm of the mobile phone positioning records for each areal unit and at each hour $t = 1, \ldots, 24$. The key independent variable is $\ln C_i$, the logarithm of the number of consumption amenities per areal unit. Apart from this, we include a vector of control variables $\mathbf{z}_i$ such as road density, number of subway stations, housing and other amenities location and temporal specific dummy variables. Both β and θ are unknown coefficients.

Two model specifications differ mainly in the way β coefficient is prescribed. β measures the percentage change of human activity as a response to a percentage increase in the number of consumption amenities. In the non-varying coefficient regression, the β coefficient is a scalar, whereas in the varying coefficient regression, the β coefficient is a function of space (a_i, b_i) and time (τ_t), which allows us to model spatial and temporal heterogeneity. For both regression models, we apply them on weekday sub-sample and weekend sub-sample separately to capture the potential weekday-weekend difference.

Variable Name	Definition	Mean	SD	Min	Max
Weekday mobile phone positioning density	Number of mobile phone positioning data per grid point per hour aggregated over weekday	5415	10075	4	225924
Weekend mobile phone positioning density	Number of mobile phone positioning data per grid point per hour aggregated over weekend	5322	9459	4	218127
Consumption amenities	Number of POIs identified as consumption amenities per grid point	75.762	174.954	0	2177
Housing amenities	Number of POIs identified as residential blocks per grid unit	1.372	3.209	0	32
Other amenities	Number of POIs not considered as consumption or housing per grid unit	18.149	41.359	0	546
Subway stations	Subway station counts per grid unit	0.121	0.833	0	42
Road Density	Total road lengths per grid unit (in km)	5.711	4.425	0	23.281
Distance to CBD	Euclidean distance from the centroid of a grid point to the centre of CBD (in km)	22.441	9.000	0	44.926
Hits	Monthly Median "hits" numbers collected by *Meituan-Dianping* for all consumption amenities per grid unit	20.6	1970	0	70442

Table 4.2: Descriptive statistics of key variables

4.4.1 Non-Varying Coefficient Estimates

Table 4.3 reports the estimates of the non-varying coefficient regression. Column (1) (3) (5) are the estimates using the weekday data and column (2) (4) (6) are the estimates using the weekend data. All the specifications are estimated by standard panel data methods (Baltagi, 2008). Due to a lack of temporal variation among the variables on the right-hand side, the fixed effects and the first difference estimators fail to work, whereas the pooled regression, the between estimator and the random effect estimator will yield the same estimates, which the Table 4.3 reports.

Specifications (1) and (2) report the coefficient estimates for consumption amenities, median of hits, residential and other amenities, number of stations, road density and distance to CBD. All the estimated coefficients, apart from the intercept, are statistically significant at 1% level. In the weekday model, the estimated coefficient for consumption amenities is 0.493. The number informs that elasticity of human activity in response to consumption amenities. Here 1% increase in the number of consumption amenities would lead to 0.493% increase in the human activity on average (spatially and temporally). The number is consistent with the empirical evidence from previous studies that suggest higher consumption amenity density is associated with more human activity. The coefficient for 'hits' is positive and significant. The variable is used to control the heterogeneity of consumption amenities. The estimate suggests that consumption amenities that are associated with higher popularity are likely to attract more activity.

For the rest of POIs, we find that the coefficients of other amenities are also positive, though the magnitude is much smaller than consumption amenities. For residential amenity, the estimate is -0.147, which suggests that a higher density of residential complex is associated with lower human activity. The counter-intuitive result is not surprising considering the residential blocks in Bejing, particularly those built after the economic liberalisation are gated communities with strictly controlled entrances. These residential complex are very different from the traditional Beijing residential communities (Hutong) as they are generally situated in less central areas with less integration of consumption amenities. These features discourage social interactions and activities. Finally, we find that better transport accessibility, such as road and subway densities are associated with positive estimates, consistent with the existing literature (Glaeser et al., 2001). Comparing

	(1) lnYd	(2) lnYn	(3) lnYd	(4) lnYn	(5) lnYd	(6) lnYn
Log consumption Amenities	0.493***	0.497***	0.487***	0.491***	0.483***	0.487***
	(0.0106)	(0.0100)	(0.0104)	(0.0099)	(0.0101)	(0.0096)
Median of Hits	0.0135***	0.0136***	0.0129***	0.0130***	0.0137***	0.0139***
	(0.0003)	(0.0004)	(0.0003)	(0.0004)	(0.0003)	(0.0004)
Log Residential Amenities	-0.147***	-0.152***	-0.145***	-0.149***	-0.0866***	-0.0865***
	(0.0111)	(0.0113)	(0.0110)	(0.0111)	(0.0085)	(0.0085)
Log Other Amenities	0.0986***	0.0942***	0.141***	0.137***	0.168***	0.166***
	(0.0036)	(0.0043)	(0.0037)	(0.0048)	(0.0041)	(0.0054)
Transport (No. of Stations)	1.535***	1.543***	1.573***	1.583***	1.050***	1.080***
	(0.0622)	(0.0690)	(0.0637)	(0.0704)	(0.0375)	(0.0476)
Road Density	0.0222***	0.0188***	0.0308***	0.0277***	0.0323***	0.0293***
	(0.0010)	(0.0007)	(0.0013)	(0.0010)	(0.0013)	(0.0010)
Distance to CBD	-0.0519***	-0.0527***	-0.0584***	-0.0595***	-0.0582***	-0.0593***
	(0.0011)	(0.0009)	(0.0013)	(0.0011)	(0.0013)	(0.0011)
Central City			-0.827***	-0.834***	1.291***	1.314***
			(0.0314)	(0.0260)	(0.0494)	(0.0421)
Inner Suburb			-0.368***	-0.383***	0.628***	0.682***
			(0.0173)	(0.0143)	(0.0329)	(0.0335)
Log Consumption* Transport					1.337***	1.159***
					(0.1193)	(0.1047)
Log Consumption* Inner City					-0.378***	-0.385***
					(0.0125)	(0.0102)
Log Consumption* Inner Suburb					-0.227***	-0.243***
					(0.0109)	(0.0106)
Constant	-0.197	-0.153	-0.0743	-0.0264	-0.128	-0.0825
	(0.2141)	(0.2096)	(0.2185)	(0.2200)	(0.2169)	(0.2115)
N	69168	69168	69168	69168	69168	69168

Robust SE in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4.3: Non-varying coefficient regression results

(1) and (2), we find that weekday and weekend sub-samples yield very similar estimates with the same significance, sign and similar magnitudes.

Column (3) and (4) are estimated with additional region-specific dummy variables. We divide metropolitan Beijing into three parts: central city, inner suburb and outer suburb. Dummies for the central city and the inner suburb are included in the regression, whereas the outer suburb serves as a benchmark. We find that key coefficients from (1) and (2) are largely robust to the inclusion of additional dummy variables. Column (5) and (6) look into some interaction effects. We find that the consumption amenities in the outer-suburb tend to have a more substantial impact on attracting human activity than the inner city and the inner suburb.

4.4.2 Spatiotemporally Varying Coefficient Estimates

The panel estimates presented in the previous section suggests that there's a strong partial correlation between the provision of consumption amenities and human activity. The estimates represent an average effect across the spatial and temporal domain. To uncover how the relationship between the two variables evolve over a 24-hour time span, and whether there exists spatial heterogeneity in such a relationship, we turn to the spatiotemporally varying coefficient regression using the proposed methodology. As the estimate of the STVCR is a trivariate function (two-dimensional space and one-dimensional time), it is difficult to visualise the whole surface. We consider the following approaches to visualise the spatial and temporal dimension of the estimate separately. First, the temporal dimension of the varying coefficient is sliced by taking the weighted average of the estimates over some meaningful spatial domain $\mathcal{D}$,

$$\hat{\beta}_{\mathcal{D}}(\tau) = \frac{1}{\#\mathcal{D}} \sum_{(a_i,b_i)\in\mathcal{D}} \hat{\beta}(a_i, b_i, \tau) \tag{4.4.19}$$

where $\#\mathcal{D}$ is the number of cross-sectional units.

Second, the spatial dimension of the varying coefficient is visualised by taking the weighted average of the estimates over some time range $\mathcal{T}$,

$$\hat{\beta}_{\mathcal{T}}(a, b) = \frac{1}{\#\mathcal{T}} \sum_{\tau_t\in\mathcal{T}} \hat{\beta}(a, b, \tau_t) \tag{4.4.20}$$

Figure 4.7 shows the temporal dimension of the STVCR estimate, with $\mathcal{D}$ defined as the (a) whole region, (b) inner suburb, (c) central city, and (4) outer suburb. The solid line in each graph represents the estimate from the weekday sample, whereas the dashed line denotes that from the weekend sample. The first thing we note is that the average estimate strongly represents the pattern in the outer suburb, due to the broad geographical coverage of the region. In both (a) and (d), elasticity drops from $0-5$am and climbs back from $5-10$am and maintains the level until midnight. The elasticity ranges from $0.35 - 0.6$, in line with panel estimates. Turning to the central city and the inner suburb where most economic activities occur, we find that the temporal variation displays different patterns. In the central city, the elasticity ranges from 0.05 to 0.16, the lowest in three regions. The temporal distribution of the elasticity during the day and evening is consistent with humans' living, working and sleeping pattern. However, it is interesting to note that there's a small peak at $2-4$am, which indicates substantial

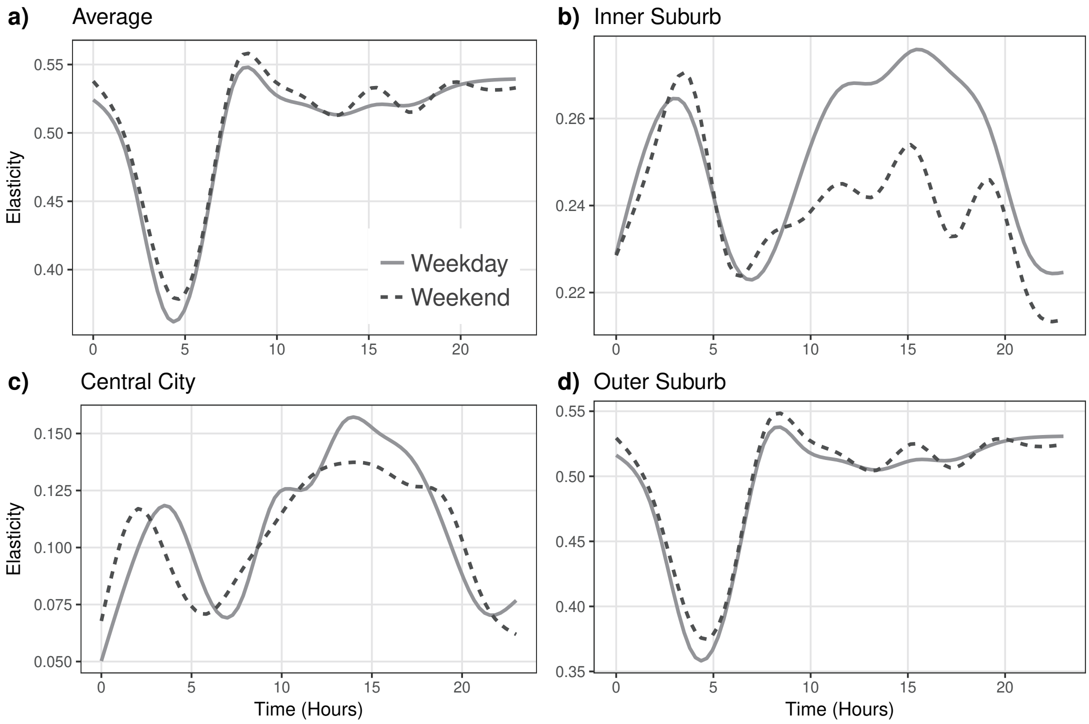

Figure 4.7: Temporal distribution of STVCR estimate

human activity in the early morning. In the inner suburb, the elasticity ranges from 0.2 to 0.28 with a trend similar to that of the central city.

We now turn to the spatial dimension of the STVCR estimate. Figure 4.8 plots the spatial dimension of the STVCR estimate. In the first graph of the upper panel, we show the urban structure of the Beijing metropolitan area. The second and third graph display the temporal average of the elasticity over Beijing with weekday sample and weekend sample, respectively. The lower panel of the figure plots the spatial distribution at hour $0, 8, 12, 20$, respectively. The figure highlights the spatial pattern of the consumption elasticity. The outer suburb such as mountains and green spaces are associated with high elasticity. The central city and the inner suburb have relatively lower elasticity estimates. This is in line with the results from Figure 4.7 and Table 4.3.

On the other hand, we find that the size of elasticity in general increases towards the outer suburb. Figure 4.9 shows the scatter plot between the weekday average elasticity per grid unit and the distance to the CBD. The solid line is the local linear smoothing curve of the scatter plot, whereas the shaded area represents the confidence band. We find that the distance to the CBD and consumption elasticity display a positive relationship. The elasticity is generally lower at locations closer to the city centre and higher at places in the outer suburbs.

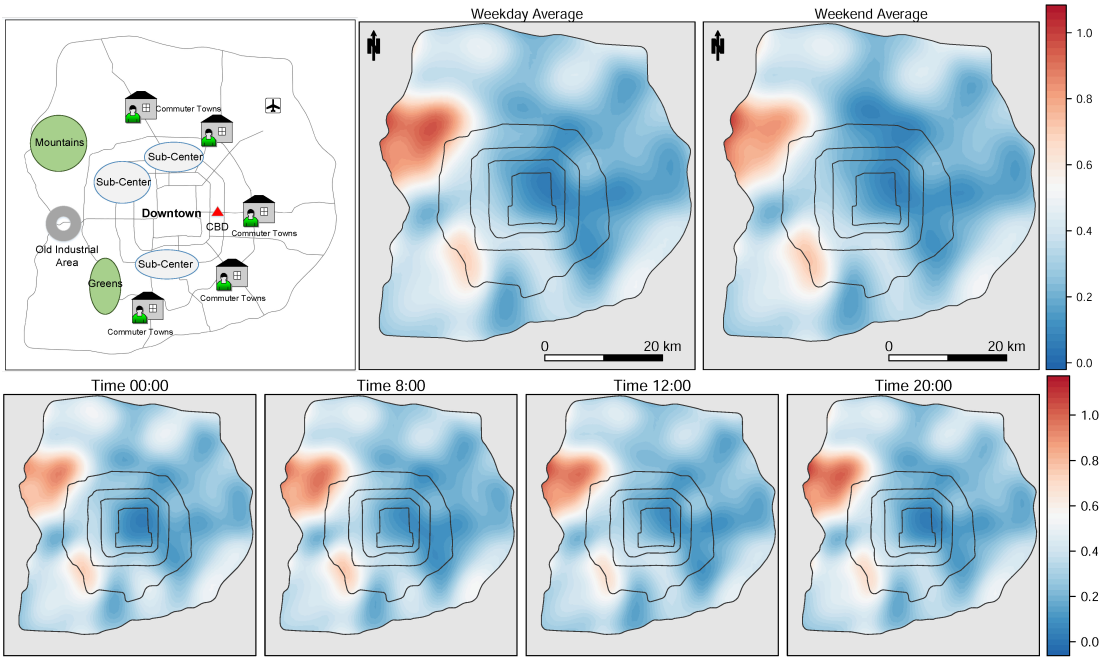

Figure 4.8: Spatial distribution of the STVCR estimate

Overall, we find strong evidence from the STVCR estimates that the interaction of consumption amenities and human activity is highly heterogeneous across time and space. On the one hand, the rise and decline of the consumption elasticity over the temporal dimension are consistent with people's living and resting behaviours across different regions. On the other hand, the size and magnitude of the elasticity vary significantly across space. Places with denser population and economic activity, such as downtown have lower elasticity, whereas less populated areas such as mountains and green spaces have significantly higher elasticity.

4.5 Conclusion

Urban consumption will remain a key driving force of economic growth in fast-urbanising China. In this chapter, we investigate the spatiotemporal relationship between the distribution of consumption amenities and human activity in Beijing. To measure these two key variables, we utilise two unique internet-based datasets, the POI data from *Meituan-Dianping* for consumption amenities and the MPP data from *WeChat* as a proxy for human activity. To estimate the spatiotemporal heterogeneity, we propose a methodology named 'spatiotemporally varying coefficient regression (STVCR)' and conduct a simulation study against geographically and temporally weighted regression (GTWR). The application of the STVCR to

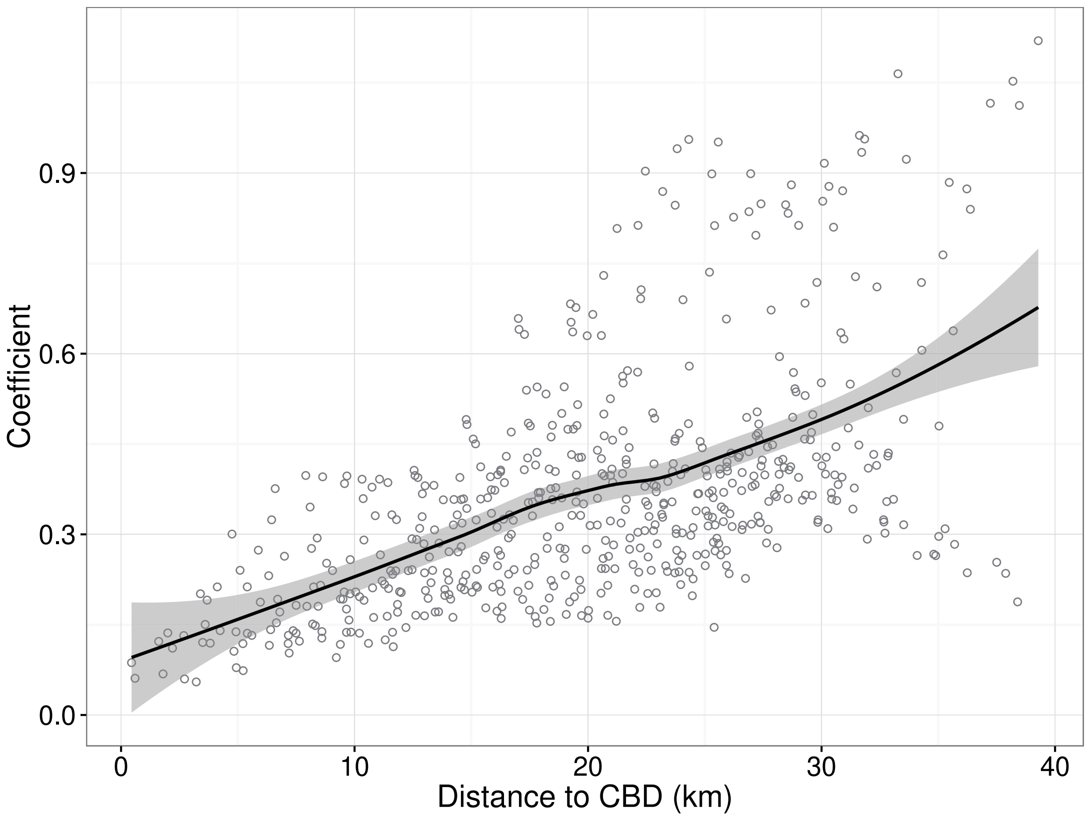

Figure 4.9: Consumption elasticity and distance to CBD

the datasets uncovers substantial heterogeneity across space and time.

In urban economics, the ever-rising unconventional big data allow economists to understand better the structure of a city and how it interacts with the residents (Batty, 2013; Wu, 2016). This study shows the enormous potential of POI and MPP data to proxy consumption amenities and human activity, respectively. Large-scale crowdsourced POI data offer comprehensive, detailed and cheap information about land use, which is particularly useful in studying cities that lack up-to-date granular public records. The eye-opening mobile phone positioning data, on the other hand, offer measures of migration and movement pattern of individuals. The data source underpins many recent studies on western cities (Tranos and Nijkamp, 2015; Jacobs-Crisioni et al., 2014). This chapter represents one of the first attempts to understand urban human activity using MPP data in the context of developing countries.

Methodologically, we propose the spatiotemporally varying coefficient regression to model the relationship between consumption amenities and human activity. The STVCR model allows the regression coefficient to change smoothly with space and time, incorporating spatial and temporal heterogeneity jointly. Conceptually, the proposed regression is similar to the setup of geographically and temporally weighted regression (Fotheringham et al., 2015). However, Fotheringham et al. (2015) applies a local constant estimator which has been pointed out in the statistical literature to have several key disadvantages (Fan and Gijbels, 1996). This point is verified in a simulation study to evaluate the statistical performances of these two approaches. Our methodology provides urban economists with a new statistical framework to go beyond the exploratory analysis of the big spatiotemporal data (see, e.g., Brunsdon et al., 2007; Chen et al., 2011) and to model regression with substantial spatiotemporal uncertainties between the dependent and independent variables.

Our empirical study shows that consumption amenities contribute significantly to the vibrancy of a city. The results confirm Glaeser et al. (2001)'s theory of consumer city that cities attract workers not only for their productive advantage but also for the consumption amenities they offer. We find substantial spatial and temporal heterogeneity of consumption elasticity of human activity. During a 24-hour span, the elasticity exhibits periodic pattern consistent with human's working and resting pattern. Over the spatial dimension, we find that the elasticity tends to be lower in the central city with abundant economic activities and higher in the outer suburb with fewer amenities. The findings provide useful insights into urban planning for fast-growing cities. In particular, an attractive mix of consumption amenities can affect dense and diverse human activity patterns to sustain urban vibrancy. This would, in turn, influence a city's productivity and social capital. A further implication points out the possibility that place-based policies (Barca et al., 2012) relating to the spatial distribution of amenities can generate observable effects on employment (Dalenberg and Partridge, 1995), firm births and agglomeration (Holl, 2004) in excess of city liveliness.

It should be pointed out that there are several limitations associated with this study. First, the analysis in this chapter is conducted on an aggregated spatiotemporal level, rather than on an individual level. We are not able to tell whether an individual shows up in a neighbourhood to engage in consumption or production activity, nor can we tell which property he or she visits. It would be interesting to associate individuals' travelling pattern with socioeconomic backgrounds to gain a deeper understanding of the factors beneath the urban residents travel motivation. Second, we acknowledge that there is a non-trivial endogeneity issue with the regression. For example, neighbourhoods with more consumption amenities tend to attract more residents, whereas more people in the area will bring more consumption amenities. The presence of such endogeneity means that the estimated coefficients should not be interpreted as the causal effects. Future studies should carefully look for exogenous variations of consumption amenities from both spatial and temporal dimension to identify the causal effects from consumption amenity innovations.

Chapter 5

Sieve Continuum GMM Estimation of Functional Linear Regression

5.1 Introduction

The chapter is concerned with the estimation of the slope function in a functional linear regression (FLR), which models the relationship between a scalar response and a functional regressor,

$$Y = \int_0^1 X(s)b(s)\mathrm{d}s + \epsilon, \tag{5.1.1}$$

where Y is a random scalar; $X(\cdot)$ is a square-integrable random function on a compact support $[0,1]$; $b(\cdot)$ is a non-random slope function also defined on $[0,1]$; ϵ is a scalar error term with mean 0 and variance $\sigma^2 > 0$. To simplify the analysis, we assume that both X and Y are centred so that intercept is not required[1]. The domain $[0,1]$ is defined for technical convenience.

As a motivating example, we consider modelling the relationship between annual precipitation and temperature profile. The precipitation is a scalar, denoted by $Prep_i$, whereas the temperature is a trajectory over a year, denoted by $Temp_i(t)$. Note that the annual precipitation is an integration of rainfall over a year,

$$Prep_i = \int_0^{365} Prep_i(t)\mathrm{d}t \tag{5.1.2}$$

[0]Earlier version of the chapter won Excellence Award (1st Prize) in 26^{th} $(EC)^2$ conference on *Theory and Practice of Spatial Econometrics.*

[1]$X = \tilde{X} - \mathbb{E}(\tilde{X})$ and $Y = \tilde{Y} - \mathbb{E}(\tilde{Y})$ where $\tilde{X}, \tilde{Y}$ refer to the uncentered variables.

The rainfall trajectory $Prep_i(t)$, on the other hand, could be modelled concurrently using the temperature of the same time,

$$Prep_i(t) = \alpha(t) + Temp_i(t)b(t) + \varepsilon_i(t), \quad t \in [0, 365] \tag{5.1.3}$$

Combine equation (5.1.2) and (5.1.3), we get a functional linear regression model,

$$Prep_i = \alpha_i + \int_0^{365} Temp_i(t)b(t)\mathrm{d}t + \varepsilon_i \tag{5.1.4}$$

where $\alpha_i = \int_0^{365} \alpha_i(t)\mathrm{d}t$ and $\varepsilon_i = \int_0^{365} \varepsilon_i(t)\mathrm{d}t$.

The functional linear regression was first introduced by Cardot et al. (1999) to model continuous-time processes and has since received a lot of attention in various areas, such as meteorology, medicine, biology, growth studies and so on (see Silverman and Ramsay, 2005). In economics and econometrics, the use of functional linear models is gaining a foothold, in particular with time series and spatial analysis (e.g. Florens and Van Bellegem, 2015; Bhattacharjee et al., 2016a). However, applications are still rare compared with the volume using mainstream econometric methodologies, partly due to a lack of easy-to-implement estimators. Besides, most FLR estimators assume the independence of X and ϵ, which is a strong assumption and unrealistic in many econometric studies. The only exception is Florens and Van Bellegem (2015) which considers the instrumental variable estimation in the functional linear regression context.

Estimating the slope function in a functional linear regression model is particularly challenging because the slope function is determined by an infinite number of parameters as opposed to finite data available. Therefore, regularisation (or dimension reduction) is required to generate meaningful estimates. So far, two general approaches have been proposed in the literature. The first approach is based on the functional principal component analysis (FPCA), also known as the spectral cut-off method (See, e.g., Dauxois et al., 1982; Hall et al., 2006; Hall and Vial, 2006; Hall and Horowitz, 2007; Silverman and Ramsay, 2005). The FPCA has been widely used as an approach to summarise and extract information from functional data (see Silverman and Ramsay, 2005). To estimate a regression slope, the FPCA method projects the infinite-dimensional parameter into a low-dimensional subspace spanned by the eigenfunctions of the empirical covariance operator of X. Hall and Horowitz (2007) shows that the FPCA estimator achieves the optimal rate of convergence under regularity assumptions. A second approach is known as Tikhonov regularisation or ridge regression (Tikhonov, 1963; Hall and

Horowitz, 2007; Florens and Van Bellegem, 2015). Unlike the FPCA, which aims to reduce the regression dimension, the Tikhonov regularisation acknowledges that the major problem of the OLS estimation of the FLR is the dis-continuity of the inverse covariance operator. The method replaces the dis-continuous covariance operator with a continuous albeit deviant operator, which gives rise to stable estimates. For the asymptotic properties of the Tikhonov regularisation, Hall and Horowitz (2005, 2007) prove that it achieves minimax rate of convergence when X is exogenous, and Florens and Van Bellegem (2015) shows the rate when the slope function is estimated by instrumental variables.

In this paper, we propose a sieve continuum generalised method of moments (SCGMM) approach to estimate a functional linear regression with orthogonality conditions from exogenous regressors or instrumental variables. To deal with the dimensionality of the slope function, we replace the infinite parameter space with a sieve parameter space spanned by a finite number of known basis functions. The sieve achieves dimension reduction by introducing some bias into the parameter space. To fully utilise all the moment conditions provided by the model, we follow the idea of the GMM with a continuum of moment conditions (Carrasco and Florens, 2000), which extends GMM's capability to cope with an infinite number of moment conditions. The estimator is produced by minimising the objective function of GMM with a continuum of moments over the sieve parameter space. We study the asymptotic properties of the estimator and derive the upper bound for the integrated mean square error. The rate coincides with the lower bound derived by Chen and Reiss (2011), which indicates that the rate is optimal.

Our study makes the following contributions. First of all, in terms of implementation, the proposed methodology is easy to apply. As the sieve method relies on a series of known basis functions, there is no need to estimate the eigenfunctions, required by the FPCA approach. Compared with the Tikhonov regularisation, which keeps an infinite dimension of the parameter space, our procedure transforms the problem of estimating a function into estimating a vector of the generalised Fourier coefficients, which leads to a closed-form GLS estimator. Furthermore, the method makes it easy to include prior constraints on the slope function, such as monotonicity, convexity, and so forth (Chen, 2007). Second, functional linear regression has been mainly studied with the implicit assumption that the error ϵ is independent of X. In this chapter, we provide a general framework which not only allows for conventional estimation with an exogenous X but also unifies the estimation of the functional instrumental variable regression. Again, the implementation of the IV estimator is more straightforward than

Florens and Van Bellegem (2015)'s Tikhonov regularisation. Third, we establish the asymptotic properties and derive the rate of convergence of our sieve continuum GMM estimator, which ensures the theoretical performance of the proposed methodology. Finally, our research enriches the study of the statistical ill-posed problem and regularisation (Kress et al., 1989; Engl et al., 1996; Carrasco et al., 2007), semi-nonparametric regression, high-dimensional regression (Belloni et al., 2011), high-dimensional IV regression (Ahrens and Bhattacharjee, 2015).

The remaining of the chapter is organised as follows. In section 5.2, we discuss the relationship between the spatial lag model and the functional linear regression. Section 5.3 presents the methodology of the functional linear regression with exogenous regressors. Section 5.4 studies the functional instrumental instrumental variable regression. Section 5.5 discusses the assumptions imposed on the model and the rate of convergence of the estimators. Section 5.6 concludes.

5.2 Functional Linear Regression and Spatial Lag Model

In this section, we discuss how a spatial lag model is related to a functional linear regression model, following Bhattacharjee et al. (2016a). To illustrate the idea, let s_1 and s_2 be two locations where we observe scalar independent variable x_1, x_2 and scalar dependent variable y_1, y_2. Assume the effect of x on y is spatially heterogeneous, a spatial lag model can be specified as follows,

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \rho W \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \begin{pmatrix} x_1\beta_1 \\ x_2\beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} \tag{5.2.5}$$

where W is a pre-determined spatial weights matrix which measures the spillover effect of y from one location to another. Let $M = (I - \rho W)^{-1}$, I being a 2×2 identity matrix, the spatial lag model is reduced to

$$\begin{aligned} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} &= M \begin{pmatrix} x_1\beta_1 \\ x_2\beta_2 \end{pmatrix} + M \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} \\ &= \begin{pmatrix} \sum_{j=1}^2 m_{1,j}x_j\beta_j \\ \sum_{j=1}^2 m_{2,j}x_j\beta_j \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \end{aligned} \tag{5.2.6}$$

where $m_{i,j}$ represents the (i, j)th element of matrix M and $[u_1, u_2]^{\mathrm{T}} = M[\epsilon_1, \epsilon_2]^{\mathrm{T}}$. Equation (5.2.6) suggests that a spatial lag model with heterogeneous effects could be reduced to a regression on a collection of the weighted $x_j\beta_j$ at the neighbourhood locations. Moreover, the model could be easily extended to n locations,

$$y_i = \sum_{j=1}^{n} m_{i,j} x_j \beta_j + u_j, \quad j = 1, \ldots, n. \tag{5.2.7}$$

If we treat the space as a continuum $\mathcal{D} \subset \mathbb{R}^2$ where m, x, β are defined, then the model (5.2.7) is turned into

$$y_i = \int_{\mathcal{D}} m_i(s) x_i(s) \beta(s) \mathrm{d}s + u_i, \quad j = 1, \ldots, n. \tag{5.2.8}$$

Note that equation (5.2.8) now forms a functional linear regression with $m_i x_i$ serving as a functional regressor, and β as a functional coefficient. The model simultaneously captures spatial autocorrelation by including an entire weighted surface of x into the regression, and spatial heterogeneity through the varying coefficient β.

5.3 Functional Linear Regression with Exogenous Regressor

We now move to the methodology of functional linear regression. First, consider a typical functional linear regression along with the assumption that X and ϵ are not correlated.

Assumption 5.3.1 (Exogeneity)**.**

$$\mathbb{E}[X(t)\epsilon] = 0, \quad \forall t \in [0, 1]. \tag{5.3.9}$$

The assumption is a form of the unconditional moment condition and is weaker than the conditional moment restriction $\mathbb{E}[\epsilon|X] = 0$, commonly used in the FLR literature. Also note that equation (5.3.9) is essentially a continuum of moment conditions (Carrasco and Florens, 2000), as it is defined for each value in $[0, 1]$.

The functional linear regression equation (5.1.1) and unconditional moment condition equation (5.3.9) imply that

$$\mathbb{E}[X(t)Y] = \int_0^1 \mathbb{E}[X(s)X(t)] b(s) \mathrm{d}s, \quad \forall t \in [0, 1], \tag{5.3.10}$$

which represents an infinite-dimensional version of the multivariate least square normal equation. Equation (5.3.10) could be written compactly as

$$Kb = g, \tag{5.3.11}$$

where K is a covariance operator, defined as $(Kf)(t) = \int_0^1 \mathbb{E}[X(s)X(t)]f(s)\mathrm{d}s$ for some function f, and $g(t) = \mathbb{E}[X(t)Y]$.

Equation (5.3.10) is known in the literature as the Fredholm integral equation of the first kind (Engl et al., 1996; Kress et al., 1989), where the value of the equation g and the linear integral operator K are estimable but the argument b is unknown. Intuitively, if we could find an estimator for the inverse of K, denoted by K^{-1}, b could be recovered. The direct inverse of an empirical version of K often leads to problems as neither K^{-1} nor the empirical counterpart are continuous operators (Kress et al., 1989). This means a small perturbation of g would be translated into considerable variation in b through K^{-1}.

To better understand the challenge, we consider a Hilbert space $\mathcal{H}$ of square integrable functions defined on a compact support $[0, 1]$

$$\mathcal{H} = \left\{ f(s), s \in [0,1] : \int_0^1 |f(s)|^2 \mathrm{d}s < \infty \right\}. \tag{5.3.12}$$

As a Hilbert space, $\mathcal{H}$ is equipped with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ defined as

$$\langle f, g \rangle_{\mathcal{H}} = \int_0^1 f(u)g(u)\mathrm{d}u, \quad \forall f, g \in \mathcal{H}. \tag{5.3.13}$$

The inner product induces a norm on $\mathcal{H}$, defined as

$$\|f\|_{\mathcal{H}} = \langle f, f \rangle_{\mathcal{H}}^{1/2}, \quad \forall f \in \mathcal{H}. \tag{5.3.14}$$

Let $\phi_1, \phi_2, \ldots$ be a sequence of orthonormal basis functions in $\mathcal{H}$, such that $\langle \phi_j, \phi_k \rangle_{\mathcal{H}} = 0$ for all $j, k \in \mathbb{N}$, $j \neq k$ and $\langle \phi_j, \phi_j \rangle_{\mathcal{H}} = 1$ for all $j \in \mathbb{N}$. Any function $f \in \mathcal{H}$ could be represented in terms of the basis functions,

$$f = \sum_{j=1}^{\infty} f_j \phi_j, \quad f_j = \langle f, \phi_j \rangle_{\mathcal{H}} \tag{5.3.15}$$

Using the terminology defined above, it is noted that the functional linear regression could be transformed into a linear regression with an infinite countable number of regressors and coefficients,

$$Y = \int_0^1 X(s)b(s)\mathrm{d}s + \epsilon = \sum_{l=1}^{\infty} b_l \chi_l + \epsilon, \tag{5.3.16}$$

where $b_l = \langle b, \phi_l \rangle_{\mathcal{H}}$, and $\chi_l = \langle X, \phi_l \rangle_{\mathcal{H}}$. The transformed regression illustrates the problem from the perspective of the multiple linear regression– the number of the unknown regression coefficients $b_1, b_2, \ldots$ is infinite but the size of the data is always finite. As a result, the rank of the data matrix will always be less than the dimension of the coefficients, violating the rank condition in the classical Gauss-Markov theorem.

To salvage the rank condition, we turn our attention to a sequence of sieve spaces that are less complex. For a given orthonormal basis function series $\{\phi_j, j = 1, 2, \ldots\}$, define a subspace $\mathcal{H}_n$ of $\mathcal{H}$ spanned by $\phi_1, \ldots, \phi_{k_n}$,

$$\mathcal{H}_n = \left\{ f \in \mathcal{H} : f = \sum_{j=1}^{k_n} a_j \phi_j, \quad a_1, \ldots, a_{k_n} \in \mathbb{R} \right\} \tag{5.3.17}$$

where $k_n \geq 1$ is a finite positive integer, non-decreasing with n. In addition, $k_n \to \infty$ as $n \to \infty$. $\{\mathcal{H}_n\}$ sequence has the following properties,

1. for a given n, the dimension of the space $\mathcal{H}_n$ is k_n, therefore finite;
2. $\{\mathcal{H}_n\}$ is increasing, $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \cdots \subset \mathcal{H}_n \subset \cdots \subset \mathcal{H}$;
3. $\{\mathcal{H}_n\}$ are dense in $\mathcal{H}$, in the sense that any function in $\mathcal{H}$ can be approximated by a function in $\mathcal{H}_n$ arbitrary well with a sufficiently large n. Furthermore, $\mathcal{H} = \lim_{n\to\infty} \cup_{j=1}^{k_n} \mathcal{H}_j$.

If we search for estimators in a sieve parameter space, then only a series of the generalised Fourier coefficients need to be estimated, which is feasible given finite available data.

Let $b_n = \sum_{j=1}^{k_n} b_j \phi_j$ and substitute b_n into (5.3.10)

$$\begin{aligned}
\mathbb{E}[X(t)Y] &= \int_0^1 \mathbb{E}[X(t)X(s)] \sum_{j=1}^{k_n} b_j \phi_j(s) \mathrm{d}s, \quad t \in [0,1] \\
&= \sum_{j=1}^{k_n} b_j \mathbb{E}\left[X(t) \int_0^1 X(s)\phi_j(s)\mathrm{d}s \right] \\
&= \sum_{j=1}^{k_n} b_j \mathbb{E}[X(t)\chi_j].
\end{aligned} \tag{5.3.18}$$

The new normal equation involves k_n parameters, determined by a continuum of moment conditions.

Let (X_i, Y_i), $i = 1, \ldots, n$ be an independent and identically distributed sample following the same joint distribution as (X, Y). Following Carrasco and Florens

(2000), the finite parameters $b_1, \ldots, b_{k_n}$ determined through (5.3.18) could be estimated by minimising the L_2 norm of the continuum of the moment conditions.

Definition 5.3.2 (Sieve Continuum GMM Estimator for FLR-EXO)**.**

$$\hat{b}_n = \arg\min_{b \in \Theta_n} \|\hat{K}b - \hat{g}\|^2_{\mathcal{H}} \tag{5.3.19}$$

where $(\hat{K}b)(t) = \int_0^1 \frac{1}{n}\sum_{i=1}^n X_i(t)\langle X_i, b\rangle_{\mathcal{H}}$, $\hat{g}(t) = \frac{1}{n}\sum_{i=1}^n X_i(t)Y_i$ *and* Θ_n *is the parameter space spanned by* $\phi_1, \ldots, \phi_{k_n}$ *with additional smoothness conditions stated in assumption (5.5.4).*

The estimator has a closed form,

$$\hat{b}_n = \boldsymbol{\phi}_n^{\mathrm{T}}(\mathcal{X}_n^{\mathrm{T}}\mathcal{W}_n\mathcal{X}_n)^{-1}\mathcal{X}_n^{\mathrm{T}}\mathcal{W}_n\mathcal{Y}_n, \tag{5.3.20}$$

where $\boldsymbol{\phi}_n = (\phi_1, \ldots, \phi_{k_n})^T$, $\mathcal{X}_n$ is a $n \times k_n$ matrix with $(i,j)th$ component $\langle X_i, \phi_j\rangle_{\mathcal{H}}$; $\mathcal{W}_n$ is a $n \times n$ matrix with (i,j)th element $\langle X_i, X_j\rangle_{\mathcal{H}}$, and $\mathcal{Y}_n = (Y_1, \ldots, Y_n)^{\mathrm{T}}$.

Proof. See Appendix. □

The estimator is composed of two parts, the first part $\boldsymbol{\phi}_n = (\phi_1, \ldots, \phi_{k_n})^T$ is the vector of basis functions used to estimate the slope function, and the second part $(\mathcal{X}_n^{\mathrm{T}}\mathcal{W}_n\mathcal{X}_n)^{-1}\mathcal{X}_n^{\mathrm{T}}\mathcal{W}_n\mathcal{Y}_n$ is the estimate of the generalised Fourier coefficients attached to the basis functions. Once the choice of the basis functions is fixed, the remaining task is simply to obtain a vector of finite coefficients, in the form of a generalised least squares.

The estimator could also be derived from the perspective of solving the linear integral equation (5.3.18). Define operator K_n as

$$(K_n f)(u) = \sum_{j=1}^{k_n} f_j \mathbb{E}[X(t)\chi_j], \tag{5.3.21}$$

where $f_j = \langle f, \phi_j\rangle_{\mathcal{H}}$ and $\chi_j = \langle X, \phi_j\rangle$. Also define the adjoint operator of K_n, denoted by K_n^* as

$$(K_n^* f)(t) = \sum_{j=1}^{k_n} \mathbb{E}[\chi_j\langle X, f\rangle]\phi_j(t). \tag{5.3.22}$$

Note that $Kb_n = K_n b_n$, therefore, equation (5.3.18) can be expressed as $K_n b_n = g$. Apply K_n^* on both sides of $K_n b_n = g$, which yields

$$K_n^* K_n b_n = K_n^* g \tag{5.3.23}$$

As $K_n^* K_n$ is a positive definite operator, the inverse exists, then b_n is solved by

$$b_n = (K_n^* K_n)^{-1} K_n^* g \tag{5.3.24}$$

A natural estimator of b_n is produced by replacing K_n, K_n^*, g with the empirical version,

$$\tilde{b}_n = (\hat{K}_n^* \hat{K}_n)^{-1} \hat{K}_n^* \hat{g} \tag{5.3.25}$$

where $(\hat{K}_n f)(u) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k_n} f_j X_i(u)\chi_{i,j}$, $(\hat{K}_n^* f)(u) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k_n} \chi_{i,j}\langle X_i, f\rangle \phi_j(u)$, $\hat{g} = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$.

Proposition 5.3.3. $\tilde{b}_n = \hat{b}_n = \boldsymbol{\phi}_n^{\mathrm{T}}(\mathcal{X}_n^{\mathrm{T}}\mathcal{W}_n\mathcal{X}_n)^{-1}\mathcal{X}_n^{\mathrm{T}}\mathcal{W}_n\mathcal{Y}_n$

Proof. See Appendix. □

While the proposed estimator is derived under the identity weighting operator, alternative weighting schemes are allowed for the continuum GMM.

$$\check{b}_n = \arg\min_{b_n} \int_0^1 \int_0^1 \left[(\hat{K}_n b_n)(t) - \hat{g}(t)\right] \omega_n(s,t) \left[(\hat{K}_n b_n)(s) - \hat{g}(s)\right] \mathrm{d}s\mathrm{d}t \tag{5.3.26}$$

where $w_n(s,t)$ is a continuous GMM weighting function defined on $[0,1]\times[0,1]$. The generalised estimator is then modified as

$$\check{b}_n = \boldsymbol{\phi}_n^{\mathrm{T}}(\mathcal{X}_n^{\mathrm{T}}\check{\mathcal{W}}_n\mathcal{X}_n)^{-1}\mathcal{X}_n^{\mathrm{T}}\check{\mathcal{W}}_n\mathcal{Y}_n \tag{5.3.27}$$

where $\check{\mathcal{W}}$ is $n \times n$ matrix with $(i,j)th$ element being $\int_0^1\int_0^1 X_i(s)\omega_n(s,t)X_j(t)\mathrm{d}s\mathrm{d}t$.

With a conventional GMM, the most efficient estimator is the one weighted by the inverse covariance of the moment condition (Hansen, 1982). With a functional linear regression, the inverse covariance of the continuum of moment conditions is once again ill-posed. To avoid additional regularisation and added complexity to the estimator, we will mainly study the identity weighting scheme.

Though the analytic form of our estimator is similar to Carrasco and Florens (2000), the implications of these two estimators diverge. In Carrasco and Florens (2000), the parameter space is a finite-dimensional Euclidean space where coefficients reside. As the sample size goes to infinity, the parameter space remains the same, and $\sqrt{n}$ consistency can be achieved. In our model, the true parameter space is infinite-dimensional, whereas the sieve parameter depends on the sample size. As the sample size increases, the complexity of the sieve parameter space increases as well. As a result, the rate of convergence will be strictly slower than $\sqrt{n}$.

5.4 Functional Instrumental Variable Regression

In this section, we consider the situation where the exogeneity assumption does not hold,

$$\mathbb{E}[X(s)\epsilon] \neq 0, \quad \forall s \in [0,1]. \tag{5.4.28}$$

Such 'endogeneity' problem often arises as a result of omitted variables, reverse causality, measurement error etc. Without the exogeneity condition, the slope function is not identified, and the proposed sieve estimator in the previous section will not be consistent. In the econometric literature, identification without exogeneity could be achieved through other channels. For example, if there exist some 'instrumental variables' that are not correlated with error, and sufficiently correlated with the endogenous regressor, then coefficient could still be identified and estimated. A classic case of the endogeneity issue comes from estimating the impact of education (X) on salary (Y) (Mincer, 1958). As education and salary levels are both correlated with one's ability (Spence, 1978), failure to include the ability variable would bias the estimator from the usual regression procedure. However, if the variation of education could be induced by a third instrumental variable (Z), correlated only with education, but not with other factors that may link to salary, then the effect of education could still be consistently estimated.

Here we follow the same logic and assume there exists an instrumental function $Z(\cdot)$, such that Z and ϵ are not correlated (exogeneity). Besides, Z and X are sufficiently correlated (relevance). To make our results general enough, we assume that Z is an element in a Hilbert space $\mathcal{D}$ of square integrable functions defined on a domain D. Let $\langle \cdot, \cdot \rangle_{\mathcal{D}}$ denote the inner product function on $\mathcal{D}$. For example, D could be defined on a compact real interval $[a,b]$, where $a < b$, and $a, b \in \mathbb{R}$. In this case, we could define $\langle f, g \rangle_{\mathcal{D}} = \int_a^b f(s)g(s)\mathrm{d}s$. When $Z = X$ and $D = [0,1]$, the functional IV regression nests the functional linear regression with exogenous regressors as a special case. D could also be defined on some product spaces, e.g., $[a,b]^p$, p being an positive integer.

Assumption 5.4.1 (IV Exogeneity)**.**

$$\mathbb{E}[Z(s)\epsilon] = 0, \quad s \in D. \tag{5.4.29}$$

The functional linear regression (5.1.1) and the IV exogeneity assumption (5.4.29) imply that

$$\mathbb{E}[Z(t)Y] = \int_0^1 \mathbb{E}[Z(t)X(s)]b(s)\mathrm{d}s, \quad t \in D \tag{5.4.30}$$

which we write compactly as

$$Tb = h, \tag{5.4.31}$$

where $(Tf)(t) = \int_0^1 \mathbb{E}[Z(t)X(s)]f(s)\mathrm{d}s$ and $h(t) = \mathbb{E}[Z(t)Y], t \in D$.

The normal equation implied by the IV exogeneity condition is also a Fredholm equation of the first kind. Therefore, the inverse of T will not be continuous and requires regularisation. A major difference between T and K is that K is a self-adjoint covariance operator of X, whereas T is the cross-covariance between X and Z, which is not self-adjoint unless $Z = X$.

Definition 5.4.2 (Sieve Continuum GMM Estimator for FLR-IV)**.**

$$\hat{b}_{iv,n} = \arg\min_{b \in \Theta_n^{IV}} \|\hat{T}b - \hat{h}\|_{\mathcal{D}} \tag{5.4.32}$$

where $\hat{h} = \frac{1}{n}\sum_{i=1}^n Z_i Y_i$*, and* $\hat{T}b = \frac{1}{n}\sum_{i=1}^n Z_i \langle X_i, b\rangle$*;* Θ_n *is the sieve parameter space spanned by* $\phi_1, \ldots, \phi_{k_n}$ *with functions satisfying smoothness conditions stated in assumption (5.5.4).*

The estimator also has a closed form

$$\hat{b}_n^{IV} = \boldsymbol{\phi}_n^{\mathrm{T}}(\mathcal{X}_n^{\mathrm{T}}\mathcal{Z}_n\mathcal{X}_n)^{-1}\mathcal{X}_n^{\mathrm{T}}\mathcal{Z}_n\mathcal{Y}_n \tag{5.4.33}$$

where $\boldsymbol{\phi}_n = (\phi_1, \ldots, \phi_{k_n})^T$, $\mathcal{X}_n$ is an $n \times k_n$ matrix with the (i,j)th component $\langle X_i, \phi_j\rangle_{\mathcal{H}}$; $\mathcal{Z}_n$ is an $n \times n$ matrix with (i,j)th element $\langle Z_i, Z_j\rangle_{\mathcal{D}}$ and $\mathcal{Y}_n = (Y_1, \ldots, Y_n)^{\mathrm{T}}$.

5.5 Rate of Convergence

In this section, we explore the rate of convergence of the proposed Sieve Continuum GMM estimators. Before laying out the assumptions for convergence results, we define the following notations.

For a function $f \in \mathcal{H}$ and non-negative integer k, define the differential operator $D_k f$ as

$$D_k f = \frac{\partial^k f(x)}{\partial x^k}, \text{ if } k > 0 \text{ and } D_0 f = f. \tag{5.5.34}$$

Given a positive integer m, define the Sobolev norm on $\mathcal{H}$ as

$$\|f\|_m = \left\{\sum_{j=0}^m \|D_j f\|_{\mathcal{H}}^2\right\}^{1/2}. \tag{5.5.35}$$

Let $\mathcal{H}_k$ be a subspace of $\mathcal{H}$ with Sobolev norm no bigger than some constant C,

$$\mathcal{H}_k = \{f \in \mathcal{H}, \|f\|_k \leq C\}. \tag{5.5.36}$$

Let $\Lambda(p, C_0)$ be a function subspace of $\mathcal{H} \otimes \mathcal{H}$ such that a bivariate function $w \in \Lambda(p, C_0)$ if and only if there exists a constant C_0 and a positive integer p such that

$$\sup_{j \leq p} \sup_{s,t \in [0,1]} |D_j f(s,t)| \leq C_0 \text{ and } \sup_{s_1,t_1,s_2,t_2 \in [0,1]} \frac{|D_p f(s_1,t_1) - D_p(s_2,t_2)|}{\|(s_1,t_1) - (s_2,t_2)\|} \leq C_0. \tag{5.5.37}$$

The assumptions for the Sieve Continuum GMM estimator for FLR-EXO are as follows.

Assumption 5.5.1. *Data (X_i, Y_i) are i.i.d and follow the same joint distribution as (X, Y).*

Assumption 5.5.2. *There exist finite constants C_2, C_3 such that*

$$\mathbb{E}Y^2 < C_2, \quad \mathbb{E}\|X\|^2 < C_3 \tag{5.5.38}$$

Assumption (5.5.1) and (5.5.2) are distributional assumptions on X and Y, in line with most of the functional linear regression literature (e.g. Cai et al., 2006; Hall and Horowitz, 2007). Assumption (5.5.1) requires (X_i, Y_i) to be independent and identically distributed, defined for technical convenience. As the proposed estimator is a type of the generalised method of moments, it is expected to work with less stringent assumptions such as independent but non-identical distributions. Assumption (5.5.2) requires the second moment of X and Y are finite. This assumption ensures the sum of the eigenvalues of the covariance operator does not diverge. This condition normally holds in practice.

Assumption 5.5.3. *There exists a constant C_1 and a positive integer r such that*

$$k(s,t) := \langle \mathbb{E}[X(s)X], \mathbb{E}[XX(t)] \rangle_{\mathcal{H}} \in \Lambda(r, C_1) \tag{5.5.39}$$

Assumption (5.5.3) defines the smoothness of the bivariate function $k(s,t)$, the kernel of the operator K^*K, in terms of the parameter r. It assumes that the derivatives of the k function up to r times be uniformly bounded and rth order derivative of k is Lipschitz continuous. The assumption requires that the k function to be a smooth function. If k is not first-order differentiable, then the assumption does not hold. The parameter r informs the number of derivatives that the kernel of the operator possess. The larger the r, the higher level of smoothness the operator has.

Assumption 5.5.4. *The sieve parameter space is*

$$\Theta_n = \mathcal{H}_n \cap \mathcal{H}_s = \left\{ h = \sum_{j=1}^{k_n} h_j \phi_j, \|h\|_s < C \right\} \tag{5.5.40}$$

for some $s \geq 2$.

Assumption 5.5.5. *With the sieve parameter space assumed in (5.5.4), a unique minimiser exists for the objective function (5.3.2).*

Assumption (5.5.4) and (5.5.5) define the sieve parameter space over which the objective function is optimised. Following Horowitz (2012b), we not only assume that the sieve parameter space is spanned by a finite number of known basis functions, but also require some level of smoothness. In cases where the constraint is bounded, the estimator will not have a closed-form, and it is very challenging to derive the asymptotic properties, suggested by Horowitz (2012b). This assumption stipulates the parameters that the methodology is able to estimate–functions that are at least twice continuously differentiable.

Assumption 5.5.6. *Define the sieve-measure of ill-posedness as*

$$\tau_n = \sup_{f \in \Theta_n} \frac{\|f\|}{\|(K^*K)f\|}, \tag{5.5.41}$$

We assume

$$\tau_n = O(k_n^r), \quad \text{and} \quad \tau_n \sup_{f \in \Theta_n} \frac{\|(K_n^* K_n - K^*K)f\|}{\|f\|} = O(k_n^{-s}) \tag{5.5.42}$$

Assumption (5.5.6) is related to the 'sieve-measure of ill-posedness', a concept introduced by Blundell et al. (2007). The measure quantifies the ill-posedness of the operator K^*K for a given choice of sieve space. Blundell et al. (2007) shows that τ_n is closely related to $\lambda_{k_n}^2$, the k_n-th eigenvalue of K^*K. As n and Θ_n grow, τ_n increases as well. The assumption $\tau_n = O(k_n^r)$ requires the τ_n to grow at a polynomial rate. Conditions are provided in Blundell et al. (2007) and Chen and Reiss (2011) under which the assumption holds. Similar assumptions are imposed on PCA-based estimators (Hall and Horowitz, 2007) where the speed at which the eigenvalue sequence decreases has a lower polynomial bound. The second part of the assumption is a smoothness restriction placed on the kernel of K^*K. It makes sure that $K_n^*K_n$ approximates K^*K sufficiently well in the sieve space. Similar assumptions are applied by Horowitz (2011) and Horowitz (2012b).

Assumption 5.5.7.

$$\|K_n^*K_n - K^*K\| = O(k_n^{-r}) \tag{5.5.43}$$

Assumption (5.5.7) is a complement of the second part of the assumption (5.5.6), which sets the approximation error bound for $K_n^*K_n$ on a larger set.

Assumption 5.5.8. *There exists a $s \geq 2$, such that (1) $b \in \mathcal{H}_s$, (2)$K^*g \in \mathcal{H}_{s+r}$, (3) for any $f \in \mathcal{H}_l$, there are coefficients f_j, $j = 1, 2, \ldots$ such that*

$$\left\| f - \sum_{j=1}^{m} f_j \phi_j \right\| \leq O(m^{-l}) \tag{5.5.44}$$

Assumption (5.5.8) is the smoothness conditions imposed on the slope function b and the K^*g, and the rate at which they can be approximated by the basis functions. The assumption is generally satisfied with the commonly used basis function system such as spline, polynomial and trigonometric bases.

Assumption 5.5.9. *K^*K is non-singular.*

Assumption (5.5.9) is concerned with the identification of the slope function. This assumption ensures that K^*K is invertible, therefore b can be solved by $b = (K^*K)^{-1}K^*g$.

Theorem 5.5.10 (Rate of Convergence for the sieve estimator of FLR-EXO)**.**

$$\|\hat{b} - b\| = O_p(k_n^{-s} + k_n^r \sqrt{k_n/n}) \tag{5.5.45}$$

Under assumption (5.3.9) (5.5.1) (5.5.3) (5.5.2) (5.5.4) (5.5.5) (5.5.6) (5.5.8) 5.5.7). If $k_n \asymp n^{1/(2s+2r+1)}$, the fastest rate of convergence is achieved as

$$\|\hat{b}_n - b\| = O_p(n^{-s/(2s+2r+1)}) \tag{5.5.46}$$

Proof. See Appendix. □

For the functional instrumental variable regression, the following additional assumptions are required.

Assumption 5.5.11. *Data (X_i, Y_i, Z_i) are i.i.d and follow the same joint distribution as (X, Y, Z).*

Assumption 5.5.12. *There exists a constant $\tilde{C}_1$ and a positive integer m such that*

$$\langle \mathbb{E}[X(s)Z], \mathbb{E}[ZX(t)] \rangle_{\mathcal{H}} \in \Lambda(m, \tilde{C}_1) \tag{5.5.47}$$

Assumption 5.5.13. *There exists a finite constant C_4 such that*

$$\mathbb{E}\|Z\|^2 < C_4 \tag{5.5.48}$$

Assumption 5.5.14. *Define the sieve-measure of ill-posedness as*

$$\delta_n = \sup_{f \in \Theta_n} \frac{\|f\|}{\|(T^*T)f\|}, \tag{5.5.49}$$

$$\delta_n = O(k_n^m), \quad \text{and} \quad \delta_n \sup_{f \in \Theta_n} \frac{\|(T_n^*T_n - T^*T)f\|}{\|f\|} = O(k_n^{-s}) \tag{5.5.50}$$

Assumption 5.5.15. $T^*g \in \mathcal{H}_{s+m}$.

Assumption 5.5.16. *T^*T is non-singular.*

Assumption 5.5.17.

$$\|T_n^*T_n - T^*T\| = O(k_n^{-m}) \tag{5.5.51}$$

Theorem 5.5.18 (Functional IV Estimator Rate of Convergence)**.**

$$\|\hat{b} - b\| = O_p(k_n^{-s} + k_n^m \sqrt{k_n/n}) \tag{5.5.52}$$

Under assumption (5.5.11) (5.5.12) (5.5.13) (5.5.14) (5.5.15) (5.5.16) (5.5.17) (5.4.29) (5.5.2) (5.5.4) (5.5.8). If $k_n \asymp n^{1/(2s+2m+1)}$, the fastest rate of convergence is achieved as

$$\|\hat{b}_{iv,n} - b\| = O_p(n^{-s/(2s+2m+1)}). \tag{5.5.53}$$

.

Proof. See Appendix □

The approaches for proving these two theorems are similar. Here we outline the strategy using the functional linear regression with exogeneity as an example.

1. The error between $\hat{b}_n$ and true b is decomposed into two parts, the estimation error between $\hat{b}_n$ and b_n, the approximation error between b_n and b, the bound of which is set in the assumption (5.5.8).

2. The error between $\hat{b}_n$ and b_n is translated into the error between $K^*K\hat{b}_n - K^*Kb_n$ through the 'sieve measure of ill-posedness' assumption (5.5.6).

3. The error between $K^*K\hat{b}_n$ and K^*Kb_n can be further decomposed into three parts involving the error between K^*Kb_n and K^*Kb, the error between $\hat{K}^*\hat{g}$ and K^*g, the error between $K^*K\hat{b}_n - \hat{K}_n^*\hat{K}_n\hat{b}_n$, each of which is bounded separately using the assumptions.

4. Combine all the error bounds together via the triangle inequality to get the rate of convergence.

The rate of convergence reveals two aspects of the functional linear estimators. First of all, the rate is by construction strictly slower than $\sqrt{n}$. This is generally in line with the results in the nonparametric literature. Second, the rate of convergence for both estimators suggest that the speed depends on two parameters. The smoothness of the slope function (s), and the relative smoothness of the operator (r or m). A smoother slope function leads to a faster rate of convergence, as it is easier to infer the shape of the function when the slope function is smooth. On the contrary, the smoother the operator, the slower the rate of convergence. This is because the smoothness of the operator is synonymous to the level of variation in X. A smooth operator means there is less information in the data and therefore more difficult to estimate the coefficient. For the functional instrumental variable regression, the parameter m informs the size of the cross-covariance between X and Z. A large m means that the instrument Z contains less information regarding X, which exacerbates the results.

5.6 Conclusion

In this chapter, we propose a class of sieve estimators for the functional linear regression models with exogenous regressor as well as function instrumental variable regression. We combine the method of sieves and the generalised method of moments with a continuum of moment conditions, which simultaneously achieves dimension reduction and utilisation of all the moment conditions. The estimator has a closed-form, making it easy to implement. We derive the minimax rate of convergence of each estimator under suitable assumptions and show that the rates are optimal.

There are still a few unanswered questions in this chapter to be studied in the future. First of all, is there a weighting scheme which outperforms other weighting schemes, and is it the inverse covariance operator of the moment conditions? The problem is particularly tricky as most interesting weighting matrix in Hansen

(1982)'s efficient GMM needs further regularisation in the framework of the functional linear regression. This adds further bias and variance into the estimator that requires a delicate balance to achieve desirable statistical properties. The second challenge is related to the choice of the tuning parameter k_n. Although many data-driven methods have been suggested to pick k_n, e.g., cross-validation criteria and information criteria, the theoretical and finite sample properties of these methods still need further exploration. Third, the assumptions required to derive the rate of convergence could be relaxed further. For example, extensions could be made to allow for spatial or temporal dependence, which would significantly improve the applicability of the functional linear regression models in economics.

Chapter 6

Conclusion

Understanding spatial heterogeneity has been a central concern in spatial and urban economic studies during the past two decades (Anselin, 1988, 2001; Brunsdon et al., 1998; Fotheringham et al., 2002; Wheeler, 2019), driven by the ever-increasing availability of geo-tagged big data, rapid progress in econometric methodologies as well as researchers' desire to understand the structural instability behind the spatial processes.

This thesis contributes to the methodology of spatial heterogeneity by developing three classes of varying coefficients regressions. In chapter 3, we propose the spatially varying coefficient regression (SVCR), designed for spatial cross-sectional data. The SVCR allows the spatial regression coefficients between a dependent variable and independent variables to vary smoothly over space, therefore capturing the spatial heterogeneous effects. Conceptually, the regression setup is similar to the geographic weighted regression (GWR), widely used in applied geographic research (Brunsdon et al., 1996, 1998; Fotheringham et al., 2002; Wheeler, 2019). We highlight the relationship and the core differences between the two methodologies and conduct detailed simulation studies to investigate their respective performances. We also address several spatial methodological issues such as smoothing over complex regions, modelling spatially auto-correlated and heteroskedastic errors, and spatial prediction with a large dataset. In chapter 4 we develop the spatiotemporally varying coefficient regression (STVCR) for spatial-temporal data. The STVCR jointly models both spatial and temporal heterogeneous effects, extending the SVCR to the temporal dimension. We compare the STVCR with the geographical and temporal weighted regression (GTWR) (Fotheringham et al., 2015) and highlight the advantage of the proposed methodology. Chapter 5 is dedicated to the methodology of the functional linear regression for spatial functional data. We propose a unified framework for estimating a functional linear regression with exogenous regressors or with instrumental variables. The method is called sieve continuum generalised method of moments (SCGMM), motivated by the combination of the sieve method (Grenander, 1981; Chen, 2007) and the GMM method with a continuum of moment conditions (Carrasco and Florens, 2000). We derive the rate of convergence of the estimator and contribute to the research of the statistical ill-posed problems (Kress et al., 1989; Carrasco et al., 2007).

Empirically, we address two important issues in urban economics, urban housing and urban consumption. In chapter 3, we investigate a housing market located in Aveiro, Portugal. We apply the SVCR that jointly models the spatial heterogeneity for coefficients and spatial autocorrelation for the error term in a hedonic regression framework. The method takes into consideration the irregularly shaped domain and estimates the complex spatial pattern of the shadow price. The estimates of the coefficients and prediction provide new insights about the urban housing market segmentation. Chapter 4 conducts an empirical study of the relationship between urban consumption amenities and human activity in China, following Glaeser et al. (2001). To facilitate the investigation, we collect two novel datasets, large scale mobile phone positioning data (MPP) and web-based point of interest (POI) data. Whereas the MPP data provide the spatiotemporal distribution of human activity on an hourly basis with high spatial resolution, the POI data inform the detailed distribution of individual consumption amenity. The incredible spatial and temporal granularity of the data allows us to apply the STVCR to uncover spatiotemporal heterogeneous effects between consumption amenities and human activity at very fine spatial and temporal scales. The study confirms that the role of consumption in facilitating urban density mostly holds in Beijing, though there exists substantial spatial and temporal uncertainty. It also illustrates the enormous potential of using internet-based data as an alternative data source in urban studies.

Admittedly, there are several limitations associated with the thesis. Methodologically, we focus mainly on the model and coefficient estimation. Many other methodological issues of importance are not covered in this thesis. For example, with SVCR and STVCR, how to conduct a homogeneity test to check whether a regression coefficient is spatially varying or non-varying? How to test whether the semiparametric model is more appropriate compared to parametric ones? With a functional linear regression model, on the other hand, we do not address several issues such as hypothesis testing, prediction. Besides, the proposed methodologies rely on the crucial assumption that the heterogeneous coefficients vary smoothly over the underlying spatial domain. If the spatial effects displays discontinuity or structural break (see, e.g., Zhu et al., 2014), or if the spatial domain is simply a small set of disjoint homogeneous areas, the proposed methodologies may not be suitable. The second limitation is associated with the complexity of implementation. Semi-nonparametric methods are considerably more complex in terms of the estimation procedures compared with their parametric counterparts. By introducing an infinite-dimensional parameter space, a semi-parametric regression relies on regularisation or smoothing to obtain stable estimates, which in turn requires parameter tuning. An ill-advised tuning parameter will either inflates the bias or the variance that makes the estimates unreliable. The third limitation is related to the empirical applications. This thesis does not place emphasis on causal inferences. Therefore the coefficients estimated from the SVCR and STVCR do not have a causal interpretation. In chapter 4, the model will suffer from endogeneity issues as consumption amenities open at places with more people around, bringing reverse causality. As a result, the coefficient does not inform the elasticity of human activity in response to opening a new restaurant.

Beyond this thesis, there are still plenty of exciting areas to be explored and studied in the future. The first area is related to the methodology and theory of semi-nonparametric econometrics in the spatial and urban setting. With SVCR and STVCR, methodologies should be extended to accommodate more sophisticated estimation and inference tasks, such as regression with a large number of dependent variables (high-dimensional model) and regression with endogeneity (nonparametric instrumental variable model). Simulation exercises could be carried out to investigate further the performance of the estimators under different data generating processes. With a functional linear regression model, substantial methodological work needs to be carried out to narrow the gap between the theory and empirical studies. For example, how to estimate an FLR model with multiple regressors, which could be a mixture of functional and non-functional components? The second area is about improving the implementation of the methodologies in this thesis. All the methods in this thesis require running bespoke Matlab scripts to carry out the estimation, which hinders the usage at different platforms. Efforts should be made to develop user-friendly statistical packages to facilitate the adoption and application of these methods. Finally, future researches should focus more on the estimation of the causal effects in an urban economic setting. With consumption and human activity, for example, one possible way is to adopt difference-in-differences (DID) approach (Abadie, 2005; Athey and Imbens, 2006), which compares the neighbourhoods with new consumption amenity opening with areas that do not experience consumption amenity change. This will bring a new understanding of how cities interact with the people and how policymakers could improve urban vibrancy.

Appendix A

Appendix of Chapter 3

A.1 Invert Large Symmetric Block Matrix

Here we present the details of the recursive method to obtain the inverse of a large symmetric block matrix of the form

$$\Sigma = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1k} \\ A_{21} & A_{22} & \cdots & A_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ A_{k1} & A_{k2} & \cdots & A_{kk} \end{pmatrix}$$

where A_{kk} has the form $A_{k,k} = a_k \boldsymbol{I}_{n_k} + b_k \mathbf{1}_{n_k} \mathbf{1}'_{n_k}$; n_k is the dimension of A_{kk}. Furthermore, $A_{ij} = A_{ji}^T$ for all $i, j = 1, \ldots, k$, where A_{ij} has a generic form $A_{ij} = \varsigma_{ij} \mathbf{1}_{n_i} \mathbf{1}'_{n_j}$. We will use the following two results, which can be found in Rao (1973).

(Rao, 1973, p.67) For a $n \times n$ matrix

$$A = a\boldsymbol{I}_n + b\mathbf{1}_n\mathbf{1}'_n,$$

its inverse is

$$A^{-1} = \tilde{a}\boldsymbol{I}_n + \tilde{b}\mathbf{1}_n\mathbf{1}'_n,$$

where

$$\tilde{a} = \frac{1}{a}, \quad \tilde{b} = \frac{-b}{a(a+nb)}$$

(Rao, 1973, p.33) For a symmetric block matrix

$$\Sigma = \begin{pmatrix} A & B \\ B' & D \end{pmatrix},$$

its inverse is

$$\Sigma^{-1} = \begin{pmatrix} A^{-1} + FE^{-1}F' & -FE^{-1} \\ -E^{-1}F' & E^{-1} \end{pmatrix}$$

where

$$E = D - B'A^{-1}B, F = A^{-1}B.$$

Now, we consider the recursive method for the inverse. The main result we establish is that the computation only involves finite scalar level operations.

In the first iteration, we immediately have the inverse of $A_{11} = a_1\boldsymbol{I}_{n_1} + b_k\mathbf{1}_{n_1}\mathbf{1}'_{n_1}$,

$$K_1 := A_{11}^{-1} = \tilde{a}_1\boldsymbol{I}_{n_1} + \tilde{b}_k\mathbf{1}_{n_1}\mathbf{1}'_{n_1}$$

In the second iteration, set

$$K_2 = \begin{pmatrix} K_1 & A_{12} \\ A'_{12} & A_{22} \end{pmatrix} = \begin{pmatrix} a_1\boldsymbol{I}_{n_1} + b_1\mathbf{1}_{n_1}\mathbf{1}'_{n_1} & \varsigma_{12}\mathbf{1}_{n_1}\mathbf{1}'_{n_2} \\ \varsigma_{12}\mathbf{1}_{n_2}\mathbf{1}'_{n_1} & a_2\boldsymbol{I}_{n_2} + b_2\mathbf{1}_{n_2}\mathbf{1}'_{n_2} \end{pmatrix}$$

Using the formula for symmetric block matrix mentioned above, it can be shown that the inverse has the same form as K_2

$$K_2^{-1} = \begin{pmatrix} p_1\boldsymbol{I}_{n_1} + q_{11}\mathbf{1}_{n_1}\mathbf{1}'_{n_1} & q_{12}\mathbf{1}_{n_1}\mathbf{1}'_{n_2} \\ q_{12}\mathbf{1}_{n_2}\mathbf{1}'_{n_1} & p_2\boldsymbol{I}_{n_2} + q_{22}\mathbf{1}_{n_2}\mathbf{1}'_{n_2} \end{pmatrix},$$

The inverse coefficients $(p_1, p_2, q_{11}, q_{12}, q_{22})$ are determined directly through the coefficients
$(a_1, a_2, b_1, b_2, \varsigma_{12}, n_1, n_2)$, which does not require matrix operation.

In the $(k+1)$th iteration, suppose K_k^{-1} has the following form,

$$K_k^{-1} = \begin{pmatrix} p_1\boldsymbol{I}_{n_1} + q_{11}\mathbf{1}_{n_1}\mathbf{1}'_{n_1} & \cdots & q_{1k}\mathbf{1}_{n_1}\mathbf{1}'_{n_k} \\ \vdots & \ddots & \vdots \\ q_{1k}\mathbf{1}_{n_k}\mathbf{1}'_{n_1} & \cdots & p_k\boldsymbol{I}_{n_k} + q_{kk}\mathbf{1}_{n_k}\mathbf{1}'_{n_k} \end{pmatrix}$$

The updated matrix in this iteration is as follows,

$$K_{k+1} = \begin{pmatrix} K_k & B_{k+1} \\ B'_{k+1} & A_{k+1,k+1} \end{pmatrix}$$

where

$$B_{k+1} = \begin{pmatrix} A_{1,k+1} \\ A_{2,k+1} \\ \vdots \\ A_{k,k+1} \end{pmatrix} = \begin{pmatrix} \varsigma_{1,k+1}\mathbf{1}_{n_1}\mathbf{1}'_{n_{k+1}} \\ \varsigma_{2,k+1}\mathbf{1}_{n_2}\mathbf{1}'_{n_{k+1}} \\ \vdots \\ \varsigma_{k,k+1}\mathbf{1}_{n_k}\mathbf{1}'_{n_{k+1}} \end{pmatrix}$$

and $A_{k+1,k+1} = a_{k+1}\boldsymbol{I}_{k+1} + b_{k+1}\mathbf{1}_{k+1}\mathbf{1}'_{k+1}$.

It can be shown that the inverse of K_{k+1} also possesses the same structure,

$$K_{k+1}^{-1} = \begin{pmatrix} r_1 \boldsymbol{I}_{n_1} + s_{1,1}\mathbf{1}_{n_1}\mathbf{1}'_{n_1} & \cdots & s_{1,k+1}\mathbf{1}_{n_1}\mathbf{1}'_{n_{k+1}} \\ \vdots & \ddots & \vdots \\ s_{1,k+1}\mathbf{1}_{n_{k+1}}\mathbf{1}'_{n_1} & \dots & r_{k+1}\boldsymbol{I}_{n_{k+1}} + s_{k+1,k+1}\mathbf{1}_{n_{k+1}}\mathbf{1}'_{n_{k+1}} \end{pmatrix}$$

where coefficients $(r_1, \dots, r_{k+1})$ and $(s_{1,1}, s_{1,2}, \dots, s_{n_{k+1},k+1})$ are computed from the coefficients of the original matrix $(p_1, \dots, p_k)$, $(q_{1,1}, q_{1,2}, \dots, q_{k,k})$, $(a_{1,k+1}, \dots, a_{k+1,k+1}, b_{k+1})$ and $(n_1, \dots, n_{k+1})$.

Appendix B

Appendix of Chapter 5

B.1 Lemma

Lemma B.1.1. *Let $\mathcal{G}$ be a Hilbert space with inner product $\langle f, g\rangle$ and norm $\|f\| = \langle f, f\rangle^{1/2}$ for $f, g \in \mathcal{G}$. Let $\{A_i, i = 1, \ldots, n\}$ be a sequence of i.i.d random functions, defined on $\mathcal{G}$.Let $\{B_i, i = 1, \ldots, n\}$ be another sequence of i.i.d random functions, defined on $\mathcal{G}$. Let $A = \mathbb{E}[A_i]$ and $B = \mathbb{E}[B_i]$, where $\mathbb{E}$ means expectation. Also define $\bar{A} = \frac{1}{n}\sum_{i=1}^{n} A_i$ and $\bar{B} = \frac{1}{n}\sum_{i=1}^{n} B_i$ then*

$$\langle \bar{A}, \bar{B}\rangle - \langle A, B\rangle = O_p(n^{-1/2}) \tag{B.1.1}$$

Proof. Note that

$$\begin{aligned}\langle \bar{A}, \bar{B}\rangle - \langle A, B\rangle &= \langle \bar{A} - A, \bar{B} - B\rangle + \langle \bar{A} - A, B\rangle + \langle A, B - \bar{B}\rangle \\ &\leq \|\bar{A} - A\|\|\bar{B} - B\| + \|\bar{A} - A\|\|B\| + \|A\|\|B - \bar{B}\|\end{aligned}$$

It is obvious that

$$\|\bar{A} - A\| = O_p(n^{-1/2}) \text{ and } \|B - \bar{B}\| = O_p(n^{-1/2}). \tag{B.1.2}$$

Besides

$$\|A\| \leq C \tag{B.1.3}$$

and

$$\|B\| \leq C \tag{B.1.4}$$

Hence $\langle \bar{A}, \bar{B}\rangle - \langle A, B\rangle = O_p(n^{-1/2})$

□

B.2 Proof 1: Derive Sieve Estimator for FLR-EXO

$$\tilde{b} = \tilde{b}_1\phi_1 + \cdots + \tilde{b}_{k_n}\phi_{k_n} := \boldsymbol{\phi}_n^{\mathrm{T}}\tilde{\mathbf{b}}_n, \tag{B.2.5}$$

where $\boldsymbol{\phi}_n = (\phi_1, \ldots, \phi_{k_n})$ and $\tilde{\mathbf{b}}_n = (\tilde{b}_1, \ldots, \tilde{b}_{k_n})$. Note that

$$\begin{aligned}
\hat{b}_n &= \arg\min_{\tilde{b}\in\mathcal{H}_n} \|\hat{K}\tilde{b} - \hat{g}\| \\
&= \arg\min_{\tilde{b}\in\mathcal{H}_n} \|\hat{K}\tilde{b} - \hat{g}\|^2 \\
&= \boldsymbol{\phi}_n^{\mathrm{T}} \arg\min_{\tilde{\mathbf{b}}_n\in\mathbb{R}^{k_n}} \|\hat{K}\tilde{b} - \hat{g}\|^2,
\end{aligned}$$

therefore, finding the function that minimise the objective function is equivalent to finding the generalised Fourier coefficients, given a sequence of basis functions.

$$\begin{aligned}
\hat{K}\tilde{b} &= \int_0^1 \frac{1}{n}\sum_{i=1}^{n} X_i(s)X_i(t)\tilde{b}(s)ds, \quad t \in [0,1] \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i(t)\int_0^1 X_i(s)\tilde{b}(s)ds \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i(t)\int_0^1 X_i(s)\left(\sum_{k=1}^{k_n}\tilde{b}_k\phi_k(s)\right)ds \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i(t)\left(\sum_{k=1}^{k_n}\tilde{b}_k\int_0^1 X_i(s)\phi_k(s)ds\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i(t)\left(\sum_{k=1}^{k_n}\tilde{b}_k\hat{c}_{i,k}\right)
\end{aligned}$$

where $\hat{c}_{i,k} = \int_0^1 X_i(s)\phi_k(s)ds$

$$\hat{g}(t) = \frac{1}{n}\sum_{i=1}^{n} X_i(t)Y_i, \quad t \in [0,1]$$

$$
\begin{aligned}
(\hat{K}\tilde{b}-\hat{g})(t) &= \left[\frac{1}{n}\sum_{i=1}^{n} X_i(t)\left(\sum_{k=1}^{k_n}\tilde{b}_k\hat{c}_{i,k}\right)\right]-\left[\frac{1}{n}\sum_{i=1}^{n} X_i(t)Y_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i(t)\left[\left(\sum_{k=1}^{k_n}\tilde{b}_k\hat{c}_{i,k}\right)-Y_i\right]
\end{aligned}
$$

$$
\|\hat{K}\tilde{b}-\hat{g}\|^2 = \int_0^1 \left\{\frac{1}{n}\sum_{i=1}^{n} X_i(t)\left[\left(\sum_{k=1}^{k_n}\tilde{b}_k\hat{c}_{i,k}\right)-Y_i\right]\right\}^2 dt
$$

$$
\begin{aligned}
\frac{\partial\|\hat{K}\tilde{b}-\hat{g}\|^2}{\partial\tilde{b}_j} &= \int_0^1 2\left\{\frac{1}{n}\sum_{i=1}^{n} X_i(t)\left[\left(\sum_{k=1}^{k_n}\tilde{b}_k\hat{c}_{i,k}\right)-Y_i\right]\right\}\left\{\frac{1}{n}\sum_{m=1}^{n} X_m(t)\hat{c}_{m,j}\right\} dt \\
&= \int_0^1 \frac{2}{n^2}\sum_{i=1}^{n}\sum_{m=1}^{n} X_m(t)X_i(t)\hat{c}_{m,j}\left[\left(\sum_{k=1}^{k_n}\tilde{b}_k\hat{c}_{i,k}\right)-Y_i\right] dt \\
&= \frac{2}{n^2}\sum_{i=1}^{n}\sum_{m=1}^{n}\left[\int_0^1 X_m(t)X_i(t)dt\right]\hat{c}_{m,j}\left[\left(\sum_{k=1}^{k_n}\tilde{b}_k\hat{c}_{i,k}\right)-Y_i\right] \\
&= \frac{2}{n^2}\sum_{i=1}^{n}\sum_{m=1}^{n} d_{m,i}\hat{c}_{m,j}\left[\left(\sum_{k=1}^{k_n}\tilde{b}_k\hat{c}_{i,k}\right)-Y_i\right]
\end{aligned}
$$

where $d_{m,i}=\int_0^1 X_m(t)X_i(t)dt$.

Let

$$
\frac{\partial\|\hat{K}\tilde{b}-\hat{g}\|^2}{\partial\tilde{b}_j}=0, \quad \text{for } j=1,\ldots,k_n
$$

$$
\sum_{i=1}^{n}\sum_{m=1}^{n} d_{m,i}\hat{c}_{m,j}\left(\sum_{k=1}^{r_n}\tilde{b}_k\hat{c}_{i,k}\right)=\sum_{i=1}^{n}\sum_{m=1}^{n} d_{m,i}\hat{c}_{m,j}Y_i, \quad j=1,\ldots,k_n
$$

which is equivalent to

$$
(\mathcal{X}_n^{\mathrm{T}}\mathcal{W}_n\mathcal{X}_n)\tilde{\boldsymbol{b}}_n=\mathcal{X}_n^{\mathrm{T}}\mathcal{W}_n\mathcal{Y}_n. \tag{B.2.6}
$$

Then

$$
\tilde{\boldsymbol{b}}_n=(\mathcal{X}_n^{\mathrm{T}}\mathcal{W}_n\mathcal{X}_n)^{-1}\mathcal{X}_n^{\mathrm{T}}\mathcal{W}_n\mathcal{Y}_n
$$

and

$$\hat{b}_n = \boldsymbol{\phi}_n^T (\mathcal{X}_n^{\mathrm{T}} \mathcal{W}_n \mathcal{X}_n)^{-1} \mathcal{X}_n^{\mathrm{T}} \mathcal{W}_n \mathcal{Y}_n$$

We could derive the Hessian matrix in the similar fashion,

$$H = \mathcal{X}_n^{\mathrm{T}} \mathcal{W}_n \mathcal{X}_n \tag{B.2.7}$$

which is positive definite.

B.3 Proof 2: Equivalence between $\hat{b}$ and $\tilde{b}$

Let $\hat{b}_{m,n} = \langle \hat{b}_n, \phi_m \rangle$, for $m = 1, \ldots, k_n$. $(\hat{b}_{1,n}, \ldots, \hat{b}_{k_n,n})^{\mathrm{T}} = (\mathcal{X}_n^{\mathrm{T}} \mathcal{W}_n \mathcal{X}_n)^{-1} \mathcal{X}_n^{\mathrm{T}} \mathcal{W}_n \mathcal{Y}_n$.

$$
\begin{aligned}
\hat{K}_n^* \hat{K}_n \hat{b}_n &= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k_n} \chi_{i,j} \langle X_i, \frac{1}{n} \sum_{l=1}^{n} \sum_{m=1}^{k_n} \hat{b}_{m,n} X_l \chi_{l,m} \rangle \phi_j \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{k_n} \sum_{l=1}^{n} \sum_{m=1}^{k_n} \chi_{i,j} \langle X_i, X_l \rangle \chi_{l,m} \hat{b}_{m,n} \phi_j \\
&= \frac{1}{n^2} \boldsymbol{\phi}_n^{\mathrm{T}} (\mathcal{X}_n^{\mathrm{T}} \mathcal{W}_n \mathcal{X}_n)(\mathcal{X}_n^{\mathrm{T}} \mathcal{W}_n \mathcal{X}_n)^{-1} \mathcal{X}_n^{\mathrm{T}} \mathcal{W}_n \mathcal{Y}_n \\
&= \frac{1}{n^2} \boldsymbol{\phi}_n^{\mathrm{T}} \mathcal{X}_n^{\mathrm{T}} \mathcal{W}_n \mathcal{Y}_n
\end{aligned}
$$

$$
\begin{aligned}
\hat{K}_n^* \hat{g} &= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k_n} \chi_{i,j} \langle X_i, \frac{1}{n} \sum_{l=1}^{n} X_l Y_l \rangle \phi_j \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{k_n} \sum_{l=1}^{n} \chi_{i,j} \langle X_i, X_l \rangle Y_l \phi_j \\
&= \frac{1}{n^2} \boldsymbol{\phi}_n^{\mathrm{T}} \mathcal{X}_n^{\mathrm{T}} \mathcal{W}_n \mathcal{Y}_n
\end{aligned}
$$

Therefore,

$$
\hat{K}_n^* \hat{K}_n \hat{b}_n = \hat{K}_n^* \hat{g} \tag{B.3.8}
$$

As $\hat{K}_n^* \hat{K}_n$ is invertible, $\hat{b}_n = \tilde{b}_n$

B.4 Proof 3: Rate of Convergence

The following proof derives the rate of convergence for functional linear regression with exogeneity. The results for functional IV regression will be similar, which will not be elaborated here.

By triangle inequality, we have

$$\|\hat{b}_n - b\| \leq \|\hat{b}_n - b_n\| + \|b_n - b\| \tag{B.4.9}$$

By Assumption (5.5.8),

$$\|b_n - b\| = O(k_n^{-s}). \tag{B.4.10}$$

For $\|\hat{b}_n - b_n\|$, notice that both $\hat{b}_n$ and b_n belong to Θ_n, by assumption (5.5.5) and (5.5.8), therefore $\hat{b}_n - b_n \in \Theta_{ns}$, and by the definition of τ_n,

$$\frac{\|\hat{b}_n - b_n\|}{\|K^*K(\hat{b}_n - b_n)\|} \leq \sup_{f \in \mathcal{H}_{ns}} \frac{\|f\|}{\|(K^*K)f\|} := \tau_n, \tag{B.4.11}$$

therefore

$$\|\hat{b}_n - b_n\| \leq \tau_n \|K^*K(\hat{b}_n - b_n)\|. \tag{B.4.12}$$

Note that $K^*K(\hat{b}_n - b_n)$ could be decomposed into

$$K^*K(\hat{b}_n - b_n) = (K^*K - \hat{K}_n^*\hat{K}_n)\hat{b}_n + (\hat{K}_n^*\hat{K}_n\hat{b}_n - K^*Kb) + K^*K(b - b_n) \tag{B.4.13}$$

Besides, $K^*Kb = K^*g$ and $\hat{K}_n^*\hat{K}_n\hat{b}_n = \hat{K}^*\hat{g}$ w.p.a 1.

By triangle inequality

$$\|K^*K(\hat{b}_n - b_n)\| = \|(K^*K - \hat{K}_n^*\hat{K}_n)\hat{b}_n\| + \|\hat{K}^*\hat{g} - K^*g\| + \|K^*K(b - b_n)\| \tag{B.4.14}$$

Let

$$\begin{aligned} \Delta_1 &:= \|(K^*K - \hat{K}_n^*\hat{K}_n)\hat{b}_n\|, \\ \Delta_2 &:= \|\hat{K}^*\hat{g} - K^*g\|, \\ \Delta_3 &:= \|K^*K(b - b_n)\|. \end{aligned}$$

Note that $K_n^* K_n b = K_n^* K_n b_n$ and $K_n^* K_n (b - b_n) = 0$, hence for Δ_3,

$$\begin{aligned}
\|K^* K(b - b_n)\| &= \|K^* K(b - b_n) - K_n^* K_n (b - b_n)\| \\
&= \|(K^* K - K_n^* K_n)(b - b_n)\| \\
&\leq \|(K^* K - K_n^* K_n)\| \|b - b_n\| \\
&= \tau_n^{-1} O(k_n^{-s})
\end{aligned}$$

by Assumption (5.5.8) and (5.5.7)

For Δ_2

$$\|\hat{K}^* \hat{g} - K^* g\| \leq \|\hat{K}^* \hat{g} - K_n^* g\| + \|K_n^* g - K^* g\| \tag{B.4.15}$$

Let $\Delta_{21} := \|\hat{K}^* \hat{g} - K_n^* g\|$, and $\Delta_{22} := \|K_n^* g - K^* g\|$.

$$\Delta_{22} = O(k_n^{-r-s}) \tag{B.4.16}$$

by Assumption (5.5.8).

For Δ_{21},

$$\hat{K}^* \hat{g} = \sum_{j=1}^{k_n} \hat{\zeta}_j \phi_j \tag{B.4.17}$$

where

$$\hat{\zeta}_j = \int_0^1 \int_0^1 \left(\frac{1}{n} \sum_{i=1}^{n} X_i(s) X_i(t) \right) \left(\frac{1}{n} \sum_{m=1}^{n} X_m(t) Y_m \right) dt \phi_j(s) ds \tag{B.4.18}$$

and

$$K_n^* g = \sum_{j=1}^{k_n} \zeta_j \phi_j \tag{B.4.19}$$

where

$$\zeta_j = \int_0^1 \int_0^1 \mathbb{E}[X(s)X(t)] \mathbb{E}[X(t)Y] dt \phi_j(s) ds \tag{B.4.20}$$

Moment calculation suggests that $\hat{\zeta}_j - \zeta_j = O_p(n^{-1/2})$. Hence

$$\begin{aligned}
\Delta_{21} &= \left[\sum_{j=1}^{k_n} (\hat{\zeta}_j - \zeta_j)^2 \right]^{1/2} \\
&= (k_n O_p(n^{-1/2})^2)^{1/2} \\
&= O_p\left(\sqrt{k_n/n}\right).
\end{aligned}$$

So

$$\Delta_2 \leq \Delta_{21} + \Delta_{22} = O_p\left(k_n^{-r-s} + \sqrt{k_n/n}\right). \tag{B.4.21}$$

For Δ_1,

$$\|(K^*K - \hat{K}_n^*\hat{K}_n)\hat{b}_n\| \leq \|(K^*K - K_n^*K_n)\hat{b}_n\| + \|(K_n^*K_n - \hat{K}_n^*\hat{K}_n)\hat{b}_n\|$$

$$\|(K^*K - K_n^*K_n)\hat{b}_n\| \leq \sup_{v\in\Theta_n}\|(K^*K - K_n^*K_n)f\| = \tau_n^{-1}O(k_n^{-s}) \tag{B.4.22}$$

by assumption (5.5.8).

$$\|(K_n^*K_n - \hat{K}_n^*\hat{K}_n)\hat{b}_n\| \leq \sup_{f\in\Theta_n}\|(K_n^*K_n - \hat{K}_n^*\hat{K}_n)f\| \tag{B.4.23}$$

where $f = \sum_{j=1}^{k_n} f_j\phi_j, v_j = \int_0^1 f(s)\phi_j(s)ds$

$$\|(K_n^*K_n - \hat{K}_n^*\hat{K}_n)\|^2 = \sum_{p=1}^{k_n}\left[\sum_{q=1}^{k_n}(\hat{c}_{q,p} - c_{q,p})v_p\right]^2$$

where

$$\hat{c}_{q,p} = \int_0^1\int_0^1\int_0^1 \frac{1}{n}\sum_{i=1}^{n}[X_i(s)X_i(t)]\frac{1}{n}\sum_{i=1}^{n}[X(u)X(t)]dtdsdu$$

$$c_{q,p} = \int_0^1\int_0^1\int_0^1 \mathbb{E}[X(s)X(t)]\mathbb{E}[X(u)X(t)]dtdsdu$$

Using the same argument,

$$\hat{c}_{p,q} - c_{p,q} = O_p(1/\sqrt{n}) \tag{B.4.24}$$

for all p, q

$$\left|\sum_{p=1}^{k_n}(\hat{c}_{q,p} - c_{q,p})v_p\right| \leq \sup_{p,q}|\hat{c}_{q,p} - c_{q,p}|\sum_{p=1}^{k_n}|v_p| = O_p(1/\sqrt{n}) \tag{B.4.25}$$

$$\sum_{q=1}^{k_n}\left[\sum_{p=1}^{k_n}(\hat{c}_{q,p} - c_{q,p})v_p\right]^2 \leq \sum_{q=1}^{k_n} O_p(1/n) = O_p(k_n/n) \tag{B.4.26}$$

Therefore

$$\|(K_n^*K_n - \hat{K}_n^*\hat{K}_n)\| = O_p(\sqrt{k_n/n})$$

Combine all the bounds, we get

$$
\begin{aligned}
\|\hat{b}_n - b\| &\leq \tau_n(\Delta_1 + \Delta_2 + \Delta_3) + O(k_n^s) \\
&= O_p(k_n^{-s} + \tau_n\sqrt{k_n/n} + \tau_n(k_n^{-r-s} + \sqrt{k_n/n}) \\
&= O_p(k_n^{-s} + k_n^r\sqrt{k_n/n})
\end{aligned}
$$

If we pick $k_n \asymp n^{1/(2s+2r+1)}$, the rate of convergence will be

$$
\|\hat{b}_n - b\| = O_p\left(n^{-s/(2s+2r+1)}\right) \tag{B.4.27}
$$

Bibliography

Abadie, Alberto (2005), "Semiparametric difference-in-differences estimators." *The Review of Economic Studies*, 72, 1–19.

Ahrens, Achim (2015), "Civil conflicts, economic shocks and night-time lights." *Peace Economics, Peace Science and Public Policy*, 21, 433–444.

Ahrens, Achim and Arnab Bhattacharjee (2015), "Two-step lasso estimation of the spatial weights matrix." *Econometrics*, 3, 128–155.

Anselin, Luc (1988), *Spatial econometrics: Methods and models*. Springer Netherlands.

Anselin, Luc (2001), "Spatial econometrics." *A Companion to Theoretical Econometrics*, 310330.

Anselin, Luc and Nancy Lozano-Gracia (2008), "Errors in variables and spatial effects in hedonic house price models of ambient air quality." *Empirical Economics*, 34, 5–34.

Anselin, Luc, Nancy Lozano-Gracia, Uwe Deichmann, and Somik Lall (2010), "Valuing access to water—a spatial hedonic approach, with an application to bangalore, india." *Spatial Economic Analysis*, 5, 161–179.

Arlot, Sylvain, Alain Celisse, et al. (2010), "A survey of cross-validation procedures for model selection." *Statistics Surveys*, 4, 40–79.

Athey, Susan and Guido W Imbens (2006), "Identification and inference in nonlinear difference-in-differences models." *Econometrica*, 74, 431–497.

Baltagi, Badi (2008), *Econometric analysis of panel data*. John Wiley & Sons.

Barca, Fabrizio, Philip McCann, and Andrés Rodríguez-Pose (2012), "The case for regional development intervention: place-based versus place-neutral approaches." *Journal of Regional Science*, 52, 134–152.

Batty, Michael (2013), "Big data, smart cities and city planning." *Dialogues in Human Geography*, 3, 274–279.

Beijing Municipal Bureau of Statistics (2018), "Population of beijing." Data retrieved from Beijing Municipal Bureau of Statistics. `http://www.bjstats.gov.cn/tjsj/cysj/201511/t20151109_311727.html`.

Belloni, Alexandre, Victor Chernozhukov, Denis Chetverikov, and Kengo Kato (2015), "Some new asymptotic theory for least squares series: Pointwise and uniform results." *Journal of Econometrics*, 186, 345–366.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen (2011), "Inference for high-dimensional sparse econometric models." *arXiv preprint arXiv:1201.0220*.

Bhaduri, Budhendra, Edward Bright, Phillip Coleman, and Jerome Dobson (2002), "Landscan." *Geoinformatics*, 5, 34–37.

Bhattacharjee, Arnab, Eduardo Castro, Taps Maiti, and João Marques (2016a), "Endogenous spatial regression and delineation of submarkets: a new framework with application to housing markets." *Journal of Applied Econometrics*, 31, 32–57.

Bhattacharjee, Arnab, Eduardo Castro, and João Marques (2012), "Spatial interactions in hedonic pricing models: The urban housing market of aveiro, portugal." *Spatial Economic Analysis*, 7, 133–167.

Bhattacharjee, Arnab, Taps Maiti, Mark Wang, and Ping Shou Zhong (2016b), "Spatially varying regression over irregularly shaped regions: Application to a hedonic house price model." *Manuscript*.

Bivand, R and D Spgwr Yu (2015), "Geographically weighted regression. r package version 0.6–28."

Blundell, Richard, Martin Browning, and Ian Crawford (2008), "Best nonparametric bounds on demand responses." *Econometrica*, 76, 1227–1262.

Blundell, Richard, Xiaohong Chen, and Dennis Kristensen (2007), "Semi-nonparametric iv estimation of shape-invariant engel curves." *Econometrica*, 75, 1613–1669.

Blundell, Richard W, Martin Browning, and Ian A Crawford (2003), “Nonparametric engel curves and revealed preference.” *Econometrica*, 71, 205–240.

Brunsdon, Chris, Jonathan Corcoran, and Gary Higgs (2007), “Visualising space and time in crime patterns: A comparison of methods.” *Computers, Environment and Urban Systems*, 31, 52–75.

Brunsdon, Chris, A Stewart Fotheringham, and Martin E Charlton (1996), “Geographically weighted regression: a method for exploring spatial nonstationarity.” *Geographical Analysis*, 28, 281–298.

Brunsdon, Chris, Stewart Fotheringham, and Martin Charlton (1998), “Geographically weighted regression.” *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47, 431–443.

Cai, T Tony, Peter Hall, et al. (2006), “Prediction in functional linear regression.” *The Annals of Statistics*, 34, 2159–2179.

Cardot, Hervé, Frédéric Ferraty, and Pascal Sarda (1999), “Functional linear model.” *Statistics & Probability Letters*, 45, 11–22.

Carrasco, Marine and Jean-Pierre Florens (2000), “Generalization of gmm to a continuum of moment conditions.” *Econometric Theory*, 16, 797–834.

Carrasco, Marine, Jean-Pierre Florens, and Eric Renault (2007), “Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization.” *Handbook of Econometrics*, 6, 5633–5751.

Chai, Andreas and Alessio Moneta (2010), “Retrospectives: Engel curves.” *Journal of Economic Perspectives*, 24, 225–40.

Chen, Jie, Shih-Lung Shaw, Hongbo Yu, Feng Lu, Yanwei Chai, and Qinglei Jia (2011), “Exploratory data analysis of activity diary data: a space–time gis approach.” *Journal of Transport Geography*, 19, 394–404.

Chen, Xiaohong (2007), “Large sample sieve estimation of semi-nonparametric models.” *Handbook of Econometrics*, 6, 5549–5632.

Chen, Xiaohong and Demian Pouzo (2012), “Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals.” *Econometrica*, 80, 277–321.

Chen, Xiaohong and Markus Reiss (2011), "On rate optimality for ill-posed inverse problems in econometrics." *Econometric Theory*, 27, 497–521.

Cleveland, William S (1979), "Robust locally weighted regression and smoothing scatterplots." *Journal of the American Statistical Association*, 74, 829–836.

Couture, Victor (2013), "Valuing the consumption benefits of urban density." *University of California, Berkeley. Processed*.

Craven, Peter and Grace Wahba (1978), "Smoothing noisy data with spline functions." *Numerische mathematik*, 31, 377–403.

Cressie, Noel (1990), "The origins of kriging." *Mathematical geology*, 22, 239–252.

Cressie, Noel (1993), *Statistics for spatial data*. New York: Wiley.

Cressie, Noel and Gardar Johannesson (2008), "Fixed rank kriging for very large spatial data sets." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 209–226.

Dalenberg, Douglas R and Mark D Partridge (1995), "The effects of taxes, expenditures, and public infrastructure on metropolitan area employment." *Journal of Regional Science*, 35, 617–640.

Darolles, Serge, Yanqin Fan, Jean-Pierre Florens, and Eric Renault (2011), "Nonparametric instrumental regression." *Econometrica*, 79, 1541–1565.

Dauxois, Jacques, Alain Pousse, and Yves Romain (1982), "Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference." *Journal of Multivariate Analysis*, 12, 136–154.

De Boor, Carl, Carl De Boor, Etats-Unis Mathématicien, Carl De Boor, and Carl De Boor (1978), *A practical guide to splines*, volume 27. Springer-Verlag New York.

Dobson, Jerome E, Edward A Bright, Phillip R Coleman, Richard C Durfee, and Brian A Worley (2000), "Landscan: a global population database for estimating populations at risk." *Photogrammetric Engineering and Remote Sensing*, 66, 849–857.

Engl, Heinz Werner, Martin Hanke, and Andreas Neubauer (1996), *Regularization of inverse problems*, volume 375. Springer Science & Business Media.

Engle, Robert F (2000), “The econometrics of ultra-high-frequency data.” *Econometrica*, 68, 1–22.

Fan, Jianqing (1992), “Design-adaptive nonparametric regression.” *Journal of the American Statistical Association*, 87, 998–1004.

Fan, Jianqing (1993), “Local linear regression smoothers and their minimax efficiencies.” *The Annals of Statistics*, 196–216.

Fan, Jianqing and Irene Gijbels (1992), “Variable bandwidth and local linear regression smoothers.” *The Annals of Statistics*, 2008–2036.

Fan, Jianqing and Irene Gijbels (1995), “Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation.” *Journal of the Royal Statistical Society. Series B (Methodological)*, 371–394.

Fan, Jianqing and Irene Gijbels (1996), *Local polynomial modelling and its applications: Monographs on statistics and applied probability*, volume 66. CRC Press.

Fan, Jianqing and Huang Tao (2005), “Profile likelihood inferences on semiparametric varying-coefficient partially linear models.” *Bernoulli*, 11, 1031–1057.

Fan, Jianqing and Wenyang Zhang (2008), “Statistical methods with varying coefficient models.” *Statistics and its Interface*, 1, 179.

Feng, Wenning, Chae Young Lim, Tapabrata Maiti, and Z Zhang (2012), “Simultaneous estimation of disease risks and spatial clustering: A hierarchical bayes approach.” Technical report, Mimeo.

Florens, Jean-Pierre and Sébastien Van Bellegem (2015), “Instrumental variable estimation in functional linear models.” *Journal of Econometrics*, 186, 465–476.

Floyd, Robert W (1962), “Algorithm 97: shortest path.” *Communications of the ACM*, 5, 345.

Follain, James R and Stephen Malpezzi (1980), “Estimates of housing inflation for 39 smsas: An alternative to the consumer price index.” *The Annals of Regional Science*, 14, 41–56.

Fotheringham, A Stewart, Ricardo Crespo, and Jing Yao (2015), “Geographical and temporal weighted regression (gtwr).” *Geographical Analysis*, 47, 431–452.

Fotheringham, A.S., C. Brunsdon, and M. Charlton (2002), *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Wiley.

Gagliardini, Patrick and Olivier Scaillet (2012), "Tikhonov regularization for nonparametric instrumental variable estimators." *Journal of Econometrics*, 167, 61–75.

Galster, George (2001), "On the nature of neighbourhood." *Urban Studies*, 38, 2111–2124.

Gillen, Kevin, Thomas Thibodeau, and Susan Wachter (2001), "Anisotropic autocorrelation in house prices." *The Journal of Real Estate Finance and Economics*, 23, 5–30.

Glaeser, Edward L (2010), "Introduction to" agglomeration economics"." In *Agglomeration economics*, 1–14, University of Chicago Press.

Glaeser, Edward L and Joshua D Gottlieb (2006), "Urban resurgence and the consumer city." *Urban Studies*, 43, 1275–1299.

Glaeser, Edward L and Matthew E Kahn (2001), "Decentralized employment and the transformation of the american city." Technical report, National Bureau of Economic Research.

Glaeser, Edward L, Hyunjin Kim, and Michael Luca (2017), "Nowcasting the local economy: Using yelp data to measure economic activity." Technical report, National Bureau of Economic Research.

Glaeser, Edward L, Jed Kolko, and Albert Saiz (2001), "Consumer city." *Journal of Economic Geography*, 1, 27–50.

Goulard, Michel, Thibault Laurent, and Christine Thomas-Agnan (2017), "About predictions in spatial autoregressive models: Optimal and almost optimal strategies." *Spatial Economic Analysis*, 12, 304–325.

Grenander, Ulf (1981), *Abstract inference*. Wiley New York.

Hall, Peter and Joel L Horowitz (2005), "Nonparametric methods for inference in the presence of instrumental variables." *The Annals of Statistics*, 33, 2904–2929.

Hall, Peter and Joel L Horowitz (2007), "Methodology and convergence rates for functional linear regression." *The Annals of Statistics*, 35, 70–91.

Hall, Peter, Hans-Georg Müller, and Jane-Ling Wang (2006), "Properties of principal component methods for functional and longitudinal data analysis." *The Annals of Statistics*, 1493–1517.

Hall, Peter and Céline Vial (2006), "Assessing the finite dimensionality of functional data." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68, 689–705.

Hansen, Lars Peter (1982), "Large sample properties of generalized method of moments estimators." *Econometrica: Journal of the Econometric Society*, 1029–1054.

Hastie, Trevor and Robert Tibshirani (1993), "Varying-coefficient models." *Journal of the Royal Statistical Society. Series B (Methodological)*, 757–796.

Hastie, Trevor, Robert Tibshirani, and Martin Wainwright (2015), *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

Henderson, Daniel J, Raymond J Carroll, and Qi Li (2008), "Nonparametric estimation and testing of fixed effects panel data models." *Journal of Econometrics*, 144, 257–275.

Henderson, J Vernon, Adam Storeygard, and David N Weil (2012), "Measuring economic growth from outer space." *American Economic Review*, 102, 994–1028.

Hjelle, Øyvind and Morten Dæhlen (2006), *Triangulations and applications*. Springer Science & Business Media.

Holl, Adelheid (2004), "Transport infrastructure, agglomeration economies, and firm birth: empirical evidence from portugal." *Journal of Regional Science*, 44, 693–712.

Horowitz, Joel L (2011), "Applied nonparametric instrumental variables estimation." *Econometrica*, 79, 347–394.

Horowitz, Joel L (2012a), *Semiparametric methods in econometrics*, volume 131. Springer Science & Business Media.

Horowitz, Joel L (2012b), "Specification testing in nonparametric instrumental variable estimation." *Journal of Econometrics*, 167, 383–396.

Ichimura, Hidehiko (1993), "Semiparametric least squares (sls) and weighted sls estimation of single-index models." *Journal of Econometrics*, 58, 71–120.

Jacobs, Jane (1962), *The death and life of great American cities*. London: Jonathan Cape.

Jacobs-Crisioni, Chris, Piet Rietveld, Eric Koomen, and Emmanouil Tranos (2014), "Evaluating the impact of land-use density and mix on spatiotemporal urban activity patterns: an exploratory study using mobile phone data." *Environment and Planning A*, 46, 2769–2785.

Jones, M Chris (1993), "Simple boundary correction for kernel density estimation." *Statistics and Computing*, 3, 135–146.

Kress, Rainer, V Maz'ya, and V Kozlov (1989), *Linear integral equations*, volume 82. Springer.

Lancaster, Kelvin J (1966), "A new approach to consumer theory." *Journal of Political Economy*, 74, 132–157.

LeSage, James and Robert Kelley Pace (2009), *Introduction to spatial econometrics*. Chapman and Hall/CRC.

Li, Linna, Michael F Goodchild, and Bo Xu (2013), "Spatial, temporal, and socioeconomic patterns in the use of twitter and flickr." *Cartography and Geographic Information Science*, 40, 61–77.

Li, Qi and Jeffrey Scott Racine (2007), *Nonparametric econometrics: theory and practice*. Princeton University Press.

Linton, Oliver and Zhijie Xiao (2019), "Efficient estimation of nonparametric regression in the presence of dynamic heteroskedasticity." *Journal of Econometrics*.

Liu, Yu, Zhengwei Sui, Chaogui Kang, and Yong Gao (2014), "Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data." *PloS One*, 9, e86026.

Lloyd, Christopher D (2010), *Local models for spatial analysis*. CRC press.

Long, Ying, Yu Zhang, and Chengyin Cui (2012), "Identifying commuting pattern of beijing using bus smart card data." *Acta Geographica Sinica*, 67, 1339–1352.

Ma, Xiaolei, Jiyu Zhang, Chuan Ding, and Yunpeng Wang (2018), "A geographically and temporally weighted regression model to explore the spatiotemporal influence of built environment on transit ridership." *Computers, Environment and Urban Systems*, 70, 113–124.

Malpezzi, Stephen (2002), "Hedonic pricing models: a selective and applied review." *Housing Economics and Public Policy*, 67–89.

Marques, J.L. (2012), "Notions of space in urban housing markets." Technical report, Unpublished PhD Thesis, University of Aveiro, Portugal.

Martins-Filho, Carlos and Feng Yao (2009), "Nonparametric regression estimation with general parametric error covariance." *Journal of Multivariate Analysis*, 100, 309–333.

Masry, Elias (1996), "Multivariate local polynomial regression for time series: uniform strong consistency and rates." *Journal of Time Series Analysis*, 17, 571–599.

Mayer-Schönberger, Viktor and Kenneth Cukier (2013), "Big data–a revolution that will transform how we live, think and work."

Michalopoulos, Stelios and Elias Papaioannou (2013), "Pre-colonial ethnic institutions and contemporary african development." *Econometrica*, 81, 113–152.

Mincer, Jacob (1958), "Investment in human capital and personal income distribution." *Journal of Political Economy*, 66, 281–302.

Müller, Hans-Georg (1991), "Smooth optimum kernel estimators near endpoints." *Biometrika*, 78, 521–530.

Nadaraya, EA (1965), "On non-parametric estimates of density functions and regression curves." *Theory of Probability & Its Applications*, 10, 186–190.

Newey, Whitney K (1997), "Convergence rates and asymptotic normality for series estimators." *Journal of Econometrics*, 79, 147–168.

Newey, Whitney K and James L Powell (2003), "Instrumental variable estimation of nonparametric models." *Econometrica*, 1565–1578.

Opsomer, Jean D, David Ruppert, Matt P Wand, Ulla Holst, and Ola Hössjer (1999), "Kriging with nonparametric variance function estimation." *Biometrics*, 55, 704–710.

Pagan, Adrian and Aman Ullah (1999), *Nonparametric econometrics*. Cambridge University Press.

Ramsay, James O and Bernard W Silverman (2005), *Functional Data Analysis*. Springer.

Ramsay, James O and Bernard W Silverman (2007), *Applied functional data analysis: methods and case studies*. Springer.

Ramsay, Tim (2002), "Spline smoothing over difficult regions." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 307–319.

Rappaport, Jordan (2008), "Consumption amenities and city population density." *Regional Science and Urban Economics*, 38, 533–552.

Robinson, Peter M (1988), "Root-n-consistent semiparametric regression." *Econometrica: Journal of the Econometric Society*, 931–954.

Rosen, Sherwin (1974), "Hedonic prices and implicit markets: product differentiation in pure competition." *Journal of Political Economy*, 82, 34–55.

Rothenberg, Jerome, George C Galster, Richard V Butler, and John R Pitkin (1991), *The maze of urban housing markets: Theory, evidence, and policy*. University of Chicago Press.

Ruppert, David, Matt P Wand, and Raymond J Carroll (2003), *Semiparametric regression*. 12, Cambridge University Press.

Ruppert, David and Matthew P Wand (1994), "Multivariate locally weighted least squares regression." *The Annals of Statistics*, 1346–1370.

Sangalli, Laura M, James O Ramsay, and Timothy O Ramsay (2013), "Spatial spline regression models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75, 681–703.

Schiff, Nathan (2014), "Cities and product variety: evidence from restaurants." *Journal of Economic Geography*, 15, 1085–1123.

Schintler, Laurie A and Manfred M Fischer (2018), "Big data and regional science: Opportunities, challenges, and directions for future research." *Working Paper*.

Schumaker, Larry (2007), *Spline functions: Basic theory*. Cambridge University Press.

Silverman, Bernard W and James O Ramsay (2005), *Functional Data Analysis*. Springer.

Spence, Michael (1978), "Job market signaling." In *Uncertainty in Economics*, 281–306, Elsevier.

Stone, Charles J (1977), "Consistent nonparametric regression." *The Annals of Statistics*, 595–620.

Stone, Charles J (1980), "Optimal rates of convergence for nonparametric estimators." *The Annals of Statistics*, 1348–1360.

Stone, Charles J (1982), "Optimal global rates of convergence for nonparametric regression." *The Annals of Statistics*, 1040–1053.

Su, Liangjun, Aman Ullah, and Yun Wang (2013), "Nonparametric regression estimation with general parametric error covariance: a more efficient two-step estimator." *Empirical Economics*, 45, 1009–1024.

The World Bank (2018), "Urban population (% of total)." URL `https://data.worldbank.org/indicator/sp.urb.totl.in.zs`.

Tikhonov, AN (1963), "On the solution of incorrectly put problems and the regularisation method." *Outlines Joint Sympos. Partial Differential Equations (Novosibirsk, 1963)*, 261–265.

Tranos, Emmanouil and Peter Nijkamp (2015), "Mobile phone usage in complex urban systems: a space–time, aggregated human activity study." *Journal of Geographical Systems*, 17, 157–185.

Varian, Hal R (2014), "Big data: New tricks for econometrics." *Journal of Economic Perspectives*, 28, 3–28.

Wang, Haonan and M Giovanna Ranalli (2007), "Low-rank smoothing splines on complicated domains." *Biometrics*, 63, 209–217.

Wang, Jane-Ling, Jeng-Min Chiou, and Hans-Georg Müller (2016), "Functional data analysis." *Annual Review of Statistics and Its Application*, 3, 257–295.

Wang, Rui (2009), "The structure of chinese urban land prices: Estimates from benchmark land price data." *The Journal of Real Estate Finance and Economics*, 39, 24–38.

Wasserman, Larry (2006), *All of nonparametric statistics*. Springer Science & Business Media.

Watson, Geoffrey S (1964), "Smooth regression analysis." *Sankhyā: The Indian Journal of Statistics, Series A*, 359–372.

Wheeler, David C (2019), "Geographically weighted regression." *Handbook of Regional Science*, 1–27.

Wilkinson, RK and Catherine A Archer (1973), "Measuring the determinants of relative house prices." *Environment and Planning A*, 5, 357–367.

Wood, Simon N, Mark V Bravington, and Sharon L Hedley (2008), "Soap film smoothing." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 931–955.

Wooldridge, Jeffrey M (2010), *Econometric analysis of cross section and panel data*. MIT press.

Wu, Wenjie (2016), *Economics of planning policies in China: Infrastructure, location and cities*. Routledge.

Wu, Wenjie, Guanpeng Dong, and Wenzhong Zhang (2017), "The puzzling heterogeneity of amenity capitalization effects on land markets." *Papers in Regional Science*, 96, S135–S153.

Wu, Wenjie, Jianghao Wang, and Tianshi Dai (2016a), "The geography of cultural ties and human mobility: Big data in urban contexts." *Annals of the American Association of Geographers*, 106, 612–630.

Wu, Wenjie, Jianghao Wang, Chengyu Li, and Mark Wang (2016b), "The geography of city liveliness and consumption: evidence from location-based big data." Technical report, London School of Economics and Political Science, LSE Library.

Yang, Yun and Surya T Tokdar (2015), "Minimax-optimal nonparametric regression in high dimensions." *The Annals of Statistics*, 43, 652–674.

Zheng, Siqi, Xiaoke Hu, Jianghao Wang, and Rui Wang (2016), "Subways near the subway: Rail transit and neighborhood catering businesses in beijing." *Transport Policy*, 51, 81–92.

Zheng, Siqi and Matthew E Kahn (2008), "Land and residential property markets in a booming economy: New evidence from beijing." *Journal of Urban Economics*, 63, 743–757.

Zhu, Hongtu, Jianqing Fan, and Linglong Kong (2014), "Spatially varying coefficient model for neuroimaging data with jump discontinuities." *Journal of the American Statistical Association*, 109, 1084–1098.

Zimmerman, Dale L and M Bridget Zimmerman (1991), "A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors." *Technometrics*, 33, 77–91.