Università
degli Studi
di Padova

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE STATISTICHE
CICLO XXVI

Behaviour Risk Factor Surveillance Data Analysis Using Varying Coefficient Models

Direttore della Scuola: Ch.ma Prof.ssa Monica Chiogna

Supervisore: Ch.mo Prof. Stefano Campostrini

Co-supervisore: Ch.mo Prof. Carlo Gaetan

Dottoranda: Shireen Assaf

31 Gennaio 2014

Acknowledgements

While completing the final version of this thesis, I cannot help but reflect on the many people who have made this journey possible. I would like to express my heartfelt gratitude to my supervisor Professor Stefano Campostrini and my co-supervisor Professor Carlo Gaetan for their insightful comments, constructive feedback and advice throughout my research period, I have learned a lot from our discussions.

From the start I was lucky to have such supportive and friendly colleagues: Luca Sartore, Roberta Pappadà, Akram Yazdani, Lorenzo Maragoni, Erlis Ruli, Ivan Luciano Danesi, and Md Abud Darda – all of who made the first year of the PhD possible and memorable. I would like to express a special thank you to Luca Sartore specifically for his help with learning and using R for my thesis, but also for his overall encouragement and support these past years and for teaching me to trust my instincts.

I would like to thank the PASSI network in Italy for providing me with the dataset used in the Italian analysis of this thesis. In addition, I am grateful to the staff of the Division of Behaviour Surveillance in the Center for Disease Control and Prevention, and especially Dr. Carol Gotway Crawford, for hosting me and providing me with the assistance I needed during my research abroad period in the United States, which provided a rich and nurturing environment to study and explore new ideas and applications. I am also grateful to the Department of Statistical Sciences for giving me this opportunity, as well as the administrative and technical staff which have been very helpful over the past three years.

A special thanks also goes to my friends and family, and especially my sister Jenin Assaf and brother-in-law Stefano Pasquini, as well as to my friends Simmaco Gionti and Dayane Favarin Pereira all of who have been very supportive during my stay in Italy.

I also want to thank my sisters, Nabila, Rema and Dena Assaf as well as my parents Karen and Said Assaf, for their encouragement and for instilling in me an interest in research and a sense of curiosity – all of which were instrumental in finalizing this thesis.

Shireen Assaf

Padova,
January 19, 2014

Abstract

There is a high potential of information available in Behaviour Risk Factor Surveillance (BRFS) data, and especially for studying trends, as these data collect information in an ongoing and almost continuous manner for long periods of time. In order to account for the complex and dynamic relationships between the variables and avoid the aggregation of measures so as not to lose information in variability, the use of varying coefficient models with non-parametric techniques have been studied. These models allow the study of the trends and inter-relationships in the effects of the variables on the outcome of interest either over time or space, therefore providing valuable information for health policy interventions.

A comparison of the possible estimation techniques, using the Italian surveillance data, has resulted in the selection of P-splines for estimation due to the flexibility in their use and the faster computation times. This estimation method was applied for a time varying coefficient model for a smoking status outcome variable using Italian surveillance data, and a time varying coefficient model for an obesity status outcome variable using U.S.A. surveillance data. The results of these models provide coefficient plots in which one can observe which subgroups of the population have an effect on the outcome which is changing over time. A spatial varying coefficient model was also studied for one point in time using smoothing spline estimation with tensor product smooths, and the maps produced from this model were able to show how the probabilities of the outcome variable (obesity) are changing across the counties of a U.S. state within each population subgroup. The strengths and limitations of these methods are discussed, as well as recommendations for further research such as the study of a spatial-temporal model using health surveillance data. Notwithstanding few limitations, the varying coefficient model represents an effective approach proving to produce interesting results (not accessible with the usual standard epidemiological approach) in this particular field of application and with BRFS data.

Sommario

C'è un alto potenziale di informazioni disponibili nei dati di sorveglianza sui fattori comportamentali di rischio, specialmente per lo studio di tendenze evolutive nella popolazione: questi dati vengono infatti raccolti in modo quasi continuo e per lunghi periodi temporali. Per spiegare le relazioni complesse e le dinamiche tra le variabili, evitando l'aggregazione di misure per non perdere l'informazione sulla variabilità, è stata studiata la possibilità di applicare a questi dati modelli a coefficienti variabili con tecniche non parametriche. Questi modelli permettono lo studio delle tendenze e delle interrelazioni negli effetti delle variabili sul risultato di interesse nel tempo o nello spazio, fornendo quindi informazioni preziose per gli interventi di politica sanitaria.

Un confronto delle possibili tecniche di stima, utilizzando i dati di sorveglianza italiani, ha portato alla selezione delle P-spline perché più flessibili nel loro utilizzo e computazionalmente più veloci. Questo metodo di stima è stato applicato ad un modello a coefficienti variabili nel tempo per lo studio di una variabile risposta sulle abitudini al fumo utilizzando i dati di sorveglianza italiani. Inoltre, è stato studiato un modello a coefficienti variabili nel tempo per l'esito di una variabile risposta sullo stato di obesità utilizzando i dati di sorveglianza statunitensi. Dai risultati derivanti dall'applicazione di questi modelli vengono prodotti grafici (di coefficienti e OR) utili per osservare quali sottogruppi della popolazione presentano effetti che stanno evolvendo nel tempo. Anche un modello a coefficienti spazialmente variabili è stato studiato (in riferimento ad un determinato momento temporale) utilizzando stime spline con lisciature fornite dal prodotto tensoriale. Le mappe prodotte da questo modello sono state in grado di evidenziare come le probabilità della variabile risposta (obesità) stanno cambiando attraverso le contee di un stato negli USA all'interno di ogni sottogruppo della popolazione. I punti di forza e i limiti di questi metodi sono stati discussi, inoltre alcune raccomandazioni per ulteriori ricerche sono state proposte per lo studio di un modello spazio-temporale utilizzando i dati di sorveglianza sanitaria. Nonostante alcune limitazioni, il modello a coefficienti variabili rappresenta un approccio efficace dimostrando di produrre risultati interessanti (non accessibili con il consueto e tipico approccio epidemiologico) in questo particolare campo applicativo.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Behaviour risk factor surveillance (BRFS) data can be a great source of information for studying changes of various health outcomes and risk factors. Results obtained from surveillance data analysis are vital for forming health policy interventions and planning, particularly when the analysis inform about temporal or spatial trends. Sometimes, the complexity of the relationship among relevant variables requires to know if and how the effects of various independent variables on a certain outcome are themselves changing over time or space. Varying coefficient models (VCM) with non-parametric techniques can be used to catch the dynamics of BRFS data, being a useful method which allows coefficients to vary with time or space using smooth functions. This allows for the study of the changing effects of possible determinants on a health outcome in order to better inform policy interventions.

Behaviour Risk Factor Surveillance (BRFS) data has a frequent data collection design (usually monthly data), and has developed from the historical registry of infectious diseases to a more detailed look at health risks and behaviours for the study of non-communicable diseases (Lee and Thacker, 2010). There are several health surveillance systems that provide this type of data, one of the longest running is the U.S.A. Behavioral Risk Factor Surveillance System (BRFSS) which has collected monthly data since 1984 (Mokdad, 2009). Another surveillance system with a similar design is the Italian Progressi delle Aziende Sanitarie per la Salute (PASSI) surveillance system which has begun collecting monthly data on health behaviours and risk factors in 2007 (PASSI - Coordinating technical group of the behavioural risk factor system, 2013). These two sources of data will be used in the analysis to demonstrate the benefit and practicality of using varying coefficient models for the analysis of trends for the study of health risks and outcomes, as well as some of the limitations.

1.2 Main Contributions of the Thesis

Although there are many articles which discuss varying coefficient models, the application of this type of model to data exclusively from a BRFS system in the health and epidemiological setting, and particularly for the analysis of large sample sizes for long periods of observation, has not been performed. The main objective of this research is to study the use of varying coefficient models on BRFS data to explore the practicality and feasibility of this method for this kind of data which can reach very large sample sizes.

One of the main traditional methods used for trend analysis for the analysis of health outcomes, other than the simple parametric regression which includes time as a covariate along with other covariates, includes time series methods in which the main assumption is the presence of a dependence structure between the observations. In time series analysis the observations are usually aggregated for each unit of time before analysis is performed; for instance for the analysis of monthly means, proportions or sums (Diggle, 1990). Aggregation of observations is also conducted in cohort trend analysis, another method often used for analysis of health outcome trends. This type of analysis can be performed on a dataset of pooled independent cross-sectional surveys in which individual observations are aggregated according to certain cohorts (such as an age cohort) and then analysed over time (Deaton, 1985). Finally, there is also longitudinal data analysis which involves using data that contain multiple measurements on the same observation with time and therefore the observations are serially correlated (Diggle, 2002). This type of analysis is usually limited to a specific health problem and has a problem of attrition due the nature of the data collection method. On the other hand, in BRFS data, the observations are not necessarily dependent since there is a new random sample of individuals taken every month. Therefore there are no problems of attrition as in longitudinal data, although causality can not be determined. BRFS data also covers a wide variety of health topics and risk factors. In addition, the type of analysis proposed using varying coefficient models does not require the aggregation of observations. This is favoured as it would be of more interest to use all the observations in the analysis to account for the variations between them. Therefore combining the use of the varying coefficient model methods and BRFS data can provide some advantages compared to the traditional methods for trend analysis.

The types of analysis discussed above through the traditional trend analysis methods all provide one value for each parameter for the entire period of observation; this assumes that the parameters are constant with time. The use of varying coefficient models however, where the parameters are allowed to vary with time, can show which parameters are actually time varying and which in fact are remaining constant. In other words, the interest is not to study the overall trend of a certain outcome with time, but whether the effects on the outcome are themselves changing with time. This kind of dynamic nature of the data would not be captured in the traditional time series, longitudinal or cohort analysis methods. The results of this kind of analysis can be very useful for identifying certain risks or characteristics that are changing for purposes of intervention and planning.

The thesis will begin with a literature review of varying coefficient models which mainly summarizes the estimation methods used for these models, namely estimation using smoothing splines and estimation using local regression methods. This will be followed by the methodology chapter to describe how the varying coefficient models were constructed for the analysis of the BRFS data. The thesis will then have two result chapters. One of the result chapters will discuss the Italian PASSI data to compare five different estimation techniques using a smoking status binary outcome variable for a time varying coefficient model. The estimation methods which will be compared are estimation using polynomial splines, penalized spline regression, P-spline estimation using the `gam` function of the `mgcv` pacakge, P-spline estimation using the `bam` function of the `mgcv` package designed for large datasets, and smoothing spline estimation with cubic regression spline and the `bam` function. While all these methods are using spline methods for estimation, the computational times as well as their construction can differ, and therefore the comparison made is to highlight the most practical and feasible method to use in this setting. The recommended method is then used to further describe the smoking status time varying coefficient model using odds ratio plots which show how the odds ratios (i.e. the coefficients) are changing over time. The second results chapter will discuss the use the U.S.A. BRFSS data to demonstrate the application of the method to a very large sample size (2,065,689 observations) for studying temporal trends, as well as applying the method to one U.S. state in one year for studying a spatial varying coefficient model, thus showing how the proposed method can also be used to study spatial trends. In the U.S. analysis, an obesity status binary outcome variable is used, and for the first part of the analysis a time varying coefficient model is constructed to again produce odds ratio plots, and in the second part a spatial varying coefficient model is constructed to produce probability maps for each of the covariates categories in the varying coefficient model. The final aim is to demonstrate whether varying coefficient models can be used as a tool for health surveillance data analysis, and provide recommendations for their use in this setting.

Chapter 2

Literature Review

2.1 Surveillance Systems

2.1.1 Background

The collection of health data for the purpose of informing public health interventions can be first tracked back to the time of the pneumonic plague in 1348 by the Venetian Republic, which monitored infected people aboard ships in order to be quarantined (Declich and Carter, 1994). This was then developed further for the purpose of monitoring infectious diseases (Lee and Thacker, 2010). However, as non-communicable diseases became more prominent the need for monitoring these diseases as well as their risk factors became more important (Campostrini *et al.*, 2011). A more encompassing surveillance system was needed to not only monitor and track diseases but also their risk factors as well as the social determinates that can effect their development. The concept of population surveillance or public health surveillance was then adopted which contains three main characteristics of: systematic or continuous collection of data; data analysis; and dissemination of information and findings (Declich and Carter, 1994). The importance of including risk factors to public health surveillance systems, and particularly the main four risks factors of smoking, physical inactivity, diet and alcohol consumption, eventually led to the name Behavioural Risk Factor Surveillance (BRFS).

One of the main purposes of health surveillance systems is to study the trends of diseases and their risk factors, as well as the social determinants that can affect disease (Declich and Carter, 1994; Campostrini *et al.*, 2011).

It would not be sufficient for this purpose to rely on cross-sectional health surveys which are usually conducted every few years, for example the Demographic and Health Surveys (DHS) which are performed in several countries usually every four years or more. Changes in behaviours could be occurring more rapidly in a population and therefore a more rapid and continuous data collection system is required to study trends in a public health setting, as is conducted by public health surveillance or BRFS systems. One of the first to establish a stable and well developed BRFS was the United States of America in 1984 (U.S. BRFSS - Behavioural Risk Factor Surveillance System) and it is still ongoing (Mokdad, 2009). However, other countries have also adopted this type of surveillance system including Italy, Canada, Brazil and Australia (Campostrini and McQueen, 2011).

2.1.2 Methodological issues

Since the main purpose of a BRFS is to collect information on changing behaviours which can affect diseases, a theoretical understanding of how and which behaviours can have this effect is essential in order to know what needs to be measured and in which manner. Therefore a public health framework is required so as not only to include questions on disease and well known risk factors, but also to include variables which measure social determinants. Therefore the BRFS surveys usually include questions on demographic aspects including sex, age marital status, education and location, as well as some measure of income status. In the U.S. BRFSS, there is a questionnaire that contains a fixed core of questions that is asked every year and a rotating core of questions asked every other year or more (Mokdad, 2009). The rotating core would include questions on specific topics which are found to require more in-depth information, or which are not expected to change very rapidly and therefore do not require more frequent measurements. In addition, while the core questions, which must include all the questions on social determinates and certain diseases, are asked in all of the 50 U.S. states and four territories, each state has the option to include in the survey additional questions from 19 optional modules on various topics (Mokdad, 2009). Therefore, one must consider which questions to select for creating the variables required in the analysis, as they may not be available in all the years or all the states and regions.

The method of data collected for BRFS is usually by telephone interview though computer-assisted telephone interview systems (CATI), and using samples selected by random-digit-dialed (RDD) method (Campostrini and McQueen, 2011; Mokdad, 2009). However, other methods can also be used depending on the country including face-to-face, web-based, mail questionnaires, and mixed mode surveying (Campostrini and McQueen, 2011). The U.S. BRFSS originally used landline telephone interviews using RDD sampling, however an increase in cell phone use as well as other factors have caused a decrease in response rates (Mokdad, 2009; Pierannunzi, 2012). This decrease can affect the quality of the estimates although perhaps to a small extent as found by Fahimi *et al.* (2008) when a comparison was made with other national surveys in the United States. However, in an effort to increase response rates and reduce bias in estimates due to the increase in cell phone only households, a mixed mode method was tested in the United States which includes cell phones, mailing of advance letters to potential sample members, and mail surveys with telephone follow-ups (Hu *et al.*, 2011; Mokdad, 2009; Pierannunzi, 2012). As a result, it was found that including cell phones in the sampling frame is capturing households previously missed by using only land lines, and as of 2011 the U.S. BRFSS public release data set began including cell phones (Center for Disease Control and Prevention (CDC), 2014). Changes in methodology however, can give rise to certain challenges. For instance, the inclusion of cellphones in the U.S. BRFSS will not only increase costs but can also have an effect on the estimates due to the changes required for weighting method used. While initially a post-stratification method was used for weighting survey data, the inclusion of cell phones required a raking or iterative proportional fitting weighting method (Pierannunzi, 2012). Comparison of estimates using these two weighting methods has shown a slight change in the prevalence estimates, however the shape and slope of the trends of these prevalences do not change much (Pierannunzi, 2012). This break in the methodology creates a problem for the study of trends using data before and after the inclusion of cell phones users, and careful consideration is required if this is attempted for analysis of trends.

2.1.3 Trend analysis using BRFS data

Many researchers have used BRFS data (mainly the U.S. BRFSS or the South Australian Monitoring and Surveillance System) to study trends and changes of various health outcomes. The statistical methods used are usually very similar, and is mainly a question of comparing the prevalence or mean estimates of a certain health related outcome for each year included in the analysis and perhaps across various characteristics (Ahluwalia *et al.*, 2005; Ashford *et al.*, 2010; Fan, 2013; Flegal *et al.*, 2002; Serdula *et al.*, 2004; Shi *et al.*, 2011; Simpson *et al.*, 2003; Taylor *et al.*, 2013; Zack *et al.*, 2004). Some articles have gone further in the analysis to produce a logistic or linear regression with year as a variable to study the trend of the outcome (Ashford *et al.*, 2010; Fan, 2013; Shi *et al.*, 2011; Taylor *et al.*, 2013; Troost *et al.*, 2012; Zack *et al.*, 2004). Simpson *et al.* (2003) did not use regression but simply observed the absolute prevalence differences between the first and last year in the analysis to observe changes. Jia and Lubetkin (2009) used time series analysis to study trends in physically and mentally unhealthy days from 1993 to 2006 of the U.S. BRFSS. The model used contained a trend component which was defined as local linear trend model that allowed for the estimation of the trend in the means of the outcome over time, and also after controlling for several independent variables and seasonal effects. Time series analysis of U.S. BRFSS data was also used to study policy interventions in the U.S. by Campostrini *et al.* (2006) to study the impact of changes in the law on drinking and driving, and by Ma *et al.* (2013) to study the impact of an increase of cigarette tax on smoking prevalence. Therefore, the common methods used for trend analysis of BRFS data are to study the trends of the outcome variable and at times across different characteristics. The trend is often assumed to be linear and non-parametric methods are not used. In addition, the measures used in the analysis are usually aggregated to compare the prevalence or means for different time periods, or for use in time series models.

2.2 Varying Coefficient Models

Varying coefficient models can be used to capture the changing affects of the covariates on the response. It is a favoured model from a practical sense as well since at times it is implausible to assume that the impacts of the coefficients on the response is constant (Fan and Zhang, 2008) and especially when long periods of observation is involved. In addition, it does not require the aggregation of measure before analysis therefore reducing loss of information. It is also a flexible method for modelling interactions between a factorial covariate for instance and a metrical one, versus the methods used in semi-parametric or generalized additive models where effects of covariates are modelled additively and without interactions (Hastie and Tibshirani, 1993; Kauermann and Tutz, 1999). There are two main non-parametric estimation methods used for varying coefficient models; estimation using smoothing splines and estimation using local regression methods. The literature on these estimation methods usually discuss the Gaussian case however all the methods can be adapted for a non-Gaussian distributed response as is usually required for the analysis of health data. The literature review gives an overview of the theory behind the estimation of these models after introducing the model below.

If we define Y as a normally distributed random variable with given covariates $(U, X_1, \ldots, X_p)^T$; then a standard varying coefficient model has the form

$$Y = \sum_{j=1}^{p} a_j(U) X_j + \varepsilon, \tag{2.1}$$

where $E(\varepsilon | U, X_1, \ldots, X_p) = 0$ and $\text{var}(\varepsilon | U, X_1, \ldots, X_p) = \sigma^2(U)$ (Hastie and Tibshirani, 1993; Fan and Zhang, 1999, 2008). The variable U is referred to as the effect modifier and can technically be any covariate including time. In addition, as described by Hastie and Tibshirani (1993), the U_j terms can be scalar or vector valued and the functions $a_j(U_j)$ can be modelled by flexible parametric functions (example polynomials, Fourier series or piecewise polynomials) or non-parametric functions. If X is a binary variable then having $Xa(U)$ means that there is a separate curve which corresponds to each value of X. More generally, if we have factor variable F where each X_j represents a coding for the levels of F then we would have $Fa(U)$ which represents an interaction between the factor F and the function $a(U)$. One important type of model discussed by Hastie and Tibshirani (1993) is if we have the same variable for the U_j then the model is a varying coefficient model with a single modifying variable. This is usually used for analysis of repeated measurements where U_j would be time so that the model would be of the form $Y = a_0(t) + X_1(t)a_1(t) + \ldots + X_p(t)a_p(t)$(Hastie and Tibshirani, 1993).

The study of the generalized varying coefficient model began with Hastie and Tibshirani (1993) who described the different forms varying coefficient models can take, although the description of the estimation method was only shown for the standard model. These models were also discussed by Fan and Zhang (2008), Cai *et al.* (1999), Cheng *et al.* (2009) and Marx (2010). Suppose that we have a variable Y with a distribution that is from an exponential family, and this distribution depends on the parameter η, then a generalized varying coefficient model has the form

$$\eta = a_0 + X_1 a_1(U_1) + \ldots + X_p a_p(U_p)$$

where $\eta = g(\mu)$ with $\mu = \mathrm{E}(Y)$, and $g()$ as the link function. The covariates are $\mathbf{x} = (X_1, \ldots, X_p)^T$ with the effect modifier covariate $\mathbf{u}$. Hastie and Tibshirani (1993) describes how the generalized varying coefficient model can be reduced to more common models, for example:

- if $a_j(U_j) = a_j$ then the model is the generalized linear model,
- if all the terms are linear and/or have the form where $X_j = c$ so that the j$^{\text{th}}$ term would be $a_j(U_j)$, then the model is a generalized additive model, and
- if we have a linear function, i.e. $a_j(U_j) = a_j U_j$ then this is a product interaction of the form $a_j X_j U_j$.

2.3 Estimation of Varying Coefficient Models

The estimation, testing and asymptotics of varying coefficient models have been covered by a number of articles. Researchers studying the above model have used two main estimations techniques: local regression and estimation using splines (polynomial spline, penalized spline regression or smoothing spline). Parametric methods are not favoured for estimation due to the lack of flexibility of these methods as well as the strong assumptions it requires which can lead to misspecification of the data and large bias (Hastie and Tibshirani, 1993; Fan and Zhang, 2008).

Most of the literature appears to focus on standard varying coefficient models with $a(U)$ modelled using non-parametric methods. However, the same estimation procedures can be adapted for the generalized varying coefficient models. For instance, Cheng *et al.* (2009) applied the generalized varying coefficient model for a binary response variable of infant mortality and Cai *et al.* (1999, 2000) used Poisson and binary response variables. In addition Marx (2010) described polynomial P-spline smoothing estimation for both the standard and generalized varying coefficient models. Although Hastie and Tibshirani (1993) discussed the generalized varying coefficient model, the estimation procedures were shown only for the standard case; however an example was provided for a binary outcome for heart disease. Hastie and Tibshirani (1993) described that extensions of the estimation used in the standard varying coefficient model to the generalized model usually involves inserting a Newton-Raphson type algorithm. What follows is a review of the literature for the non-parametric methods used for estimating both the standard and generalized varying coefficient models.

2.3.1 Estimation using splines

Before discussing the various forms of estimation using splines, a short description of splines is needed. To define splines we begin with a regression model of the form

$$y_i = s(x_i) + \epsilon_i,$$

which is minimized to find the estimates $\hat{s}_n(x_i)$ by

$$\sum_{i=1}^{n} (y_i - \hat{s}_n(x_i))^2$$

(Wasserman, 2006). The function $s(x_i)$ can be any function including linear functions, polynomials or splines. For instance, if we have a linear function, then the we have the simple linear regression model with the minimization leading to the least squares estimator. Splines are then special piecewise polynomials joined by knots (i.e. specific position points), and the various types of splines depend on the basis defined as well as the order of the spline and the knot placement. To have an M^{th}-order spline, there is a piecewise $M-1$ degree polynomial with $M-2$ continuous derivatives at the knots (Wasserman, 2006). For example, starting with a set of ordered knots in a certain interval, in a cubic spline ($M = 4$) the function $s(\cdot)$ is a continuous cubic polynomial function over the knots, and it has continuous first and second derivatives at these knots (Wasserman, 2006).

A commonly used spline is that using a B-spline basis functions which are favoured because they have compact support that can speed up calculations

(Wasserman, 2006; Hastie *et al.*, 2009). To describe B-splines the knot sequence first needs to be defined. Starting with the boundary knots $\xi_0 < \xi_1$ and $\xi_K < \xi_{K+1}$, there is an augmented knot sequence τ such that

$$\begin{aligned} \tau_1 &\leq \tau_2 \leq \ldots \leq \tau_M \leq \xi_0, \\ \tau_{j+M} &= \xi_j, \text{ for } j = 1, \ldots, K, \quad \text{and} \\ \xi_{K+1} &\leq \tau_{K+M+1} \leq \tau_{K+M+2} \leq \ldots \tau_{K+2M} \end{aligned}$$

(Hastie *et al.*, 2009). Usually the extra τ knots that are beyond the boundary knots are set to be equal, and equal to ξ_0 and ξ_{K+1} respectively, i.e. $\tau_1 = \ldots = \tau_M = \xi_0$ and $\xi_{K+1} = \tau_{K+M+1} = \ldots = \tau_{K+2M}$ (Hastie *et al.*, 2009). The i^{th} B-spline basis function is then denoted by $B_{i,m}(x)$ for a basis of order m with the knot-sequence $\tau, m \leq M$. The B-spline basis functions are then defined recursively as follows:

$$B_{i,1}(x) = \begin{cases} 1 & \text{if } \tau_i \leq x < \tau_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

for $i = 1, \ldots, K + 2M - 1$. Then for $m \leq M$ we have:

$$B_{i,m} = \frac{x - \tau_i}{\tau_{i+m-1} - \tau_i} B_{i,m-1} + \frac{\tau_{i+m} - x}{\tau_{i+m} - \tau_{i+1}} B_{i+1,m-1}$$

for $i = 1, \ldots, K + 2M - m$ (Hastie *et al.*, 2009). Therefore, this recursive process can be used for any order B-spline.

Now to find the estimates $\hat{s}(x)$ using spline estimation, we first write

$$s_n(x) = \sum_{j=1}^{N} \gamma_j B_j(x) \tag{2.2}$$

with $B_1, \ldots, B_N$ defined as the basis for splines such as B-splines. Then to find the coefficients $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_N)^T$, the following is minimized

$$(\mathbf{y} - \mathbf{B}\boldsymbol{\gamma})^T(\mathbf{y} - \mathbf{B}\boldsymbol{\gamma}) + \lambda\boldsymbol{\Omega}$$

(Wasserman, 2006). In smoothing spline estimation as shown below, a penalty is added to the objective function 2.2 above. In this case the solution to the minimization is $\hat{\boldsymbol{\gamma}} = \left(\mathbf{B}^T\mathbf{B} + \lambda\boldsymbol{\Omega}\right)^{-1}\mathbf{B}^T\mathbf{y}$, where $\lambda\boldsymbol{\Omega}$ is the added penalty which shrinks the regression coefficients towards a subspace resulting in a smoother fit (Wasserman, 2006) and is described further in the smoothing spline estimation section.

In the case of varying coefficient models, the coefficients are smooth functions and therefore the model in equation 2.1 can be written as $y_i = s(x_{i1} \ldots, x_{ip}, u_i) + \varepsilon_i$, where $s(x_{i1} \ldots, x_{ip}, u_i) = \sum_{j=1}^{p} a_j(u_i)x_{ij}$. The goal is then to find $\hat{a}_j(U)$, and with spline methods this is performed either with added penalties (as in smoothing spline and penalized regression spline estimation), or with no penalties (as in polynomial spline estimation).

Polynomial Spline Estimation

Polynomial splines, which are piecewise polynomials joined together smoothly at a set of interior knot points, was used for estimation by Huang *et al.* (2004, 2002) for analysis of time-varying coefficient models using longitudinal data. Therefore, the effect modifier variable of time (t) in this case replaces the above notation of having U represent the effect modifier variable. It is important to note that the data in these longitudinal studies are data where the same measurements are made on the same subject with time, and this differs from data where the measurements are repeated with time but with a new random sample of subjects for each time period. However, the same estimation procedures are applied, although in longitudinal data there is the issue of having dependence between the observations.

Polynomial spline estimation applies to both time-invariant and time-dependent covariates. The procedure is very similar to the smoothing spline estimation, however there is no penalty and the smoothing parameter is defined differently. The estimation begins by approximating the functional coefficients a_l to be estimated by spline functions, i.e. $a_l(t) \approx \sum_{k=1}^{K_l} \gamma_{lk} B_{lk}(t)$, as was done in the description of the spline estimator in equation 2.2. In polynomial spline estimation, for each $l = 0, \ldots, L$ there is a basis $\{B_{lk}(\cdot), k = 1, \ldots, K_l\}$ from a linear space $\mathcal{G}_l$ of spline functions with a fixed degree and knot sequence. Differently from the smoothing spline method where λ is the smoothing parameter (as shown in the next section), in this method the K_l (or the number of knots) play the role of the smoothing parameters and can be selected by using the cross-validation method, AIC or BIC (Huang and Shen, 2004). (Huang and Shen, 2004) showed that the selection of models using AIC was the preferred method. Having a different number of knots for each coefficient function allows for these functions to have different amounts of smoothing. Huang *et al.* (2004, 2002) both used B-splines for the basis functions $B_{lk}(\cdot)$ in their estimation due to their good numerical properties. After substituting the coefficient functions a_l with the basis approximation, the varying coefficient model for a specific knot sequence becomes

$$y_{ij} \approx \sum_{l=0}^{L} \sum_{k=1}^{K_l} x_{ijl} B_{lk}(t_{ij}) \gamma_{lk} + \epsilon_{ij}.$$

The extra subscript in this model compared to the smoothing spline model is due to the use of longitudinal data by Huang *et al.* (2004, 2002), and therefore it is used to indicate the subject which is followed over time. This would not be the case for instance when analysing surveillance data where the subjects are not followed with time. From the model above, the estimate of γ_{lk} can be found by minimizing

$$\sum_{i=1}^{n} w_i \sum_{j=1}^{n_i} \left(y_{ij} - \sum_{l=0}^{L} \sum_{k=1}^{K_l} x_{ijl} B_{lk}(t_{ij}) \gamma_{lk} \right)^2 ,$$

where w_i are the weights which can be one if equal weight is given to each single observation or $1/n_i$ to give equal weight to each subject (since in longitudinal data one subject has repeated observations). To write this in matrix notation the following terms are defined

$$\mathbf{B}(t) = \begin{pmatrix} B_{01}(t) & \dots & B_{0K_0}(t) & 0 & \dots & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & \dots & 0 & B_{L1}(t) & \dots & B_{LK_L}(t) \end{pmatrix},$$

for $i = 1, \dots, n, j = 1, \dots, n_i$, and $l = 0, \dots, L$. Also we define

$$\begin{aligned} \mathbf{R}_{ij}^T &= \mathbf{X}_i^T(t_{ij}) \mathbf{B}(t_{ij}), & \mathbf{R}_i &= (\mathbf{R}_{i1}, \dots, \mathbf{R}_{in_i})^T \\ \boldsymbol{\gamma}_l &= (\gamma_{l0}, \dots, \gamma_{iK_l})^T, & \boldsymbol{\gamma} &= (\boldsymbol{\gamma}_0^T, \dots, \boldsymbol{\gamma}_L^T)^T \\ \mathbf{y}_i &= (y_{i1}, \dots, y_{in_i})^T, & W_i &= \operatorname{diag}(w_i, \dots, w_i) \end{aligned}$$

then we have

$$\sum_{i=1}^{n} (\mathbf{y}_i - \mathbf{R}_i \boldsymbol{\gamma})^T \mathbf{W}_i (\mathbf{y}_i - \mathbf{R}_i \boldsymbol{\gamma}).$$

This is minimized to give the estimator

$$\hat{\gamma} = \left(\sum_i \mathbf{R}_i^T \mathbf{W}_i \mathbf{R}_i \right)^{-1} \sum_i \mathbf{R}_i^T \mathbf{W}_i \mathbf{y}_i.$$

The spline estimate of $\mathbf{a}(t)$ can be found by $\hat{\mathbf{a}}(t) = \mathbf{B}(t)\hat{\boldsymbol{\gamma}} = (\hat{a}_0(t), \ldots, \hat{a}_L(t))^T$, where $\hat{a}_l(t) = \sum_{k=1}^{K_l} \hat{\gamma}_{lk} B_{lk}(t)$ (Huang *et al.*, 2004).

Smoothing Spline

Smoothing spline estimation was used as a the method for estimation of varying coefficient models by Hastie and Tibshirani (1993) which was the one of the original articles on varying-coefficient models. It was also used by Hoover *et al.* (1998) and Chiang *et al.* (2001) for longitudinal data analysis.

The estimation details begin with the observations $y_1, \ldots, y_n$ and x_{ij} and u_{ij} of the predictors X_j and U_j for the model $y_i = x_{i1}a_1(u_{i1}) + \ldots + x_{ip}a_p(u_{ip}) + \epsilon_i$ (data are not longitudinal). To find the estimators, the penalized sum of square residuals

$$\sum_{i=1}^{n} \left\{ y_i - \sum_{j=1}^{p} x_{ij} a_j(u_{ij}) \right\}^2 + \sum_{j=1}^{p} \lambda_j \int a_j''(u_j)^2 \mathrm{d}u_j$$

is minimized, where λ_j is the smoothing parameter which penalizes the roughness of the functional coefficients a_j. The first term is the summation of the square residuals and the second term is the summation of the penalties ($\lambda\boldsymbol{\Omega}$ presented previously) for each coefficient function. The a_j are again expressed in terms of basis functions $a_j(u_{ij}) = \sum_{l=1}^{n_j} \gamma_{ij} B_{jl}(u_{ij})$, where n_j are the number of unique values of U_j (i.e. the number of knots). The $B_{jl}(U_j)$ are basis functions for the j$^{\text{th}}$ variable, these functions can be polynomial bases, Fourier bases, natural cubic splines or B-splines functions. By letting $\mathbf{a}_j$ represent $a_j(u_{ij})$ evaluated at the n observed values of U_j so that $\mathbf{a}_j = \mathbf{B}_j\boldsymbol{\gamma}_j$, where $\mathbf{B}_j$ is a matrix of spline functions, we can rewrite the above penalized least squares equation above in matrix form as

$$\left\| \mathbf{y} - \sum_{j=1}^{p} \mathbf{D}_j \mathbf{B}_j \boldsymbol{\gamma}_j \right\|^2 + \sum_{j=1}^{p} \lambda_j \parallel \boldsymbol{\gamma}_j \parallel_{\boldsymbol{\Omega}_J}^2,$$

where $\mathbf{D}_j$ is the diagonal matrix with the n observed values of X_j on the diagonal. The last term contains the penalty seminorm $\parallel \boldsymbol{\gamma}_j \parallel^2_{\Omega_J}$, with $\boldsymbol{\Omega}_j$ having the ik$^{\text{th}}$ element as $\int B^{i"}_j(r)B^{k"}_j(u)\mathrm{d}u$. Minimizing the above gives $\hat{\gamma}$ which can then be used to find $\hat{a_j}$ by $\hat{a}_j(u_{ij}) = \sum_{l=1}^{n_j} \hat{\gamma}_{ij}B_{jl}(u_{ij})$ using backfitting procedures. The smoothing parameters $\lambda_1 \ldots \lambda_p$ are fixed, they control the amount of smoothing and they must be estimated usually by cross-validation or generalized cross-validation.

Penalized Spline Regression

The polynomial spline approach described above required knowledge of the location and the number of knots, however this is unknown and needs to be found by cross-validation or AIC. Similar to smoothing spline estimation, in penalized spline regression, a penalty is added to allow for automatic knot selection by using all or a reasonable number of knots but then constrain their influence (Ruppert *et al.*, 2003). The type of penalty however differs from the penalty used in smoothing spline estimation. Here the penalty used is similar to that used in ridge regression. Following the matrix notation described for smoothing spline estimation, the model to be minimized is

$$\left\| \mathbf{y} - \sum_{j=1}^{p} \mathbf{D}_j \mathbf{a}_j \right\|^2 + \sum_{j=1}^{p} \lambda_j^2 \mathbf{a}_j^T \mathbf{P} \ \mathbf{a}_j,$$

where $\mathbf{P}$ can be an identity matrix or other matrix of operators such as difference operators as used in P-splines as will be shown. This type of penalized estimation will shrink all coefficients of the spline basis functions toward zero. The smoothing parameter λ is chosen by cross-validation as in smoothing spline estimation.

P-splines Estimation:

As described by Eilers and Marx (1996) the main advantages of using P-splines versus polynomial spline estimation using B-splines are that P-splines have no boundary effects, they conserve moments of the data and have a polynomial curve fits as limits. For finding the estimates and achieving smoothness the method first uses a rich regression basis to overfit the smooth coefficient vector with a modest number of equally spaced B-splines, then to ensure the proper amount of smoothness P-splines are added which are constructed by placing a difference penalty on the coefficients of adjacent B-splines. The penalized loss in this case can be written as

$$\left\|\mathbf{y}-\sum_{j=1}^{p}\mathbf{D}_j\mathbf{B}_j\boldsymbol{\gamma}_j\right\|^2+\sum_{j=1}^{p}\lambda_j\parallel \boldsymbol{\Delta}_d\boldsymbol{\gamma}_j\parallel^2,$$

where again $\mathbf{D}_j$ is the diagonal matrix with the n observed values of X_j on the diagonal as in the smoothing spline estimation, $\mathbf{B}_j$ is a matrix of spline functions, and $\boldsymbol{\Delta}_d$ is a matrix which constructs the d^{th} differences of $\boldsymbol{\gamma}$. Then by defining $\mathbf{R}=\mathbf{D}_j\mathbf{B}_j$ as in the polynomial spline estimation, the above loss is minimized to find $\hat{\boldsymbol{\gamma}}$ by

$$\hat{\gamma}=\left(\mathbf{R}^T\mathbf{R}+\mathbf{P}\right)^{-1}\mathbf{R}^T\mathbf{y}.$$

The matrix $\mathbf{P}$ = block diag$(\lambda_0\boldsymbol{\Delta}_d^T\boldsymbol{\Delta}_d,\ldots,\lambda_p\boldsymbol{\Delta}_d^T\boldsymbol{\Delta}_d)$ has a block diagonal structure that breaks the linkage of the penalization from one smooth term to the next. There is a separate λ for each term and this is chosen by cross-validation or minimum AIC (Marx, 2010). The penalty in P-spline estimation can have different degrees so that the first, second or third difference can be taken. Eilers and Marx (2002) recommend using at least a second degree difference penalty with either a quadratic or cubic B-spline basis.

The adaptation of the P-spline to a generalized varying coefficient model was also described by Eilers and Marx (1996) and Marx (2010) and this simply involves the maximization of

$$l(\boldsymbol{\gamma})-\sum_{j=1}^{p}\lambda_j\parallel \boldsymbol{\Delta}^d\boldsymbol{\gamma}_j\parallel^2,$$

where $l(\boldsymbol{\gamma})$ is the log-likelihood function. Here the penalty term is subtracted from the log-likelihood function to discourage roughness of any varying coefficient vector. A Fisher's schoring algorithm is used to find the estimates.

Wang *et al.* (2008) used penalized spline regression estimation for estimating a varying coefficient model for the analysis of repeated measurements in longitudinal data. However, in addition to the penalty, knot selection was still performed. The authors describe that the purpose of the penalty is for the selection of variables. This is performed by adding a regularization penalty to the minimization of a weighted sum of square residuals (as in the polynomial spline estimation) so that non-relevant variables assumed to have zero coefficient functions are estimated as identically zero. Wang *et al.* (2008) used function space notation to write the estimation procedure by first letting $\mathcal{G}_k$ denote all functions that have the form $\sum_{l=1}^{K_l} \gamma_{lk} B_{lk}(t)$ where $B_{lk}(t)$ are again basis functions. Then $g_k(t)$ was defined as $g_k(t) = \sum_{l=1}^{K_l} \gamma_{lk} B_{lk}(t) \in \mathcal{G}_k$ and $\| g_k \|$ the L_2-norm of the function g_k. Also $p_\lambda(u), u \geq 0$, was defined as the penalty function with penalty parameter λ; this function could be defined in many ways, and in the application of the method Wang *et al.* (2008) used a quadratic spline function known as the SCAD (smoothly clipped absolute deviation) for the penalty function. This leads to the penalized weighted sum of square residuals to be rewritten as

$$\frac{1}{n} \sum_{i=1}^{n} w_i \sum_{j=1}^{J_i} \left\{ y_i(t_{ij}) - \sum_{k=1}^{p} g_k(t_{ij}) x_i^{(k)}(t_{ij}) \right\}^2 + \sum_{k=1}^{p} p_\lambda(\| g_k \|),$$

where n are the number of subjects and J_i is the number of observations for the i^{th} subject. Here the smoothness of the coefficient functions is controlled by the K_l and the λ parameter decides the variable selection. The interior knots can be equally spaced or placed on the sample quantiles of the data. The knots were selected by an approximated cross-validation criterion proposed by the authors. The above is then solved by an iterative algorithm to find the estimates.

2.3.2 Estimation using local regression:

Local regression can be seen as an effective alternative to smoothing splines. It involves calculations around a neighbourhood so only part of the data is used and therefore it is considered less computationally intensive then spline methods (Fan and Zhang, 2000). Fan and Zhang (1999) used this method in a two-step estimation procedure to resolve the issue of having different degrees of smoothness of the different coefficient functions, and Hoover *et al.* (1998); Wu *et al.* (2000); Wu and Chiang (2000) applied the procedure to longitudinal data.

Instead of using splines for estimation, we have kernels which are described using Wasserman (2006). A Kernel is defined as any smooth function

K such that $K(x) \geq 0$ and satisfies the following conditions:

$$\int K(x)\ dx = 1,$$
$$\int xK(x)\ dx = 0, \text{ and}$$
$$\int x^2 K(x)\ dx > 0.$$

An example of a kernel function is the Gaussian kernel

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp^{-x^2/2},$$

or the commonly used Epanechnikov kernel

$$K(x) = \frac{3}{4}(1 - x^2) I(x)$$

where

$$I(x) = \begin{cases} 1, & \text{if } |x| \leq 1, \\ 0, & \text{if } |x| > 1. \end{cases}$$

For using kernels in local nonparametric regression, a weighted average of the y_is are taken to give higher weights to points near x. In other words we define the kernel estimator as

$$\hat{r}_n(x) = \sum_{i=1}^{n} \frac{K\left(\frac{x - x_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x - x_j}{h}\right)} y_i$$

for a positive valued bandwidth h. The choice of kernel is not as important as the choice of the bandwidth which controls the amount of smoothing. A small bandwidth can give rough estimates while a larger one gives smoother estimates, and therefore the bandwidth is usually chosen by cross-validation (Wasserman, 2006).

The local regression estimation procedures used in the literature for varying coefficient models involves the use of kernel-local polynomials. This is because kernel estimators can suffer from boundary bias which can be reduced by using local polynomial regression (Wasserman, 2006). As described by Wasserman (2006), to find the estimate $\hat{r}_n(x)$ we take a smooth regression function $r(u)$ in the neighbourhood of the target value x and approximate it with a polynomial $P_x(u;a)$. This polynomial can be defined using a Taylor series expansion, i.e.

$$P_x(u;a) = a_0 + a_1(u-x) + \ldots + a_p\frac{(u-x)^p}{p!}.$$

The estimates $\hat{a} = (\hat{a}_0, \ldots, \hat{a}_p)^T$ are found by minimizing the locally weighted sum of squares

$$\sum_{i=1}^{n} w_i(x)(y_i - P_x(x_i;a))^2,$$

where $w_i(x) = K((x_i - x)/h)$. Then the local estimate of r is $\hat{r}_n(u) = P_x(u;\hat{a})$ and for $u = x$ we have $\hat{r}_n(x) = P_x(x;\hat{a}) = \hat{a}_0(x)$.

One step local regression estimation

For a model $Y = \sum_{j=1}^{p} a_j(U)X_j + \varepsilon$ with a sample $\{(u_i, x_{i1}, \ldots, x_{ip}, y_i)\}_{i=1}^{n}$, one can approximate the coefficient functions $a_j(\cdot)(j = 1, \ldots, p)$ for each given point u_0 by a truncated Taylor series expansion by $a_j(u) \approx a_j + b_j(u - u_0)$ for u in a neighbourhood of u_0 (Fan and Zhang, 1999). Then the least-squares problem is to minimize

$$\sum_{i=1}^{n}\left[y_i - \sum_{j=1}^{p}\{a_j + b_j(u_i - u_0)\}x_{ij}\right]^2 K_h(u_i - u_0)$$

where $K_h(\cdot) = K(\cdot/h)/h$ is a kernel function usually taken to be the Epanechnikove kernel, with bandwidth h (Cleveland *et al.*, 1992; Fan and Zhang, 2008, 1999). Bandwidth selection can be conducted by cross-validation method (Hoover *et al.*, 1998). Then the linear estimator $\hat{\mathbf{a}}(u)$ of $\mathbf{a}(u) = \sum_{j=1}^{p} a_j(u)$ in matrix form is

$$\hat{\mathbf{a}}(u) = (\mathbf{I}_p, \mathbf{0}_p)(\boldsymbol{\Gamma}_u^T\mathbf{W}_u\boldsymbol{\Gamma}_u)^{-1}\boldsymbol{\Gamma}_u^T\mathbf{W}_u\mathbf{y},$$

where

$$\begin{aligned} \mathbf{X} &= (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T, & \mathbf{U}_u &= \mathrm{diag}(u_1 - u, \ldots, u_n - u) \\ \boldsymbol{\Gamma}_u &= (\mathbf{X}, \mathbf{U}_u\mathbf{X}), & \mathbf{y} &= (y_1, \ldots, y_n)^T, \\ \mathbf{W}_u &= \mathrm{diag}\left(K_h(u_1 - u), \ldots, K_h(u_n - u)\right), \end{aligned}$$

with $(u_1, \ldots, u_n)$ in a neighbourhood of u, $\mathbf{I}_p$ is a p size identity matrix, and $\mathbf{0}_p$ is a p size matrix of 0 entries. The estimator $\hat{\mathbf{a}}(u)$ is asymptotically normally distributed (Fan and Zhang, 2008).

Two-step local regression estimation

A shortcoming of the above local regression method is that it involves only one smoothing parameter, and this can cause an undersmoothing of some coefficient functions when they have different degrees of smoothness (Hoover *et al.*, 1998). With the one-step method, one bandwidth is used for estimation, however a larger bandwidth would be required for smoother components, and a smaller bandwidth is needed for rougher components. A proposed solution for this issue is used by Fan and Zhang (2000, 1999) which suggest using a two-step estimating procedure. To describe this procedure assume that we have the model

$$y_i = \sum_{j=1}^{p-1} a_j(u_i)x_{ij} + a_p(u_i)x_{ip} + \epsilon_i, \qquad i = 1, \ldots, n.$$

Here we assume that $a_p(\cdot)$ is smoother than any $a_j(\cdot), j = 1, \ldots, p-1$ which have the same smoothness (Fan and Zhang, 2008). First the local regression estimation procedure described above is used with a small bandwidth to obtain the initial estimator of $\mathbf{a}(u)$ where $\mathbf{a}(u) = (a_1(\cdot), \ldots, a_p(\cdot))^T$. This gives an estimate with a larger variance but a smaller bias. Then the estimator $\tilde{a}_j(u_i)$ replaces $a_j(u_i)$ for $j = 1, \ldots, p-1$, which gives

$$y_i - \sum_{j=1}^{p-1} \tilde{a}_j(u_i)x_{ij} = a_p(u_i)x_{ip} + \epsilon_i, \qquad i = 1, \ldots, n.$$

Since $a_p(\cdot)$ is the smoother component, assumed to have a fourth derivative, by Taylor expansion it can be represented by $a_p(u_i) \approx \sum_{k=0}^{3}(k!)^{-1}a_p^{(k)}(u)(u_i - u)^k$ with u_i in a neighbourhood of u with length $2h_1$. Then by using a larger bandwidth this leads to the minimization of

$$\sum_{i=1}^{n} \left\{ y_i - \sum_{j=1}^{p-1} \tilde{a}_j(u_i)x_{ij} - x_{ip}\sum_{k=0}^{3} a_{p,k}(u_i - u)^k \right\}^2 \times K_{h_1}(u_i - u)$$

with respect to $(a_{p,0}, a_{p,1}, a_{p,2}, a_{p,3})$ to find the estimator of $a_p(u)$ which corresponds to $a_{p,0}$. This final estimator is

$$\hat{a}_p(u) = e_{1,4}^T (G^T W_1 G)^{-1} G^T W_1 \tilde{\mathbf{y}}$$

where

$$\tilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_n)^T, \qquad \tilde{y}_i = y_i - \sum_{j=1}^{p-1} \tilde{a}_j(u_i) x_{ij}$$

$$\mathbf{W}_1 = \operatorname{diag}\left(K_{h_1}(u_1 - u), \ldots, K_{h_1}(u_n - u)\right), \quad \mathbf{G} = \operatorname{diag}(x_{1p}, \ldots, x_{np})\mathbf{Q},$$

and

$$Q = \begin{pmatrix} 1 & u_1 - u & (u_1 - u)^2 & (u_1 - u)^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & u_n - u & (u_n - u)^2 & (u_n - u)^3 \end{pmatrix}$$

The two-step estimation procedure was shown to outperform the one-step estimation procedure when the residual sum of squares were compared; this was true even in the case where all coefficients had the same level of smoothness. However, even when the two-step estimation procedure is used this gives only two levels of smoothness compared to spline methods which provide a separate level of smoothness for each coefficient function. In addition, Huang *et al.* (2002) criticized the approach of Fan and Zhang (2000) by stating that the two-step estimation binned data from adjacent time points but the methods of bin selection were not studied. Huang *et al.* (2002) demonstrated that his proposed approach of the one-step polynomial spline estimation through basis expansions, as described above, allows different amounts of smoothing for different individual coefficient curves and requires no binning of data when the observations are sparse at distinct observation times (Huang *et al.*, 2002).

Semi-varying coefficient models in local regression estimation

In practical problems it is very plausible to have some coefficients which are varying and some which are not. This translates to a semi-varying coefficient model with the form

$$Y = Z_1\mathbf{a}_1(U) + Z_2\mathbf{a}_2 + \varepsilon,$$

where $(Z_1, Z_2)^T = X$ with Z_i a p_i dimensional covariate, $i = 1, 2$, and $p_1 + p_2 = p$ (Fan and Zhang, 2008; Zhang *et al.*, 2002). The model has a non-parametric component with the coefficient functions $\mathbf{a}_1(U)$, and a linear component with the constant functions $\mathbf{a}_2$ (Zhang *et al.*, 2002).

Zhang *et al.* (2002) proposed a two-step procedure for this model which involves first estimating the coefficients in the linear component of the model using a small bandwidth and the same procedure of the local polynomial estimation described previously. To estimate $\mathbf{a}_2$ first, it is treated as a functional parameter and then using the local polynomial estimation the initial estimator of $\mathbf{a}_2(u_i)$ was found to be

$$\tilde{\mathbf{a}}_2(u_i) = (\mathbf{0}_{p_2\times p_1}, \mathbf{I}_{p_2}, \mathbf{0}_{p_2\times p}) \left(\mathbf{\Gamma}_{u_i}^T \mathbf{W}_{u_i} \mathbf{\Gamma}_{u_i}\right)^{-1} \mathbf{\Gamma}_{u_i}^T \mathbf{W}_{u_i} \mathbf{y},$$

with the terms having the same definition as described in the one-step local regression estimation. Then to obtain $\hat{\mathbf{a}}_2$ the initial estimator is averaged over $i = 1, \ldots, n$ as follows

$$\hat{\mathbf{a}}_2 = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{0}_{p_2\times p_1}, \mathbf{I}_{p_2}, \mathbf{0}_{p_2\times p}) \left(\mathbf{\Gamma}_{u_i}^T \mathbf{W}_{u_i} \mathbf{\Gamma}_{u_i}\right)^{-1} \mathbf{\Gamma}_{u_i}^T \mathbf{W}_{u_i} \mathbf{y}.$$

This final estimate of $\hat{\mathbf{a}}_2$ is then substituted in the original model which then transforms it to a standard varying coefficient model. Zhang *et al.* (2002) then used Fan and Zhang (2000, 1999) two-step estimation method to estimate the coefficient functions of the non-parametric part of the model.

Xia *et al.* (2004) also studied the estimation of semi-varying models, however they proposed using a semi-local least squares estimation by estimating $\mathbf{a}_1(U)$ locally and $\mathbf{a}_2$ globally. Their procedure is similar to that of Zhang *et al.* (2002), in that the estimate for $\mathbf{a}_2$ is found first and then substituted this into the model to obtain a standard varying coefficient model which is then estimated using the procedure of Fan and Zhang (2000, 1999). Xia *et al.* (2004) also described a model selection process using cross-validation. Model selection is crucial in semi-varying coefficient models since we need to test which coefficients are varying and which are not. Other methods of hypothesis testing for model selection is discussed in more detail further on.

A different approach for estimating semi-varying coefficient models was through the use of profile least-square estimation by Fan and Huang (2005). This begins with assumption that $\mathbf{a}_2$ is known, which allows the semi-varying model to be rewritten as

$$y_i - z_{i2}\mathbf{a}_2 = z_{i1}\mathbf{a}_1(Uu_i) + \epsilon_i, \qquad i = 1, \ldots, n,$$

where $(z_{i1}, z_{i2})^T = \mathbf{x}_i$ Fan and Zhang (2008); Fan and Huang (2005). The local regression estimation can then be applied to find the estimator of $\mathbf{a}_1(u_i)$ which is

$$\tilde{\mathbf{a}}_1(u_i) = (\mathbf{I}_{p_1}, \mathbf{0}_{p_1}) \left(\tilde{\mathbf{\Gamma}}_{u_i}^T \mathbf{W}_{u_i} \tilde{\mathbf{\Gamma}}_{u_i}\right)^{-1} \tilde{\mathbf{\Gamma}}_{u_i}^T \mathbf{W}_{u_i} \tilde{\mathbf{y}},$$

where $\tilde{\mathbf{\Gamma}}_u$ is the same as the previously defined $\mathbf{\Gamma}_u$ in the local regression estimation above but with the $\mathbf{X}$ replaced by $\mathbf{Z}_1$, i.e.

$$\begin{aligned}
&\tilde{\mathbf{\Gamma}}_u = (\mathbf{Z}_1, \mathbf{U}_u\mathbf{Z}_1), \quad \text{also we have} \\
&\mathbf{Z}_1 = (z_{11}, \ldots, z_{n1})^T, \quad \mathbf{Z}_2 = (z_{12}, \ldots, z_{n2})^T, \\
&\ \tilde{\mathbf{y}} = \mathbf{y} - \mathbf{Z}_2\mathbf{a}_2.
\end{aligned}$$

After substituting $\tilde{\mathbf{a}}_1(u_i)$ for $\mathbf{a}_1(u_i)$ in the model above, a least squares estimation can be used to find $\hat{\mathbf{a}}_2$.

Local regression estimation for generalized VCMs

The local regression method equivalent for generalized varying coefficient models involves the maximization of

$$l_n(\mathbf{a}, \mathbf{b}) = \frac{1}{n} \sum_{i=1}^{n} l \left[g^{-1} \left\{ \sum_{j=1}^{p} (a_j + b_j(u_i - u_0)) x_{ij} \right\}, y_i \right] \times K_h(u_i - u_0),$$

where $K_h(\cdot)$ is a kernel function with a bandwidth h, and $a_j(u) \approx a_j + b_j(u - u_0)$ as was described in the standard case. Cai *et al.* (2000) explains that in order to find the estimates one needs to maximize the local likelihood above for perhaps hundreds of distinct values of u_0 each requiring an iterative algorithm. To save the computational cost a one-step Newton-Raphson estimator was proposed. If we let $\boldsymbol{\beta} = \boldsymbol{\beta}(u_0) = (a_1, \ldots, a_p, b_1, \ldots, b_p)^T$ and $l_n'(\boldsymbol{\beta})$ and $l_n''(\boldsymbol{\beta})$ be the gradient and Hessian matrix of the local log-likelihood $l_n(\boldsymbol{\beta})$, then the one-step of the Newton-Rapshon algorithm gives the updated estimator

$$\hat{\boldsymbol{\beta}}_{OS} = \hat{\boldsymbol{\beta}}_0 - \{l''_n(\hat{\boldsymbol{\beta}}_0)\}^{-1} l'_n(\hat{\boldsymbol{\beta}}_0),$$

where $\hat{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\beta}}_0(u_0) = (\hat{\mathbf{a}}(u_0)^T, \hat{\mathbf{b}}(u_0)^T)$ is the initial estimator. For example in a logistic regression, the one-step estimator would be given by

$$\hat{\boldsymbol{\beta}}_{OS} = \hat{\boldsymbol{\beta}}_0 + \begin{pmatrix} \mathbf{H}_{n,0}, & \mathbf{H}_{n,1} \\ \mathbf{H}_{n,1} & \mathbf{H}_{n,2} \end{pmatrix}^{-1} + \begin{pmatrix} \mathbf{v}_{n,0} \\ \mathbf{v}_{n,1} \end{pmatrix},$$

where

$$\mathbf{H}_{n,j} = \sum_{i=1}^{n} K_h(u_i - u_0)\hat{p}_{i0}(1 - \hat{p}_{i0})(u_i - u_0)^j \mathbf{x}_i \mathbf{x}_i^T, \qquad j = 0, 1, 2,$$

satisfies

$$\begin{aligned} \text{logit}(\hat{p}_{i0}) &= \sum_{j=1}^{p} \{\hat{a}_{j,0} + \hat{b}_{j,0}(u_i - u_0)\} x_{ij}, \qquad \text{and} \\ \mathbf{v}_{n,j} &= \sum_{i=1}^{n} K_h(u_i - u_0)(y_i - \hat{p}_{i,0})(u_i - u_0)^j \mathbf{x}_i, \qquad j = 0, 1. \end{aligned}$$

Cheng *et al.* (2009) proposed an approach to find both constant and functional parameters, which is similar to the estimation of the semi-varying coefficient models described earlier in the standard estimation. Li and Liang (2008) also described an estimation procedure for generalized case of semi-varying coefficient model using a penalized quasi likelihoods for the main purpose of variable selection, which will be discussed in the next section. Both Cheng *et al.* (2009) and Li and Liang (2008) used local likelihood estimation as was shown in the semi-varying coefficient model estimation, however there were some difference in the approaches used. Cheng *et al.* (2009) used a similar method to Zhang *et al.* (2002) which was to estimate the constant vector parameter $\boldsymbol{\theta}$ first by treating it as an unknown function of U, and then plug this estimate to the local likelihood function to estimate the functions $\mathbf{a}_j(\cdot)$. The details of this estimation begins with the conditional log-likelihood function

$$L_0(\boldsymbol{\theta}, \mathbf{a}) = \sum_{i=1}^{n} \log\ f\left(y_i, \mathbf{x}_i, \boldsymbol{\theta}, \mathbf{x}_{i,1}^T \mathbf{a}_1(u_i), \ldots, \mathbf{x}_{i,l}^T \mathbf{a}_l(u_i)\right),$$

where f is a known parametric density function, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^T$ is the unknown constant vector and $\mathbf{a}_j(\cdot) = (a_{j1}(\cdot), \ldots, a_{jp_j}(\cdot))^T$ is the unknown function. As was done in the standard case, a truncated Taylor's expansion was used for the approximation $\mathbf{a}_j(u_i) \approx \mathbf{a}_j(u) + \mathbf{b}_j(u)(u_i - u)$, with u_i in a a neighbourhood of u. Then the local log-likelihood function was written as

$$\sum_{i=1}^{n} K_h(u_i - u) \log f\left(y_i, \mathbf{x}_i, \boldsymbol{\theta}, \mathbf{x}_{i,1}^T\{\mathbf{a}_1 + \mathbf{b}_1(u_i - u)\}, \ldots, \mathbf{x}_{i,l}^T\{\mathbf{a}_l + \mathbf{b}_l(u_i - u)\right),$$

where $K_h(\cdot)$ is a kernel function with bandwidth h. Maximizing the above gives $\left(\tilde{\boldsymbol{\theta}}(u)^T,\ \tilde{\mathbf{a}}_1(u)^T, \tilde{\mathbf{b}}_1(u)^T, \ldots, \tilde{\mathbf{a}}_l(u)^T, \tilde{\mathbf{b}}_l(u)^T\right)^T$, then by averaging the $\tilde{\theta}(u_i)$ over $i = 1, \ldots, n$ the final estimator is found, i.e.

$$\hat{\boldsymbol{\theta}} = \frac{1}{n} \sum_{i=1}^{n} \tilde{\boldsymbol{\theta}}(u_i).$$

To obtain the functional parameters, $\hat{\boldsymbol{\theta}}$ was replaced for $\boldsymbol{\theta}$ in the local log-likelihood and maximizing this gave the estimators

$$\left(\hat{\mathbf{a}}_1(u)^T, \hat{\mathbf{b}}_1(u)^T, \ldots, \hat{\mathbf{a}}_l(u)^T, \hat{\mathbf{b}}_l(u)^T\right)^T.$$

Alternately, Li and Liang (2008) did not average the constant parameter as shown above but used a penalized likelihood to find the final estimator for the constant parameter. This procedure begins by first maximizing the local likelihood function

$$\sum_{i=1}^{n} Q\left[g^{-1}\{\mathbf{a}^T \mathbf{x}_i + \mathbf{b}^T \mathbf{x}_i(u_i - u) + \mathbf{Z}_i \boldsymbol{\theta}\}, y_i\right] K_h(u_i - u),$$

where again the functional parameters are approximated by a truncated Taylor's expansion, with $\mathbf{a} = (a_1, \ldots, a_p)^T$ and $\mathbf{b} = (b_1, \ldots, b_p)^T$. Here the constant parameter $\boldsymbol{\theta}$ has the covariates $\mathbf{Z}$. The solution to the above maximization gives $\tilde{\mathbf{a}}, \tilde{\mathbf{b}}, \tilde{\boldsymbol{\theta}}$ which is then replaced to obtain the penalized likelihood

$$\sum_{i=1}^{n} Q\left\{g^{-1}\left(\tilde{\mathbf{a}}^T\mathbf{x}_i + \tilde{\mathbf{b}}^T\mathbf{x}_i(u_i - u) + \mathbf{Z}_i\boldsymbol{\theta}\right), y_i\right\} - n\sum_{j=1}^{p} p_{\lambda_j}(|\theta_j|),$$

with $p_{\lambda_j}(\cdot)$ a pre-specified penalty function (such as the SCAD penalty) with regularization parameter λ_j. This penalized likelihood is maximized to obtain $\hat{\boldsymbol{\theta}}$, which also performs a variable selection for the constant parameters. Finally to obtain the functional parameters, $\hat{\boldsymbol{\theta}}$ is substituted for $\boldsymbol{\theta}$ to obtain the following local likelihood function,

$$\sum_{i=1}^{n} Q\left[g^{-1}\{\mathbf{a}^T + \mathbf{b}^T\mathbf{x}_i(u_i - u) + \mathbf{Z}_i\hat{\boldsymbol{\theta}}\}, y_i\right] K_h(u_i - u),$$

which can now be maximized to find the final estimators of $\{\hat{\mathbf{a}}, \hat{\mathbf{b}}\}$.

2.3.3 Bayesian approach to VCMs

Hastie and Tibshirani (1993) briefly described a Bayesian approach for estimating the varying coefficient model. For a simple standard varying coefficient model $Y = Xa(u) + \varepsilon$ with normally distributed ε, a prior is placed on $a(u)$. West and Harrison (1997) described this approach in more detail by defining

$$\begin{aligned} Y_t &= X_t a_t + \nu_t, & \nu_t &\sim N(0, V_t), \\ a_t &= G_t a_{t-1} + t\omega_t, & \omega_t &\sim N(0, W_t) \end{aligned}$$

where t refers to time in this case. The first equation is the observation equation and the second is the evolution equation, the regression parameter here is a function of time defined by a Markov process as shown in the evolution equation. Having a Markov formulation in the second version makes it convenient to make inference sequentially, by an updating formula based on the Kalman filter. However, the Markov assumption and normality assumptions might not always hold, and other Bayesian approaches may be required (Hastie and Tibshirani, 1993).

Fahrmeir and Lang (2001) describes the Bayesian approach for generalized varying coefficient models with applications to temporal and spatial effects. Beginning with the observations $(y_i, x_{i1}, \ldots, x_{ip}, w_i)$ for $i =$

$1, \ldots, n$, lets assume y_i belongs to an exponential family with mean $\mu_i = \mathrm{E}(y_i|x_i, w_i) = h(\eta_i)$, where h is a known link function and x_i and w_i are covariates, the model described is written as

$$\eta_i = f_1(x_{i1})z_{i1} + \ldots + f_p(x_{ip})z_{ip} + \mathbf{w}_i'\boldsymbol{\beta} + b_{gi},$$

where $f_1, \ldots, f_p$ are unknown smooth functions of the covariates. The b_{gi} terms are unit or group specific random effects for any $g = 1, \ldots, G$, and $\mathbf{z} = (z_1, \ldots, z_p)$ is a design vector that could contain components of $\mathbf{x}$ or $\mathbf{w}$. The covariates $x_1, \ldots, x_p$ described are either metrically or spatially correlated and have a non-linear effect on the outcome, and $\mathbf{w}$ is a vector of further covariates who are assumed to have a linear effect. In the Bayesian framework, the functions $f_1, \ldots, f_p$, the parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_r)^T$ and the random effects $\mathbf{b} = (b(1), \ldots, b(G))^T$ are all random variables that require appropriate priors. For timescales and metrical covariates, priors are based on Gaussian smoothness priors while for spatial covariates Gaussian Markov random fields priors are used. For instance, if time was the covariate for x which are equally spaced observations, then the common priors for the smooth functions are first or second order random walk models with normally distributed errors. For inference, Gibbs sampling is used for Gaussian responses and Metropolis-Hastings algorithms are used for non-Gaussian responses.

The use of a Bayesian framework for estimation of varying coefficient models (or GAMs which can be easily extended to varying coefficient models) was described by Biller and Fahrmeir (2001); Brezger and Lang (2006); Fahrmeir and Lang (2001) for studying temporal trends, by Assunçao (2003); Congdon (2003); Gamerman *et al.* (2003); Fahrmeir *et al.* (2004) for studying spatial trends, and for spatial-temporal trends as well by Brezger and Lang (2006); Fahrmeir *et al.* (2004, 2000). For the spatial-temporal models, the effects of time and space are modelling additively, although interactions between space and time were discussed by Fahrmeir *et al.* (2000) by transforming time into a categorical variable of two periods.

2.3.4 Spatial VCMs using tensor product smooths

Bayesian methods are usually used for the analysis of spatially varying coefficients, however their application may not always be feasible and especially for large datasets. An alternative, that is more consistent with the estimation methods discussed previously, is to use nonparametric methods using local regression or splines as discussed previously. Local regression techniques were used for estimation of spatially varying coefficient models by Brunsdon *et al.* (1996); Muller (2007); Wheeler and Páez (2010) where they are referred to as geographically weighted regression techniques. Young *et al.* (2008) applied this method using U.S. BRFSS data and other sources of data to study the association between myocardial infarctions and ambient ozone levels to demonstrate variation with space. An alternative estimation method is described by Wood (2006a), which shows how spatially varying coefficient models can also be estimated using spline estimation methods, as this can also provide a flexible model for studying spatial variations. This method was used by Augustin *et al.* (2009) for studying forest health and Augustin *et al.* (2013) for studying fishery management, both of which used a tensor products of smooths of spatial coordinates as well as time to study spatial-temporal trends. Wood (2006b) and Wood *et al.* (2013) also applied a spatial-temporal varying coefficient model and showed how the method can be used in mixed models, and Heim *et al.* (2007) and Eilers *et al.* (2005) showed how this method could be applied to 3-dimensional cases. The spatial-temporal models using this method are not modelled additively, but are taken as a tensor product smooth of space and time, and are seen as a useful method as the tensor products are invariant to the units in which the covariates are measured.

Tensor products smooths as discussed by Wood (2006a) are used to build smooths of several variables, for instance when we have the variables u_x and u_y as the latitude and longitude coordinates in spatial data. Although, this method may not capture boundary effects, it is still a very flexible model and it is a compromise for the reduced computation time compared to using MCMC methods to estimate the spatial coefficients. To demonstrate how these products are formed, we begin with the assumed low rank bases for representing the smooth functions s_x and s_y for each variable. Then, similar to what was shown previously to describe splines, we have

$$s_{u_x}(u_x) = \sum_{j=1}^{J} \gamma_j B_j(u_x) \quad \text{and} \quad s_{u_y}(u_y) = \sum_{l=1}^{L} \delta_l C_l(u_y),$$

where γ_j and δ_l are parameters with $B_j(u_x)$ and $C_l(u_y)$ the known basis functions. In order to write the tensor product, we need to convert the smooth function $s_{u_x}(u_x)$ to a smooth function of u_x and u_y. This can be achieved by allowing the parameters γ_j to vary smoothly with u_y, i.e. $\gamma_j(u_y) = \sum_{l=l}^{L} \delta_{jl} C_l(u_y)$, which then gives us

$$s_{u_x u_y}(u_x, u_y) = \sum_{j=1}^{J} \sum_{l=1}^{L} \delta_{jl} C_l(u_y) \gamma_j(u_x).$$

This procedure can be followed to construct tensor products for any number of variables required for the analysis (i.e. for 3-dimensional cases or more).

The same concept can be followed to construct the tensor product penalties. To find the penalty of the smooth $s_{u_x u_y}(u_x, u_y)$ for instance, we can write

$$J(s_{u_x u_y}) = \lambda_{u_x} \int_{u_y} J_{u_x}(s_{u_x|u_y}) du_y + \lambda_{u_y} \int_{u_x} J_{u_y}(s_{u_y|u_x}) du_x.$$

Where λ_{u_x} and λ_{u_x} are the smoothing parameters that control the tradeoff between the wiggliness in different directions, and allows the penalty to be invariant to the relative scaling of the variables. For example, if a cubic spline is used, then we have the penalty

$$J(s) = \int_{u_x, u_y} \lambda_{u_x} \left(\frac{\partial^2 s}{\partial u_x^2} \right)^2 + \lambda_{u_y} \left(\frac{\partial^2 s}{\partial u_y^2} \right)^2 du_x du_y.$$

A penalty of the type defined by Eilers and Marx (2002) for P-splines can also be used.

With these components defined, the tensor product smooths can be estimated using the methods described for estimation with splines.

2.4 Hypothesis Testing

Hypothesis testing in varying coefficient models attempts to find whether coefficients are actually varying or not, and also if certain covariates are statistically significant. This translates to testing two different null hypotheses

$$H_0 : a_k(\cdot) = a_k, \qquad k = 1, \ldots, p, \qquad \text{and}$$
$$H_0 : a_k(\cdot) = 0, \qquad \text{for certain k}$$

(Fan and Zhang, 2008). The first null hypothesis involves testing the parametric null hypothesis against the non-parametric alternative hypothesis. The second null hypothesis involves testing a non-parametric null hypothesis against a non-parametric alternative hypothesis, since it contains unknown non-parametric components $a_j(\cdot)$ for $j \neq k$ (Fan and Zhang, 2008).

Cai *et al.* (1999, 2000) discussed the use of a non-parametric generalized maximum likelihood ratio test statistic for testing the hypotheses above. The test statistic is $T = 2\{\ell(H_1) - \ell(H_0)\}$ where $\ell(H_0)$ and $\ell(H_1)$ are the log-likelihood functions under the null and alternative hypotheses or more specifically

$$T = \sum_{i=1}^{n} \left(\ell[g^{-1}\{\mathbf{x}_i^T \hat{\mathbf{a}}(u_i)\}, y_i] - \ell\{g^{-1}(\mathbf{x}_i^T \hat{\mathbf{a}}, y_i)\} \right),$$

where $\hat{\mathbf{a}}(\cdot)$ is the local maximum likelihood estimator of the functional coefficient $\mathbf{a}(\cdot)$ under the alternative hypothesis, and $\hat{\mathbf{a}}$ is the maximum likelihood estimator of the constant vector $\mathbf{a} = (a_1, \ldots, a_p)^T$ under the null hypothesis (Fan and Zhang, 2008). Then the null hypothesis is rejected when $T > c_\alpha$ for a critical value c_α computed by either the asymptotic distribution of T or bootstrap under the null hypothesis (Fan and Zhang, 2008).

For constructing a test statistic, Cai *et al.* (2000) explains that while in parametric models the likelihood ratio test is asymptotically chi-squared, for nonparametric models the number of parameters tends to infinity and so the test statistic would be asymptotically normal and also independent from the value of $\mathbf{a}$. Therefore, a conditional bootstrap can be used to construct the null distribution of the test statistic T. Cai *et al.* (2000) describes the procedure as first obtaining the estimates under the null hypothesis (say $\hat{a}_1, \ldots, \hat{a}_p$) and then generating a bootstrap sample for the Y^* from the generalized linear model $\hat{\eta}(u_i, \mathbf{x}_i) = \sum_{j=1}^{p} \hat{a}_j x_{ij}$ to compute T^*. Then the distribution of T^* is used as an approximation to the distribution of T, which is valid since the asymptotic null distribution does not depend on the values of $\{a_j\}$. The same procedure applies when it is required to test $a_p(\cdot) = 0$, however here the data should be generated from the mean function $g\{m(\mathbf{u}, \mathbf{u})\} = \sum_{j=1}^{p-1} \hat{a}_j(\mathbf{u})x_j$, where $\hat{a}_j(\cdot)$ is an estimate under the null hypothesis. The conditional bootstrap method described above also applies readily to the Poisson and Bernoulli distributions since there is no dispersion parameter involved. If there is a dispersion parameter involved, then this has to be estimated as well.

Li and Liang (2008) described a generalized quasi-likelihood ratio test for selection of significant variables of the functional parameters. As described previously the constant parameters were selected using a penalized quasi likelihood estimation procedure (Li and Liang, 2008). This leads to testing the two hypotheses

$$\ell(H_1) = \sum_{i=1}^{n} Q\left\{g^{-1}\left(\hat{\boldsymbol{\alpha}}^T(u_i)\mathbf{x}_i^T + \mathbf{z}_i^T\hat{\boldsymbol{\theta}}\right), y_i\right\},$$

and

$$\ell(H_0) = \sum_{i=1}^{n} Q\left\{g^{-1}\left(\mathbf{z}_i^T\bar{\boldsymbol{\theta}}\right), y_i\right\},$$

for the null hypothesis (which is equivalent to the second null hypothesis shown above). This gives the generalized quasi-likelihood ratio test statistic

$$T_{GLR} = r_K\{\ell(H_1) - \ell(H_0)\},$$

where $r_K = \left\{\int K(0) - 0.5\int K^2(u)du\right\}\left\{\int\{K(u) - 0.5K * K(u)\}du\right\}^{-1}$. As was done by Cai *et al.* (2000), the null distribution of T_{GLR} is estimated by Monte Carlo simulation or the bootstrap procedure.

Kauermann and Tutz (1999) used a different approach and proposed a graphical technique for testing between a parametric model and a varying coefficient model. They express the varying coefficient model as

$$\mathrm{E}(Y|X, U) = h\{\mathbf{Z}(X)a(U)\}$$

where $h(\cdot)$ is the inverse link function, $\mathbf{Z}(X)$ is a design matrix of the covariates X, and $a(U)$ as an unknown and smooth function in U. A parametric form of this model was expressed by replacing the varying coefficient $a(U)$ with the parametric function $\mathbf{V}(U)\beta$ where $\mathbf{V}(U)$ is a matrix built from U. This leads to the ability to check between the parametric and varying coefficient models by studying the local discrepancy

$$\rho(U) := \gamma(U) - \mathbf{V}(U)\beta;$$

if $\rho(u) \equiv 0$ then we have a parametric model. Kauermann and Tutz (1999) showed how $\rho(u)$ can be studied by graphical methods for testing between the parametric and varying-coefficient models.

Chapter 3

Methodology

3.1 Constructing the Varying Coefficient Model

Estimation methods using splines was used for estimation of varying coefficient models (VCM) in the analysis. This includes polynomial spline estimation, smoothing spline estimation, and P-spline estimation methods. Local polynomial regression was not used as this provides only one or two levels of smoothing for each coefficient function as discussed in the literature review. Spline methods were preferred as they provide a separate level of smoothing for each coefficient function. The results as well the the computational feasibility and practicality of these spline methods for use in surveillance data analysis were compared. This chapter describes the methodology in which the models were constructed and applied to the Italian PASSI data for a smoking status binary outcome variable, and the U.S. BRFSS data for a obesity status binary outcome variable. These datasets and the variables used in the analysis are described in more detail in the results chapters.

The methodology for constructing the varying coefficient model using the different spline estimation methods are very similar. For all these methods, the modifying variable is time, which for the Italian PASSI surveillance data for instance are the months of observation from 2008 to 2011. This gives 44 months in which a new random sample was extracted at each month, as July and August were combined in the data collection phase. Therefore, the maximum number of knots that can be used for estimation is 44 knots for this dataset. In polynomial spline estimation, there is an additional step to the methodology which requires the selection of the number of knots K_l for each variable. However, in smoothing spline, penalized spline regression and P-spline estimation, this step is not required as a high number of knots (or all the knots) are placed with an added penalty to control the smoothing.

In the first step of constructing the varying coefficient model, a varying coefficient model is fit in which each variable alone is allowed to have varying coefficients while all other variables have constant coefficients. The second step then involves testing these models against a parametric model (or a model in which all the coefficients are constant) to see if the coefficients are actually varying. The final step then combines all the significantly varying coefficients in a step wise selection method to find the final varying coefficient model. The final varying coefficient model that needs to be found, could be theoretically written as

$$\log\left(\frac{P(Y=1|\mathbf{Z}=z,\mathbf{X}=x,T=t)}{1-P(Y=1|\mathbf{Z}=z,\mathbf{X}=x,T=t)}\right)=b_0+\sum_{j=1}^{p}b_jZ_j+a_0(t)+\sum_{k=1}^{q}a_k(t)X_k,$$

where Y is the outcome binary variable, Z_j are the independent variables with constant parameters b_j, and X_k are the independent variables with varying coefficients $a_k(t)$. The left expression of the model above will be represented in the remaining of the thesis by logit(Y) for simplicity, where `logit` is the log odds of the binary outcome variable. For the spatial varying coefficient model, `t` in the above model is replaced by `s` to represent space.

3.1.1 Step One: Fitting a VCM for each variable

For the estimation using the polynomial estimation method, the number of knots was selected for each variable using the AIC criterion. Having a separate number of knots for each variable's coefficient functions allow for different degrees of smoothness. A third degree B-spline was used, and therefore the knots to be checked were from 5 to 44 knots; a minimum of five knots is used since three knots are required for the degrees of the spline and two knots for the boundaries of the domain. This procedure was conducting using `R` software with the `mgcv` package (R Core Team, 2012; Wood, 2007) used for generalized additive models. To use the `gam` function for estimating a varying coefficient model the `''by''` option is used as shown in the following example:

```
gam(SMK ~ Zj + income + s(time, bs = "bs", fx=TRUE, k = i,
                  by = income), family = binomial("logit")),
```

where `SMK` is the binary response variable for smoking status for instance, `income` is the independent variable for income status, Z_j are all the other independent variables, `k="bs"` is to select the B-spline, `fx=TRUE` indicates that there is no penalty, and the `by` option allows for defining varying coefficients. B-splines are used due to their good numerical properties, and as this spline is not found in the options of the `gam` function, it added from the `splines` package using the `smooth.construct` function in the `mgcv` package, which allows for the construction of other types of spline functions. In the above example, `‘‘i"` refers to the number of knots to be checked from 5 to 44. Therefore the same model above is fit 40 times, each time selecting a different number of knots from 5 to 44. The model which gives the lowest AIC value is selected to determine the number of knots for that variable, and this model is then used in the test discussed below to indicate if the coefficients are constant or varying. Although knot selection is not required for the penalized spline regression, this was performed so as to compare with the polynomial spline estimation method as the only difference is the addition of the penalty. Therefore, the code used to fit the model is the same as shown in the example above except for the change `fx=FALSE` which is the default setting in the `gam` function.

For the estimation using the smoothing spline or P-spline methods, there is no need to perform the procedure for the selection of knots. Both these estimation methods contain a penalty with a smoothing parameter λ. For the P-spline estimation the R code to fit the model for each variable is as shown in the following example:

```
gam(SMK ~ Zj + income + s(time, bs = "ps", k = 44, m=c(3,2),
                  by = income), family = binomial("logit")).
```

Here `k="ps"` is to select for using P-splines, and `m=c(3,2)` indicates that a third degree B-spline is used with a second order difference penalty. The maximum number knots of 44 were used in the estimation. The same code as above was used but with the `bam` function of the `mgcv` package in order to observe if this function can significantly improve computation time. This function is ideal for large datasets as it can reduce computation time by reducing the model matrix required for finding the estimates (Wood, 2007).

The `bam` function uses a fast restricted maximum likelihood (REML) method to selection the smoothing parameter, whereas models fit with the `gam` function use the default GCV method to select the smoothing parameter (Wood, 2011, 2007).

The final method to be compared is using the smoothing spline estimation. Earlier computations not shown here have indicated that this estimation method has a high computation time due to the type of penalty used, and therefore only the `bam` function was used to save computational time. A cubic regression spline was also used and therefore the following code was used:

```
bam(SMK ~ Zj + income + s(time, bs = "cr", k = 44, by = income),
    family = binomial("logit")),
```

where `bs="cr"` selects a cubic regression spline.

3.1.2 Step Two: Test for varying coefficients

For the second step, the model selected is tested against the parametric model, or a model where all the independent variables have constant coefficients, to see if the variable has coefficients that are actually varying. For instance if we an independent variable X_1, and again a smoking status binary outcome variable, the following alternative hypothesis is tested against the parametric null hypothesis

$$\mathbf{H_0}: \ \text{logit}(SMK) = \sum_{j=1}^{p} b_j Z_j,$$

and

$$\mathbf{H_1}: \ \text{logit}(SMK) = \sum_{j=1}^{p} b_j Z_j + a_1(t) X_1,$$

where Z_j are the variables from the parametric model with constant coefficients b_j, and $a_1(t)$ are the functional coefficients of X_1. This test would show if the varying coefficients $a_1(\cdot)$ are actually varying or should remain constant, and it is performed using the a chi-square test with the function `anova`. This test can be used since there are a limited number of parameters to be estimated (parameters are not increasing with increasing n), and the sample size is large enough to guarantee the asymptotic chi-square distributed of the test statistics. Any variable which gives a significant p-value is then included in the forward selection process of the final step, and any variable which gives a non-significant p-value indicates that it has coefficients that are constant.

3.1.3 Step Three: Constructing the final VCM using forward selection

The final step involves finding the full varying coefficient model where more than one variable with varying coefficients are required in the model. This is conducted using a stepwise method beginning with the varying coefficient model from step one which gave the largest deviance explained. To know which variable should be added next, the residuals of the first model (or each previous model in the step-wise process) is fit with each of the remaining variables and the variable from the model which provides the best explanation for these residuals is added next. For polynomial spline estimation, knot selected is then performed again. However, in this step, the best combination of the number of knots when two or more variables with varying coefficients are included in the model needs to be found. For each combination of knots a varying coefficient model is fit and the model which gives the minimum AIC is selected to determine the number of knots and the model to be examined. The same procedure is conducted using the penalized spline regression method. However, in the P-spline and smoothing spline methods there is no knot selection performed and the maximum number of knots is used again for each variable with varying coefficients.

Once the models are fit, a new test is performed to see if an additional variable with varying coefficients should be added when a previous one is already in the model, i.e. continuing with the example above the test is

$$\mathbf{H_0}: \ \text{logit}(SMK) = \sum_{j=1}^{p} b_j Z_j + a_1(t)X_1,$$

$$\mathbf{H_1}: \ \text{logit}(SMK) = \sum_{j=1}^{p} b_j Z_j + a_1(t)X_1 + a_2(t)X_2,$$

where X_2 is another variable which has coefficients that are actually varying according to the test in step two. The final varying coefficient model is found by testing each additional variable with varying coefficients using the above test until all the coefficient functions for all the independent variables are tested.

3.2 Odds Ratio Plots

From the results, plots of the varying coefficients over time can be constructed to observe and understand the trends of the effects of the variables on the outcome. The plots provided by the `mgcv` package show the changing coefficient estimates with time, and these kinds of plots will be shown for all the time varying coefficients to compare the results of the five methods used. However, to understand these plots better and particularly for using in public health interventions, it is preferable to look at odds ratio plots. These plots are constructed by adding the constant estimate of a certain category to the spline plot of that category. This is done because the variables in the varying coefficient models have a constant coefficient found in b_j as shown in the models above, as well as the time varying coefficients found in $\mathbf{a}(t)$ if this was found to be significant in the tests. Since a logistic model is being used, the odds ratios are then constructed by taking the exponential of the coefficients to produce odds ratio plots which are easier to interpret, the plots are therefore on an exponential scale. These plots can be interpreted as the change with time in the odds ratio of a certain category compared to the reference category on the outcome. For plots produced by the `plot.gam` function of the `mgcv` package, Bayesian confidence intervals for the smooth terms are added to the plot, these confidence intervals can be obtained by simulating from the posterior distribution of the functional coefficients (Wood, 2006a).

3.3 Spatial VCMs

The methods discussed for constructing the time varying coefficient model can also be used to estimate a spatial varying coefficient model by defining spatial coordinates as the modifying variables. The same basic procedure is followed (steps 1 to 3), however here we have a tensor product of smooths of the spatial coordinates as discussed in the literature review. This procedure is described by Wood (2006a) and can also be extended for creating a tensor product of spatial coordinates and time for the study of a spatial-temporal varying coefficient model. Again using the mgcv package, the following code can be used to fit the spatial varying coefficient models

```
bam(SMK ~ Zj + income + te(x,y, k = 10, by = income),
                  family = binomial("logit")).
```

Here te represents tensor product smooths and the longitude and latitude are represented by the variables x and y. The coordinates can be taken to be the centroids of specific geographical locations (for example regions or counties), or population centroids, i.e. coordinates of highest population locations. The code above uses the default cubic regression spline, however other splines can also be used. The number of knots can be defined for each coordinate separately; in the above example placing one value for the number of knots implies using the same number for both coordinates. Therefore, there are 100 knots used in this example as there are 10 knots for each coordinate.

To represent the results, maps are constructed for each category with space varying coefficients from the final model. The maps represent the probabilities of the outcome variable for the indicated category at each centroid while keeping all the other categories constant at the reference level. This is conducted by using the final model to predict the probabilities of the outcome at the centroid coordinates of the regions or counties under study. The predictions are first made for estimating the probabilities when all the categories are kept at the reference level (the reference map). Then to observe the spatial variation for each category of a variable, a new prediction is made by changing the variable in question from the reference category to another category again while keeping all the other categories at the reference level. The resulting maps can show whether the probabilities are changing spatially. The application of this method is discussed in the USA analysis results chapter.

Chapter 4

Results: Italian Analysis

4.1 Italian PASSI Data

The data used for analysis is from the PASSI (Progressi delle Aziende Sanitarie per la Salute or Progress in the Italian Local Health Units) surveillance system in Italy, the details of which are described by Baldissera *et al.* (2011) as well as Minardi *et al.* (2011); Binkin *et al.* (2010) who have used the PASSI data for analysis. The unit used for data collection is the local health unit which are found in all the 21 Italian regions. Each region in Italy has between 1 and 22 local health units and these units are responsible for providing health services for its residents. Data collection began in mid 2007 and is still ongoing and is conducted by each local health unit participating in the surveillance system, which is over 90% of the Italian local health units. A monthly random sample is chosen from a list of residents from each local health unit aged 18-69 years and an interview with those selected is conducted by telephone. The questionnaire used in the telephone interview covers a wide variety of behavioural and preventive topics and the interview lasts for a median of 20 minutes.

4.2 Data Preparation

For the current analysis, the Italian PASSI surveillance data for the years 2008 to 2011 is combined for a total sample size of 148,266 observations. The variables used in the analysis are found in Table A.1. Most of these variables are self explanatory, namely age, sex, marital status, education, region, citizenship, and income level (or economic difficulty). However, the variables smoking, alcohol consumption, physical activity and depression require more than one question to be constructed as described below.

The smoking variable required three questions from the smoking section of the questionnaire to be constructed. The first question asked the subject whether they had smoked at least 100 cigarettes in their entire life, the second question asks whether they are currently smoking, and the third questions asks if during the last 12 months they had stopped smoking at least one day with the intention to quit smoking. Four categories can then be created from these questions as follows:

- if the subject indicates that they have not smoked 100 cigarettes in their life then they are a non-smoker,
- if they have smoked 100 cigarettes in their life but they are currently not smoking then they are ex-smokers,
- if they have smoked 100 cigarettes in their life, are currently not smoking and have stopped smoking at least one day in the 12 months for the intention to quit smoking then they are persons attempting to quit or quitters,
- if they have smoked 100 cigarettes in their life and are currently smoking then they are smokers.

To construct the binary smoking status variable, the smoking variable is reduced to two categories where smokers and quitters are combined to be the current smoker category, and ex-smokers and non-smokers are combined to be the new non-smoker category.

For alcohol consumption, the questionnaire used in the interview asks how many days did the subject drink at least one unit of alcohol in the last 30 days, if the answer is zero days than this was classified as a non-drinker. Otherwise, the following question asks on average how many units of alcohol they drank per day. If for females this was more than one on average per day and for males more than two on average per day, then the person is classified as a high risk drinker. If less than this amount, then the person is classified as a low risk drinker. Therefore three categories are created for the alcohol variable: non-drinker, low risk drinker, and high risk drinker.

For the depression variable two questions are used to construct the variable following the technique of Binkin *et al.* (2010), who have also used PASSI data for studying depression. The first question asks for how many days in the last two weeks did the person feel they had little interest and did not want to do anything, and the second question asks for how many days in the last two weeks did the person feel depressed or that their moral was low. For both variables, the responses were categorized as: 0 for 0-1 days, 1 for 2-6 days, 2 for 7-11 days and 3 for 12-14 days. These categories were then combined; if the total was from 3-6 then the person was categorized as depressed, and from 0-2 as non-depressed.

The physical activity variable was constructed by the PASSI team and found in the dataset. There are three categories to this variable: active, partially active, and sedentary. An active person is considered a person that performs heavy work or has a job that requires a lot of physical effort, that performs moderate physical activity for at least five days a week for 30 minutes, or performs vigorous activity at least three days a week for more than 20 minutes. A partially active person is a person who does not have a heavy physical job but still does some physical activity in their free time, however without reaching the recommended physical activity guideline levels. A sedentary person is a person who does not have a heavy physical job and also does not exercise in their free time.

4.3 Parametric Smoking Model

Before conducting the analysis required for the varying coefficients, a parametric model was found to describe the independent variables involved and their effect on the smoking status binary outcome variable. The proportion of smokers with time was found to be decreasing as shown in Figure 4.1. For performing the analysis, no observations were removed and the sample size was 148,266 individuals from the years $2008 - 2011$. The variables used for this model are summarized in Table A.1 in the Appendix. The proportions of each independent variable by the outcome status is shown in Table A.2. Also shown are the unadjusted and adjusted odds ratios with 95% confidence intervals, which were produced using a logistic model with the smoking status outcome variable. The results in Table A.2 indicate that all the independent variables were significant in the unadjusted models. In

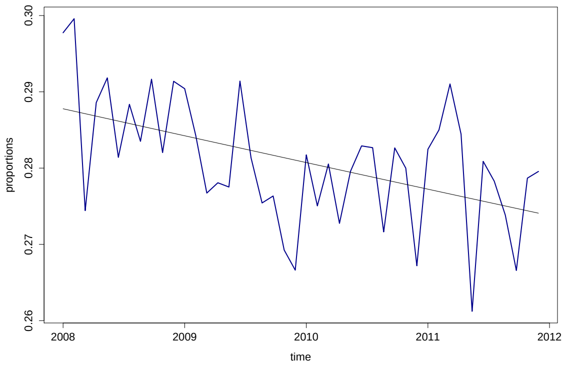


Figure 4.1:]
Proportion of smokers with time with a logistic trend line

the adjusted model, only the citizenship variable and the 30-39 age group category lost their significance. The highest odds ratios were found for high risk drinkers with an odds ratio of 2.31 (C.I. 2.22-2.40), which indicates that high risk drinkers have more than twice the odds of being smokers than those who do not drink. The best parametric model was found by both forward and backward selections using AIC, and this was found to be the model which excluded the citizenship variable. This model had practically the same estimates as those shown in the Table A.2 and therefore are not presented here. Therefore, the smoking status varying coefficient models constructed will exclude the citizenship variable.

4.4 Comparison of Methods for Smoking VCM

The analysis is performed on a smoking status outcome variable with a comparison of different non-parametric estimation methods used for varying coefficient models. The five methods compared are:

Method I: Polynomial spline,

Method II: Penalized spline regression,

Method III: P-spline using the `gam` function,

Method IV: P-spline using the `bam` function,

Method V: Smoothing spline with a cubic spline using the `bam` function.

Knot selection was performed for Method I and Method II which requires more computation time. Method III, IV and V do not require knot selection, and the maximum number of knots of 44 knots were used to fit the models using these methods. Method III and IV are the same except for the function used from the `mgcv` package. The `bam` function should give very similar results to the `gam` function but with much faster computation. The same basic methodology is used for the different varying coefficient model estimation methods, and the comparison is made between these methods particularly for studying the practicality and the computational feasibility of these methods for use in surveillance data analysis. In addition the estimates and plots are compared to observe any significant changes between these methods. The recommended method is then described in more detail in a separate section.

The steps described in the methodology chapter were used for the five methods to be compared. The main purpose of this comparison was to study the reliability and practicality of these methods in its application to health surveillance data. For large data sets such as in health surveillance data, a main obstacle can be computation time. For polynomial spline estimation, while using fewer knots and no penalty can decrease computation time, the computation time required for selection of knots can be large. While parallel computation is used for the selection of knots of different variable simultaneously, the computation time can still be large when there are many variables in the model. In penalized regression spline estimation, a penalty is added and the knots are again selected, this increases the computation time even further. Another option was to use P-splines which does not require selection of knots since a penalty is added. Finally the smoothing spline estimation method can also be used which also has a penalty which differs from the penalty used in P-splines. The results and computation times are compared in further detail in this section.

Following the first and second step described in the methodology chapter, each independent variable is tested to see if it has coefficients that are time varying. To perform this test, a model is found in which each independent variable is allowed to have time varying coefficients while all other coefficients are constant. Once this model is found, the test is performed. Table 4.1 shows the number of knots and the time in minutes to fit of each of these models as described in step one of the methodology. Also shown in the table are the p-values of the chi-square test performed (with p-values less than 0.1 indicated in bold); this tests the model against the parametric null hypothesis as explained in step two of the methodology chapter.

Table 4.1: Comparison of methods in finding the coefficients of variables which are time varying using a chi-square test (with reported p-values). Models are for each variable with varying coefficients and remaining variables with constant coefficients using a smoking status dependent variable.

Model	Method I			Method II			Method III		Method IV		Method V	
for var*	k	t	p-value	k	t	p-value	t	p-value	t	p-value	t	p-value
age	5	109.0	**0.020**	9	660.9	**0.001**	54.7	**0.001**	2.8	**0.002**	2.5	**0.002**
sex	5	43.0	0.448	9	105.2	0.10	10.5	0.143	0.9	0.178	0.9	0.178
mstatus	5	61.6	0.588	5	213.1	0.136	39.7	0.216	1.5	0.212	1.3	0.212
edu	5	66.9	0.367	12	455.5	**0.040**	48.8	**0.066**	2.0	0.212	1.9	0.212
inc	5	55.4	0.150	5	264.5	**0.015**	30.5	**0.035**	1.4	**0.034**	1.3	**0.035**
work	5	33.6	0.770	16	129.0	0.144	18.6	0.381	0.9	0.375	0.8	0.375
region	5	49.2	0.297	19	276.8	**0.036**	23.9	**0.062**	1.8	0.351	1.3	0.351
phy	5	53.4	0.159	18	236.5	**0.002**	26.4	**0.008**	1.5	**0.062**	1.2	**0.062**
alco	5	55.1	**0.060**	7	282.1	**0.018**	38.5	**0.023**	1.4	**0.060**	1.2	**0.060**
depress	8	33.6	0.453	41	127.1	**0.030**	17.9	0.295	0.9	0.293	0.8	0.293
Total		560.9			2750.7		309.5		15.1		11.9	

Notes: * Variable abbreviations found in Table A.1. k is for knots and t is for time. Method III, IV and V all used 44 knots.

As shown in the Table 4.1, all the methods selected age and alcohol consumption to have significant time varying coefficients at the 0.1 level. Methods II, III, IV and V also selected income and physical activity as having time varying coefficients. Method II and III gave significant p-values in the test for education and region, and Method II included depression as well. In addition to the listed independent variables in Table 4.1, a test was performed for a model with varying coefficients for the time variable alone (i.e. to have a time varying intercept) against the parametric model as the null hypothesis. This model was rejected for all five methods.

In terms of computational time, the largest were for Method I and II since these methods required knot selection, with Method II having a higher computation time since there is also an added penalty. Method III used the P-spline method with 44 knots and the `gam` function and this was faster than

Method I and II, however the computation time was still relatively large. Using the `bam` function significantly decreased the computational time and provided similar results as Method III. Method IV and V have very similar computation times and results, both methods use a penalty and 44 knots with the `bam` function.

Once the variables with varying coefficients are found, the next step requires the building of the varying coefficient model as described in step three of the methodology chapter. This was done by starting with the variable which gave the highest percent of deviance explained; for all the methods this was the variable age. Then to see which variable to add next, the residuals of the model with age varying coefficients was fit with each of the remaining variables that gave a significant p-value in Table 4.1 to see which variable gives the best explanation. The percent deviance explained were compared to know which variable gave the best explanation and therefore should be added next. The order of adding the variables differed from one method to the next as shown in Table 4.2. The table also shows the time required to fit each model as well as the p-value test of testing each model with the lower model in the step wise building of the model.

The results of the model building process for the five methods resulted in the selection of the same model for Methods II, III, IV and V. This model includes time varying coefficients for the variables age, alcohol consumption, income level and physical activity, and can be written as

$$\begin{aligned}\text{logit}(SMK) = b_0 + \sum_{j=1}^{p} b_j Z_j + a_1(t)age + a_2(t)alcohol + a_3(t)income\\ + a_4(t)phyical,\end{aligned}$$

where Z_j are all the remaining independent variables with constant coefficients b_j. Although, as shown in Table 4.2, the selected model for Method II was Model 3, it is possible that Model 4 or 5 could have been selected. However, is unknown since these models were not completed and required a large computation time ($> 80,000$ minutes) and memory (> 70GB) to complete, and it was therefore terminated. However, it is not expected that Model 4 or 5 would have been chosen since most of the methods selected Model 3, and therefore this is the likely model for this application. In Method I, which uses polynomial spline estimation, a different model was selected. The selected model for this method, Model 1, includes time varying coefficients for only the age and alcohol variables and can be written as

$$\text{logit}(SMK) = b_0 + \sum_{j=1}^{p} b_j Z_j + a_1(t)age + a_2(t)alcohol.$$

Another step was performed in which the addition of the time varying coefficient for the intercept was tested again for inclusion in the final models. For Methods I and II the inclusion of a spline of time was rejected, however for Methods III, IV and V, the p-value of the test was significant although borderline for Methods IV and V (p-values 0.092 and 0.095 respectively). Therefore, the final model now becomes

$$\begin{aligned}\text{logit}(SMK) = b_0 + \sum_{j=1}^{p} b_j Z_j + a_0(t) + a_1(t)age + a_2(t)alcohol + a_3(t)income \\ + a_4(t)phyical,\end{aligned}$$

with $a_0(t)$ the new addition which allows the intercept to vary with time as well. The results of the estimates and the spline plots did not change much with this addition.

As for the computation times, since Method I only required the addition of one more variable with a small number of knots, the computation time was very small. The largest computation time was for Method II, and the total time shown in Table 4.2 is only for reaching Model 3. Methods IV and V had the lowest computation times as these methods both use the `bam` function for large datasets designed to cut computation time. The benefits of using the `bam` function is clear and especially when comparing Method III and IV, which use the same estimation methods with P-splines. As shown in Table 4.2, Method III, which uses the `gam` function, required over 5000 minutes to find the final varying coefficient model while Method IV, which uses the `bam` function, only required 29 minutes. Based on the computational time and practicality of the method, it is recommended to use Method IV which uses the P-spline method with 44 knots and the `bam` function. Although Method V has similar computation time, Method IV allows the selection of the degree of the spline and the order of the penalty separately, which gives more flexibility to researchers in applying this method to different settings.

Table 4.2: Comparison of methods in building of the smoking status time VCM, starting with a VCM with age time varying coefficients.

	Method I			Method II			Method III			Method IV			Method V		
Model	vars	t	p-value	vars	t	p-value	vars	t	p-value	vars	t	p-value	vars	t	p-value
1	+alco	0.9	**0.032**	+ inc	20.1	**0.012**	+ phy	135.6	**0.007**	+ inc	5.3	**0.029**	+ inc	4.7	**0.029**
2	-	-	-	+ inc + phy	780.5	**0.002**	+ phy + inc	462.7	**0.034**	+ inc + alco	9.1	**0.045**	+ inc + phy	5.3	**0.048**
3	-	-	-	+ inc + phy + alco	6624.0	**0.013**	+ phy + inc + alco	1141.1	**0.011**	+ inc + alco + phy	14.6	**0.035**	+ inc + phy + alco	14.4	**0.035**
4	-	-	-	+ inc + phy + alco + depress	NC	NC	+ phy + inc + alco + region	1877.2	0.632	-	-	-	-	-	-
5	-	-	-	+ inc + phy + alco + depress + edu	NC	NC	+ phy + inc + alco + edu	2426.5	0.343	-	-	-	-	-	-
Total		0.9			7424.6*			5016.4			29.0			24.4	

Notes: Variable abbreviations found in Table A.1. t is for time. NC - not completed due to models running more than 80,000 minutes.

Once the final varying coefficient models have been selected for each method, the models can then be examined further to observe if there are any significant differences between the methods. The estimates of the constant coefficients are not discussed here as these do not change very much from one method to the next and is not the main interest here. Table 4.3 gives the p-values of the splines for all the five methods with p-values less than 0.1 indicated in bold. These p-values are not very reliable as they are underestimated so that any border line significant p-value could actually be non-significant (Wood, 2006a). The p-values would behave correctly for un-penalized models, however when a penalty is added the calculation of the p-values neglects the smoothing parameter uncertainty in the reference distributions used for testing (Wood, 2006a). Therefore, if there is a non-significant p-value in Table 4.3 then the term is probably not needed in the model. However, if there is a significant p-value, this should be taken with caution and especially if the p-value is borderline significant such as that of the low income category in Method III and IV. Therefore it is better to rely on the spline plots to observe any changes in the coefficients over time.

To see the differences in the results of the splines estimations between the methods more clearly, plots of the coefficients over time are shown in the Appendix in Figures A.1 for the alcohol categories, A.2 for the age categories, A.3 for the income categories, and A.4 for the physical activity categories. Note that for Methods III, IV, and V, the models include a time varying intercept (plots for this not shown) while for Methods I and II there is no time varying intercept in the model. The plots have a horizontal line at zero so as to clearly see where the coefficients are negative or positive, and also show how they are changing from the constant line. From the plots, one can see that excluding Method I, the plots look very similar and especially for Methods III, IV and V. The plots indicate that there are some trends that are not constant for many of the categories of the variables. Most of these trends appear linear, however for some categories such as high risk drinker and low income in Methods III, IV and V, a non linear trend is observed in which there is an increase and then a decrease in the trend after the year 2010.

Table 4.3: Reported p-values for the spline estimates of the selected smoking status VCM for each of the five methods used in the analysis.

Variable	Method I	Method II	Method III	Method IV	Method V
Age					
18-29	0.134	0.202	0.380	0.388	**0.012**
30-39	**0.022**	**<0.001**	**0.017**	**0.017**	0.300
40-49	0.100	0.777	0.865	0.834	**0.062**
50-59	**0.017**	0.036	0.166	0.159	0.992
60-69	**0.017**	0.920	0.953	0.924	0.160
Income					
High	-	0.491	0.324	**0.054**	0.363
Medium	-	0.803	0.926	**0.047**	0.991
Low	-	**0.015**	**0.068**	**0.093**	0.266
Physical activity					
Active	-	0.636	0.992	0.556	0.557
Partially active	-	**0.017**	0.154	**0.016**	**0.016**
Sedentary	-	**0.070**	0.403	0.989	0.985
Alcohol					
Non-drinker	**0.023**	**0.004**	0.319	0.839	0.246
Low risk drinker	**0.028**	**0.071**	0.986	0.478	0.819
High risk drinker	0.209	0.338	**0.077**	0.155	0.139

4.5 Description of Smoking VCM from the Recommended Method

The recommended method resulting from the comparison of the five methods in the previous chapter, is Method IV which uses P-spline estimation and the `bam` function. This is mainly due to the relatively fast computation time, and the flexibility the method provides in selecting the degree of the spline and the order of the penalty separately. The selected varying coefficient model for this method was the same as the model selected for Method II, III and V, and is the model which includes varying coefficients for the variables age, income, alcohol and physical activity and also with the addition of a time varying intercept, i.e.

$$\begin{aligned}\text{logit}(SMK) = b_0 + \sum_{j=1}^{p} b_j Z_j + a_0(t) + a_1(t)age + a_2(t)alcohol + a_3(t)income \\ + a_4(t)phyical.\end{aligned}$$

The previous section concentrated on the comparison of the five methods in obtaining the final varying coefficient model. In this section however, the details of the model from the recommended method will be discussed. The summary estimates of this selected model, including the odds ratios (OR) with 90% confidence intervals and p-values, are shown in Table 4.4. The constant estimates in the model did not change significantly when compared to the parametric model (not shown). The only main change is for the age category 30-39, which was not significant in the parametric model with an OR of 0.9, and was now found to be significant with an OR of 1.1. Again caution should be taken for the p-values of the splines as they can be underestimated, therefore the p-values for the income categories splines reported in Table 4.4 could all be in-fact non-significant.

To understand the trends in the varying coefficients, odds ratio plots are produced for all the varying coefficients in the selected model as shown in Figure 4.2. Unlike the time varying coefficient plots shown in the appendix for the five methods, odds ratio plots cannot be produced for the reference categories. To observe the changes in the reference categories the coefficient plots in the appendix can be examined; for instance, the reference category for age 18-29 has a slightly decreasing trend. The odds ratio plots are on an exponential scale since the exponential of the estimates are taken to obtain the odds ratios, and all the plots contain Bayesian confidence intervals.

For the age categories, the categories 40-49 and 60-69 have a constant trend and therefore these ORs are not changing over time. However, the age category 30-39 has a clear linear and increasing trend above one. This indicates that those in this category are increasing their odds with time of becoming smokers compared to the 18-29 age group. The age category 50-59 has a slightly increasing linear trend with ORs below one. This indicates that while those in this age category have lower odds of being smokers than the reference group, this odds is increasing, and therefore this age category is becoming more similar to the reference group over time.

Table 4.4: Summary of estimates with reported odds ratios and p-values for the selected smoking status VCM - Model 3 of Method IV (p-values < 0.1 indicated in bold).

Variable	**OR** (90% **C.I.**)	**p-value**
Age (Ref: 18-20)		
30-39	1.11 (1.05-1.16)	**0.011**
40-49	0.89 (0.85-0.94)	**0.004**
50-59	0.85 (0.81-0.90)	<**0.001**
60-69	0.52 (0.49-0.56)	<**0.001**
s(time):18-29	-	0.388
s(time):30-39	-	**0.017**
s(time):40-49	-	0.834
s(time):50-59	-	0.159
s(time):60-69	-	0.966
Sex (Ref: Female)		
Male	1.58 (1.55-1.60)	<**0.001**
Marital Status (Ref: Married)		
Single	1.47 (1.44-1.50)	<**0.001**
Widowed/Divorced	1.84 (1.79-1.90)	<**0.001**
Education (Ref: University or higher)		
High school	1.35 (1.32-1.39)	<**0.001**
Middle school	1.80 (1.75-1.85)	<**0.001**
Primary school or less	1.41 (1.36-1.46)	<**0.001**
Income (Ref: High)		
Medium	1.24 (1.18-1.30)	<**0.001**
Low	1.50 (1.38-1.63)	<**0.001**
s(time):High	-	**0.054**
s(time):Medium	-	**0.047**
s(time):Low	-	**0.093**
Work status (Ref: Work)		
Do not work	0.75 (0.73-0.76)	<**0.001**
Region (Ref: North)		
Centre	1.21 (1.18-1.23)	<**0.001**
South	1.10 (1.08-1.13)	<**0.001**
Physical activity (Ref.: Active)		
Partially active	0.90 (0.87-0.94)	<**0.001**
Sedentary	1.17 (1.12-1.22)	<**0.001**
s(time):Active	-	0.556
s(time):Partially	-	**0.016**
s(time):Sedentary	-	0.989
Alcohol consumption (Ref: non-drinker)		
Low risk drinker	1.55 (1.50-1.61)	<**0.001**
High risk drinker	2.23 (2.05-2.43)	<**0.001**
s(time):Non-drinker	-	0.839
s(time):Low risk drinker	-	0.478
s(time):High risk drinker	-	0.155
Depression (Ref: Not depressed)		
Depressed	1.43 (1.38-1.47)	<**0.001**

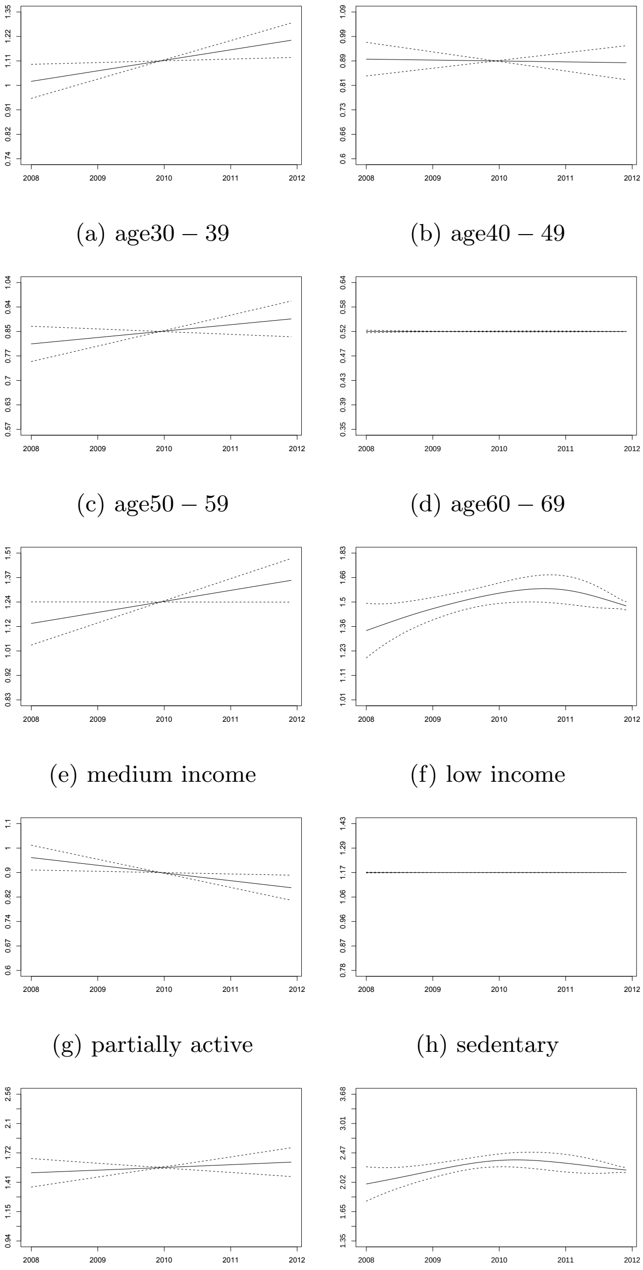
Notes: Ref - reference category, OR - odds ratios with 90% confidence intervals.

For the income categories, the medium income category shows an increasing OR linear trend that is above one. This indicates that those with medium income have increasing odds over time of becoming smokers compared to those with high income. A slight non-linear trend can be observed for low income category where the trend is increasing until 2010 after which it appears to change direction to be constant or decrease. More years of observation may be required to see if this trend is actually decreasing.

For the physical activity categories, we can see that only the partially active category gives a linear trend while the sedentary category remains constant over time. For the partially active category, the trend is decreasing and below one. This indicates that those who are partially active are decreasing their odds of being smokers compared to active persons.

Finally in the alcohol categories we have a relatively constant trend for the low risk drinker category. However, we can observe a slight non-linear trend for the high risk drink category. For this category the trend was slightly increasing until the year 2010 after which it starts to become constant. Therefore for high risk drinkers, while the odds of being a smoker was increasing over time compared to non-drinkers, this increase apparently ceases after 2010.

Performing the usual model diagnostics by plotting the residuals versus the fitted values is not useful in the case of having binary data (Wood, 2006a). Therefore, another technique as proposed by Wood (2006a, p. 115) can be used in which we take the fitted values of the final time varying coefficient model and produce simulated binary independent data from these values. The same final time varying coefficient model is then fit using the simulated data, then the residuals of the model from the simulated data are compared to the residuals from the final model. This is conducted by ordering both residuals according to the fitted values, and then observing whether one has fewer or more values above or below zero. For this comparison, an initial check of the plot of each of these residuals versus the fitted values produced very similar plots as shown in Figure 4.3. A further check was made by fitting a gam model of the residuals over zero with a spline of the fitted values (comparable to the usual plot of residuals versus fitted values in Gaussian case). This also produced very similar plots and estimates; therefore indicating that the residuals from the final varying coefficient model above and the residuals from the same model fit with the simulated inde-



(a) age30 − 39

(b) age40 − 49

(c) age50 − 59

(d) age60 − 69

(e) medium income

(f) low income

(g) partially active

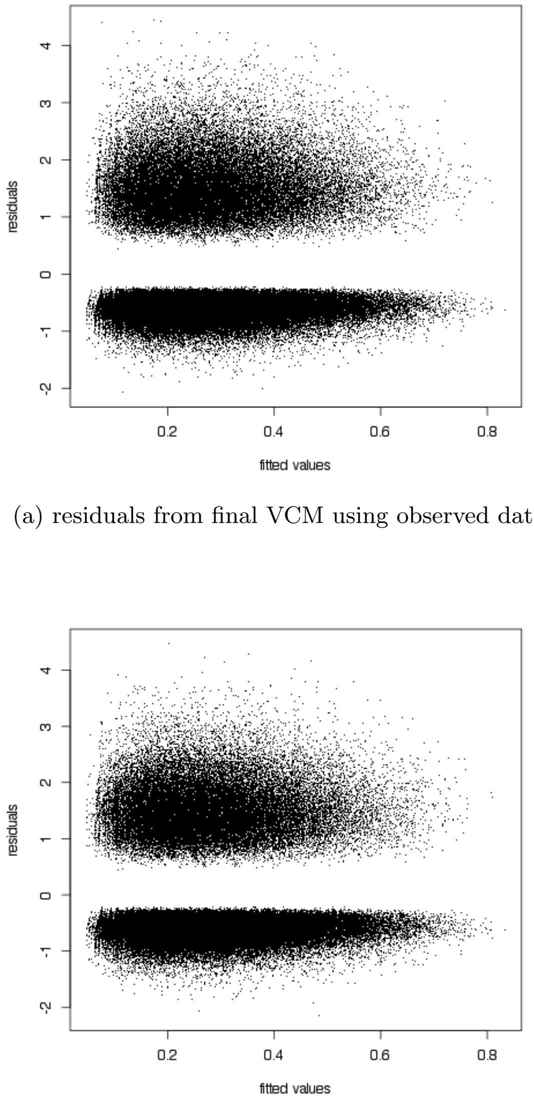
(h) sedentary

(i) low risk drinker

(j) high risk drinker

Figure 4.2: Odds ratio plots of the smoking status VCM.

pendent data are similar. This could provide sufficient evidence that the final varying coefficient model has residuals that are independent. This was conducted using one simulation however, as there is a high computation time required to perform several simulations due to the size of the model.



(a) residuals from final VCM using observed data

(b) residuals from final VCM using simulated data

Figure 4.3: Comparison of residuals from final VCM model between observed and simulated data.

Chapter 5

Results: USA Analysis

Two types of analysis were carried out using U.S. BRFSS data. Part I involves all the U.S.A. states for the years between 1993 and 2009 for a time varying coefficient model, and Part II involves using the county level data for Florida state for the year 2010 for a spatial varying coefficient model. Both analyses used spline estimation for fitting varying coefficient models for an obesity status binary outcome variable, with some differences as will be shown in more detail. This chapter will begin with a description of the data used for each part of the analysis, followed by the methodology used to construct the models (with only slight changes from the methodology chapter), and finally the results.

5.1 U.S. BRFSS Data

The data used for the analysis in this chapter is from the U.S. BRFSS (Behavior Risk Factor Surveillance System) data of the Center of Disease Control in the United States of America (Center for Disease Control and Prevention (CDC), 2014). The U.S. BRFSS currently covers all the states and territories of the U.S.A., and has collected monthly data since 1984, making it the largest and longest running BRFS of its kind (Mokdad, 2009).

For Part I, the data used for analysis was restricted to include the years from 1993 to 2009 with gap years of 2004, 2006, and 2008 (i.e. 14 years of monthly data). This was due to the availability of the questions that were asked by each state. Some of the variables required in the analysis were derived from questions that were not asked each year, or not asked by all the states. For instance the fruit variable is derived from questions concerning the fruit and vegetable intake of individuals and these questions were asked on the years 1994, 1996, 1998, 2000, 2002, 2003, 2005, 2007, and 2009 for all the states, and on 1993, 1995, 1997, and 1999 for some of the states. It was not asked for any of the states in 2004, 2006, 2008 and 2010, therefore this restricted the analysis to not include these years. The same was true for the question regarding physical activity except that physical activity was asked in the year 2010. However, most of the variables used in the analysis were derived from questions asked each year for all the states, i.e. core questions. These include questions on socio-demographic aspects (age, sex, marital status, etc), as well as smoking status. The questions on whether the individual has access to health care and on their general health started in 1993, and therefore the analysis was selected to begin in this year. As for the outcome of interest, the proportion of obesity in the USA between 1993-2009 was found to be steadily increasing as shown in Figure 5.1 below.

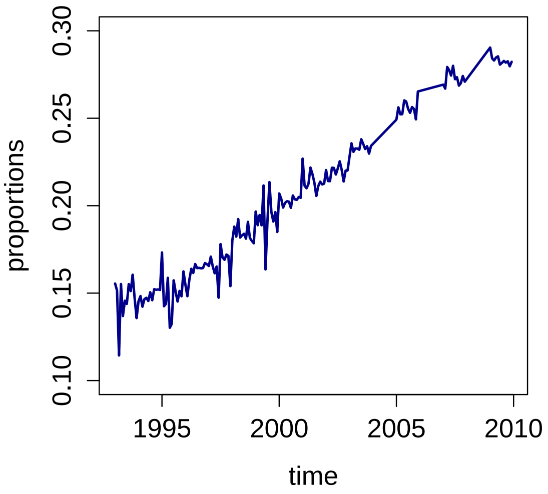


Figure 5.1: Proportion of obesity with time

For Part II, the 2010 county level data used for Florida state was selected, as this dataset contains a reasonable sample size per county to be able to conduct the analysis. Using this data, with the 67 counties of Florida, a spatial varying coefficient model could be constructed. The variables used in this analysis were the same variables as those found in Part I, however the fruit variable was not included as this was found in 2010, and of course the region variable was not applicable to this case. The proportions of obesity for each county in 2010 are shown in Figure 5.2, which shows that the proportions of obesity are varying by county.

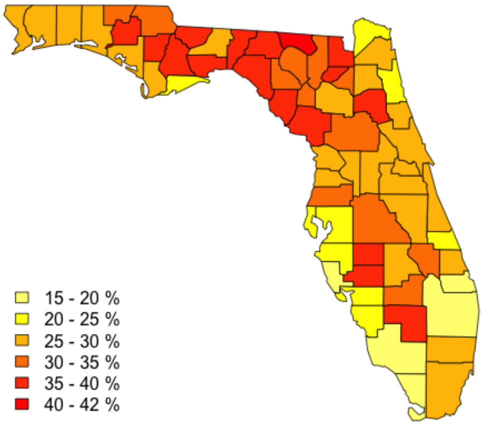


Figure 5.2: Proportion of obesity by county in Florida in 2010.

5.2 Data Preparation

5.2.1 Part I: Time VCM

Initially, after eliminating all the missing values for the variables used in the analysis, a total sample size of 1,813,072 observations remained for the 14 years of available data. However, one variable derived from a question on household total earnings, used to construct the income variable, had a large number of missing values (13.8%). To reduce the number of missing values of this variable in order to include it in the analysis, a prediction was made with the available data using a model with income as the outcome variable and several socio-demographic variables as the predictors (in this case an ordered logistic regression was used). This model was then used to impute the missing values for the income variable. After the imputation of the missing values, the new sample size increased to 2,065,689 observations available for analysis. The variables used in the analysis are found in Table B.1 of the Appendix, and mainly involved combining certain categories in the original questions. The income variable however was constructed from household earning while taking into consideration the poverty threshold for each year as well as the number of persons in a household (US Department of Health and Human Services, 2013). If the total earnings was below the poverty threshold for that year, then this was classified as a low income category. If the total earnings was three times more than the poverty threshold, then this was classified as high income, and all remaining values between these two extremes was classified as medium income.

5.2.2 Part II: Spatial VCM

For the spatial analysis using Florida state county level data for 2010, the sample size before imputing the missing values for the income variable, as was conducted in Part I, was 27,678 observations, and after imputing the missing values it was 32,110 observations. The same variables as the those found in the analysis of Part I were used except for the absence of the fruit and region variables. The descriptives of the variables in this part of the analysis is found in Table B.2 in the Appendix.

5.3 Constructing the Varying Coefficient Models

For the analysis of the USA data, there was no comparison of methods as was conducted with with Italian data analysis. Therefore, P-spline estimation was used for Part I and a smoothing spline estimation using cubic regression splines and tensor products was used for Part II. To construct the model for both Part I and Part II, the same procedure was followed as described in the methodology chapter. Briefly, the first step was to fit a varying coefficient model for each variable while leaving all other variables with constant coefficients (as in the examples shown below), then this model was tested against the null hypothesis which contains a model where all the variables have constant coefficients. A significant p-value would indicate that the coefficients for the tested variable are actually varying. This procedure is followed for each variable before beginning to build the varying coefficient model. To build the varying coefficient model, each new addition of a variable with varying coefficients is tested with a lower model where the variable being tested has constant coefficients. This procedure is performed until the full model is reached. Further detail on the methodology used for Part I and Part II is described below.

5.3.1 Part I: Time VCM

For this part of the analysis the modifying variable is time, and for the 14 years of observation of monthly data, this provides 168 months. The models were fit using a P-spline method using a third degree B-spline and a second order difference penalty. A higher number of knots would greatly increase the computation time and particularity for large datasets, and therefore placing the maximum number of knots of 168 knots would be computationally infeasible. Therefore, 42 knots were selected to fit the models (three for each year), and the sufficiency of the number of knots was checked using the `gam.check` function of the `mgcv` package in the `R` software program. As described previously in methodology chapter, the following example `R` code was used to fit the varying coefficient models using the `mgcv` pacakge:

```
bam(OBS ~ Zj + income + s(time, bs = "ps", k = 42, m=c(3,2),
                   by = income), family = binomial("logit")).
```

Here `OBS` is the binary outcome variable for obesity status. The `bam` function is used for fitting GAM models for large datasets, `k="ps"` is for P-splines, `k=42` is to select 42 knots, and `m=c(3,2)` indicates that a third degree B-spline is used with a second order difference penalty. The example above shows that we are using obesity as an outcome and income as the covariate which contains varying coefficients.

Once the final time varying coefficient is found, odds ratio plots can then be constructed for each category of the variable which were found to have varying coefficients. These plots can then be used to see the changing odds ratio trends of these categories on the outcome with respect to the reference category.

5.3.2 Part II: Spatial VCM

For the space varying coefficient model for Florida state, the modifying variable is space. Here the space variable is represented by the longitude and latitude coordinates of the centroids of each county in Florida. These were obtained from the National Center for Ecological Analysis and Synthesis of the University of California in Santa Barbara (NCEAS, 2013). The file obtained from this source contained the centroids as well as the fips codes (Federal Information Processing Standards) for each county in the USA.

This was matched with the fips codes in the Florida dataset in order to import the coordinates; a Florida map showing the location of the resulting coordinates was checked to ensure accuracy. The models were then fit using smoothing spline estimation with tensor products of cubic regression splines. The number of knots used was ten for each coordinate, therefore creating 100 knots for each tensor product. The following is an example of the code that was used:

```
bam(OBS ~ Zj + income + te(x,y, k = 10, by = income),
                    family = binomial("logit")).
```

In this example `te` represents tensor product smooths, and the longitude and latitude of the centroids are represented by the variables `x` and `y`. Here we are again using obesity as the outcome with an example shown for the income covariate.

The function `te` is invariant to linear rescaling of covariates but not to the rotation of the covariate space (i.e. it is not isotropic). Thin plate splines on the other hand are isotropic, but are not invariant to rescaling of covariates. Tensor products can also use variables measured in different units (for example if the analysis requires a smooth of space and time), and they are also more computationally efficient than thin plate splines (Wood, 2006a). To select between the use of the tensor product or the thin plate splines, a chi-square test was used between the two models, which were the same except for the type of function used (`te` versus `s` with bs=”tp”) (Wood, 2006a). This resulted in the rejection of the model which uses thin plan splines.

The maps used to represent the results show the probabilities of obesity for the indicated category at each centroid while keeping all the other categories constant at the reference level. A reference map is also constructed which contains the probabilities of obesity with all the categories at the reference level. This map can be used to compare to the remaining maps to see how the probabilities differ across categories. In addition, each map can be observed separately to see any spatial variations of the probabilities of obesity for that category. In other words, observing different probabilities across the counties show how the coefficient is varying spatially.

5.4 Results

The results focus on discussing the varying coefficient models from the analyses in Part I and Part II. The parametric results, which involves fitting a logistic model with obesity as the binary outcome, are found in Table B.3 for Part I and Table B.4 for Part II in the Appendix. These tables summarize the unadjusted and adjusted odds ratios of the constant terms with their confidence intervals, these results are discussed briefly below.

In general for Part I, all the variables gave significant unadjusted and adjusted odds ratios in the logistic model except for the race category other race (non-Hispanic minority groups). The highest odds ratios belonged to the race category black (adjusted OR 1.71 C.I. 1.69-1.73), and for having poor to fair health (adjusted OR 1.86 C.I. 1.84-1.87). Some trends in the ordered categories were found such as increasing odds ratio of obesity with age (except for the 65 and older age category), and with decreasing income level. For the health risk variables, not performing any physical exercise gave a higher odds ratio of obesity (1.48 C.I. 1.47-1.49) compared to those who do exercise. The analysis also found that current smokers had a lower odds of obesity than non-smokers (OR 0.61 C.I. 0.60-0.61).

For Part II, using Florida county level data for the year 2010, all the variables gave significant odds ratios in the unadjusted models except for the categories of other for the race variable, and the combined widowed, divorced, or separated category for the marital status variable. In the adjusted model however, several categories lost their significance. All the categories for marital status, the high income category, the does not work category, the Hispanic and other race categories, and the have a health plan category did not have significant odds ratios in the adjusted model. There was a general agreement in the odds ratios between the results of Part I and Part II except for the 65 and older category which now had a lower odds of obesity compared to the reference group in Florida. Note that in Part II the reference category for the variable sex and smoking status was changed compared to the analysis in Part I, and this was mainly due to the presence of spatial variation in the female and non-smoker reference groups in the final spatially varying coefficient model (there was little or no spatial variation in males or current-smokers). Therefore, the reference categories were changed in order to highlight these spatial variations in these categories. Finally, it was found that the variable marital status was not required in the parametric Florida model, and therefore the building of the spatially varying coefficient model does not include this variable. The coefficients in the parametric logistic model excluding the martial status variable do not change very much from those in Table B.4, and are therefore not reported.

5.4.1 Part I: Time VCM for USA

To begin the procedure of constructing the varying coefficient model, a model was required which contains all the independent variables with constant coefficients. This model represents the model used in the null hypothesis of the tests which determine whether the coefficients are actually varying. In the Italian data analysis, this was the parametric logistic model since in the first step a time varying intercept was rejected. However, the analysis for the U.S. data has shown that the intercept was significantly changing with time, therefore the model used as the null hypothesis to test each variable for varying coefficients is:

$$\text{logit}(OBS) = b_0 + \sum_{j=1}^{p} b_j Z_j + a_0(t), \tag{5.1}$$

where `OBS` is the outcome variable obesity, $a_0(t)$ is the time varying coefficients for the intercept, and Z_j are the covariates with their constant coefficients b_j. The following step involved the testing of the addition of a time varying coefficient for each independent variable using a chi-square test with the function `anova` in `R`. This is then followed by the testing of nested models to find the final time varying coefficient model. Table 5.1 summarizes the process of building the model with p-values less than 0.05 indicated in bold.

As can be seen in the table, in the first step, all the variables gave a significant p-value when tested against Model 5.1. In the second step however, the smoke variable was found to not have time varying coefficients when other time varying coefficients were present in the model. The test for Model XII is not available, as this model required over 300 hours to process. The final model was therefore selected to be Model XI, which is the model that contains varying coefficients for all the variables except for the smoke and fruit variables. This model can be written as:

$$\textbf{Model XI} : \text{logit}(OBS) = b_0 + \sum_{j=1}^{p} b_j Z_j + a_0(t) + a_1(t)age + a_2(t)income + a_3(t)phy$$
$$+ a_4(t)edu + a_5(t)mstatus + a_6(t)region + a_7(t)race$$
$$+ a_8(t)sex + a_9(t)hplan + a_{10}(t)work + a_{11}(t)genhealth,$$

Table 5.1: Constructing the time VCM for obesity outcome in the USA (Part I).

Model	Description	time (min)	p-value	H_0 used
Selection of variables that have varying coefficients				
Model age	LM + s(t):age	13.0	<**0.001**	LM
Model sex	LM + s(t):sex	5.5	<**0.001**	LM
Model mstatus	LM + s(t):mstatus	7.7	<**0.001**	LM
Model edu	LM + s(t):edu	12.6	<**0.001**	LM
Model work	LM + s(t):work	12.6	<**0.001**	LM
Model race	LM + s(t):race	13.1	<**0.001**	LM
Model income	LM + s(t):income	14.8	<**0.001**	LM
Model region	LM + s(t):region	11.3	<**0.001**	LM
Model phy	LM + s(t):phy	7.0	<**0.001**	LM
Model smoke	LM + s(t):smoke	6.9	**0.041**	LM
Model fruit	LM + s(t):fruit	1.4	**0.009**	LM
Model hplan	LM + s(t):hplan	5.6	<**0.001**	LM
Model genhealth	LM + s(t):genhealth	6.9	<**0.001**	LM
Finding the full varying coefficient mode				
Model I	Model age + s(t):income	26.9	<**0.001**	Model age
Model II	Model I + s(t):phy	70.2	<**0.001**	Model I
Model III	Model II + s(t):edu:	235.6	<**0.001**	Model II
Model IV	Model III + s(t):mstatus	325.3	<**0.001**	Model III
Model V	Model IV + s(t):region	540.5	<**0.001**	Model IV
Model VI	Model V + s(t):race	656.1	<**0.001**	Model V
Model VII	Model VI + s(t):sex	785.3	<**0.001**	Model VI
Model VIII	Model VII + s(t):hplan	943.0	<**0.001**	Model VII
Model IX	Model VIII + s(t):smoke	957.2	0.588	Model VIII
Model X	Model VIII + s(t):work	1129.6	<**0.001**	Model VIII
Model XI	Model X + s(t):genhealth	1113.8	<**0.001**	Model X
Model XII	Model XI + s(t):fruit	> 18000.0	NC	Model XI

Notes: LM - Model 5.1, s(t) - spline of time, variable abbreviations found in Table B.1, NC - not completed.

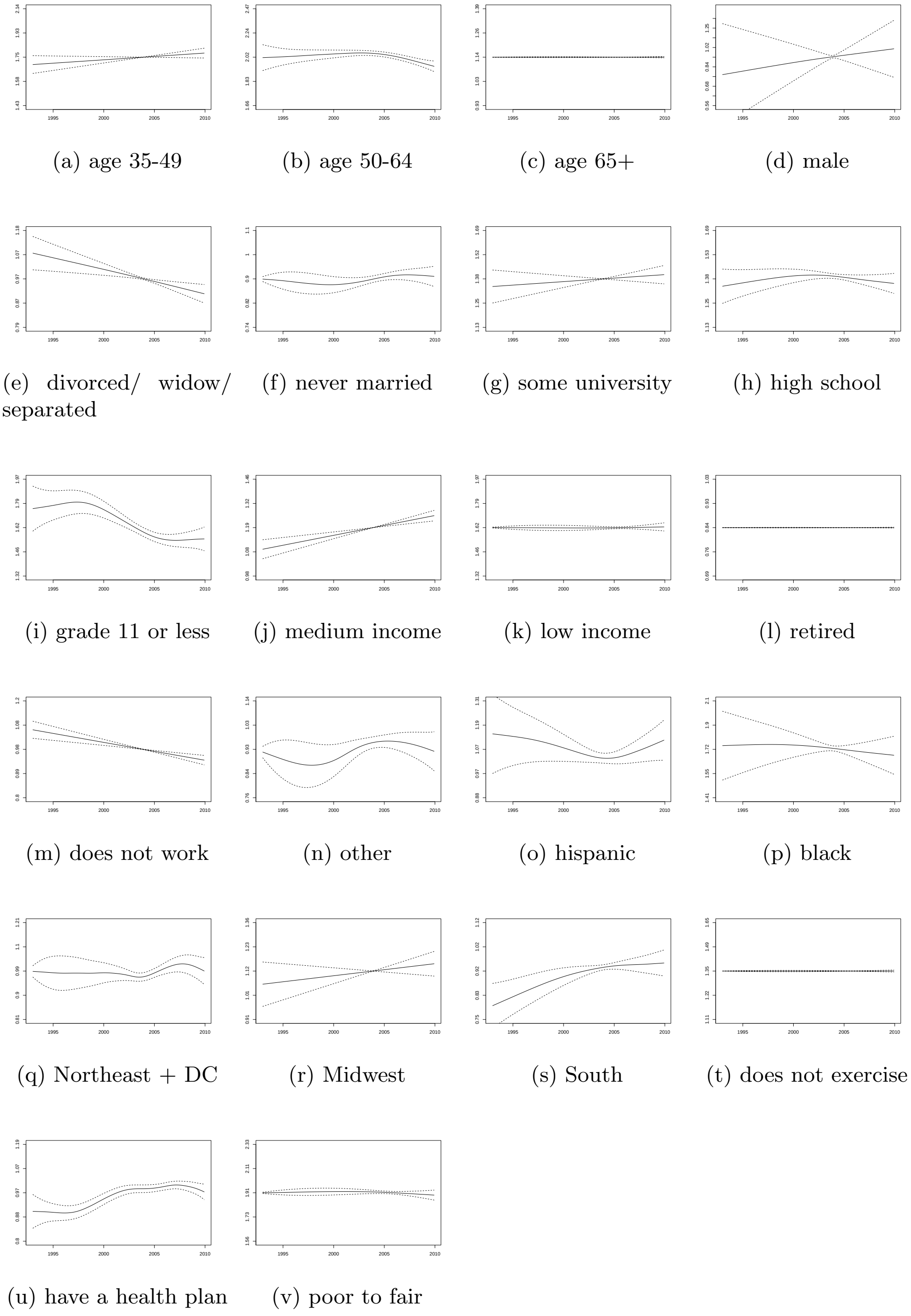
where again OBS is the outcome variable for obesity, and Z_j are the covariates with constant parameters b_j. The time varying coefficients, $\sum_{i=1}^{11} a_k(t)$, were found for the variables age, income, physical activity, education, marital status, region, race, sex, access to health care, work status, and general health status. The time varying coefficient $a_0(t)$ allows the intercepts to vary with time. The high number of covariates that were found to be significantly varying with time may be due to the large observation period used in the analysis. For this length of time, it is expected that some change would occur in most of the coefficients which is captured by the model.

To understand these results, one can observe the trends in the time varying coefficient plots of Figures B.1 and B.2 in the Appendix. These plots show how each time varying coefficient $\sum_{i=1}^{11} a_k(t)$ is changing over time, and which coefficients were found to be actually constant. Odds ratio plots are also produced for the categories in Figure 5.3. These plots were produced in the same manner as was done with the Italian smoking status varying coefficient model, where the constant term of the variable is added to the spline and then the exponential is taken to produce the odds ratio plot. The reference categories of the variables do not have odds ratio plots, however one can observe the coefficient plots to see any changes in the reference categories.

The odds ratio plots show that although a variable was found to be time varying in the tests of Table 5.1, not all the categories of that variable have coefficients that are varying. For example, the age category 65 and over shows a constant odds ratio trend, and the age category 50-64 was also constant until approximately the year 2005 where the odds of being obese began to drop slightly with time compared to the 18-34 reference age group. However, the age group 35-49 was found to have a slightly increasing odds ratio trend, indicating that the odds of being obese for this age category is increasing with time compared to the reference category. The reference age category of 18-34 does not have an odds ratio plot, but the coefficient plot of this category in Figure B.1 shows an increasing trend.

Other interesting trends for the socio-economic and demographic variables, include the male category in which there is a increasing odds ratio trend which begins from an odds ratio below one to an odds ratio above one compared to the female reference category. The divorced, widowed, or separated marital status category shows a decreasing odds ratio trend for being obese compared to the the reference married category, while the never married category showed a constant odds ratio trend. The reference category for martial status also has a decreasing coefficient trend as shown in Figure B.1. For the education variable, all the categories including the reference category have more or less a constant trend except for the education category of grade 11 or less. In Figure 5.3, we see a non-linear trend for this category that begins to decrease after the year 2000, but then becomes constant after the year 2005. The medium income level category has an increasing odds ratio trend which is above one, while the low income level trend is constant compared to the high income reference category (which has an increasing coefficient trend). The does not work category has a decreasing odds ratio trend that begins from from odds ratios slightly above one and then decreases to odds ratios below one, while the retired have a constant odds ratio trend compared to the working reference category. For the race variable, the black category has a relatively constant odds ratio trend as does the coefficient trend of the white reference race category. However, the other race category has a non-linear odds ratio trend that increases then decreases after the year 2005, and is always below an odds ratio of one. The Hispanic race category also has a non-linear odds ratio trend that decreases then increases after the year 2005 (odds ratios above one). Finally, for the region variable, the Midwest and South regions have increasing odds ratio trends compared to the West reference category.

For the remaining variables in the model, only having a health plan has a non-constant odds ratio trend as shown in Figure 5.3. The does not exercise category has a constant odds ratio trend, although the does perform exercise reference category has a decreasing coefficient trend in Figure B.2. Therefore, this indicates that not performing exercise has a steadily higher odds of being obese compared to performing exercise. Both categories of the general health variable have constant trends. For the health plan variable, the category ' have a health plan' has an increasing odds ratio trend that is below one, however the increase begins to level off around the year 2005 compared to the 'have no health plan' category. This indicates that having a health plan while being protective for being obese (odds ratios below one) compared to having no health plan, these odds ratios are moving towards one in the most recent years.



(a) age 35-49

(b) age 50-64

(c) age 65+

(d) male

(e) divorced/ widow/ separated

(f) never married

(g) some university

(h) high school

(i) grade 11 or less

(j) medium income

(k) low income

(l) retired

(m) does not work

(n) other

(o) hispanic

(p) black

(q) Northeast + DC

(r) Midwest

(s) South

(t) does not exercise

(u) have a health plan

(v) poor to fair

Figure 5.3: Odds ratio plots of the obesity time VCM for the USA.

As was conducted for the final varying coefficient model chosen from the Italian data analysis, a comparison of the residuals from the final model above and the residuals from the same model fit using simulated independent data was performed. As in the Italian case, the plots of these residuals versus the obesity fitted values produced very similar plots, therefore possibly indicating that the residuals are in fact independent.

5.4.2 Part II: Spatial VCM for Florida

For the spatially varying coefficient models, the modifying variable is space which is represented by `s` (replacing the modifying variable of time `t` found in the previous models). For the analysis space was represented by the longitude and latitude coordinates of the centroids of all 67 Florida counties. As was conducted with the time varying coefficients for the USA, a model is required that contains variables which are constant to represent the model in the null hypothesis in the tests. This was found to be the model similar to Model 5.1 which contains a spatially varying intercept as this was found to be significantly spatially varying when tested against the parametric logistic model. This model can be written as:

$$\text{logit}(OBS) = b_0 + \sum_{j=1}^{p} b_j Z_j + a_0(s). \tag{5.2}$$

Again we have that `OBS` is the outcome variable obesity, $a_0(s)$ is the spatially varying intercept, and Z_j are the covariates with their constant coefficients b_j. The same procedure was followed to build the varying coefficient model with the results summarized in Table 5.2, with p-values less than 0.05 indicated in bold.

As shown in the first part of Table 5.2, the variables age, sex, education, income, physical activity, smoke, and general health were found to have significantly spatially varying coefficients. The variables work, race, and health plan were found to have spatially constant coefficients, and were therefore not including in the building of the model in the second part of Table 5.2. The final spatial varying coefficient model found in the stepwise building of the model, was Model V which contains spatially varying coefficients for the variables age, income, education, sex, physical activity and smoke. The general health variable was found to have spatially constant

Table 5.2: Constructing the spatial VCM for obesity outcome in the Florida (Part II).

Model	Description	time (min)	p-value	H_0 used
Selection of variables that have varying coefficients				
Model age	LM + te(x,y):age	3.5	**0.017**	LM
Model sex	LM + te(x,y):sex	1.4	<**0.001**	LM
Model edu	LM + te(x,y):edu	3.7	**0.045**	LM
Model work	LM + te(x,y):work	2.2	0.247	LM
Model race	LM + te(x,y):race	5.0	0.060	LM
Model income	LM + te(x,y):income	2.3	**0.004**	LM
Model phy	LM + te(x,y):phy	1.3	**0.032**	LM
Model smoke	LM + te(x,y):smoke	1.4	**0.006**	LM
Model hplan	LM + te(x,y):hplan	1.4	0.444	LM
Model genhealth	LM + te(x,y):genhealth	1.5	**0.022**	LM
Finding the full varying coefficient mode				
Model I	Model age + te(x,y):income	11.3	**0.009**	Model age
Model II	Model I + te(x,y):edu	32.7	**0.014**	Model I
Model III	Model II + te(x,y):sex:	53.5	<**0.001**	Model II
Model IV	Model III + te(x,y):phy	94.4	**0.026**	Model III
Model V	Model IV + te(x,y):smoke	209.9	**0.003**	Model IV
Model VI	Model V + te(x,y):genhealth	340.1	0.166	Model V

Notes: LM - Model 5.2, te(x,y) - tensor product of coordinates, variable abbreviations found in Table B.2.

coefficients when other spatially varying coefficients were found in the model. The final model found can be written as:

$$\textbf{Model V} : \text{logit}(OBS) = b_0 + \sum_{j=1}^{p} b_j Z_j + a_0(s) + a_1(s)age + a_2(s)income \\ + a_3(s)phy + a_4(s)edu + a_5(s)sex + a_6(s)smoke,$$

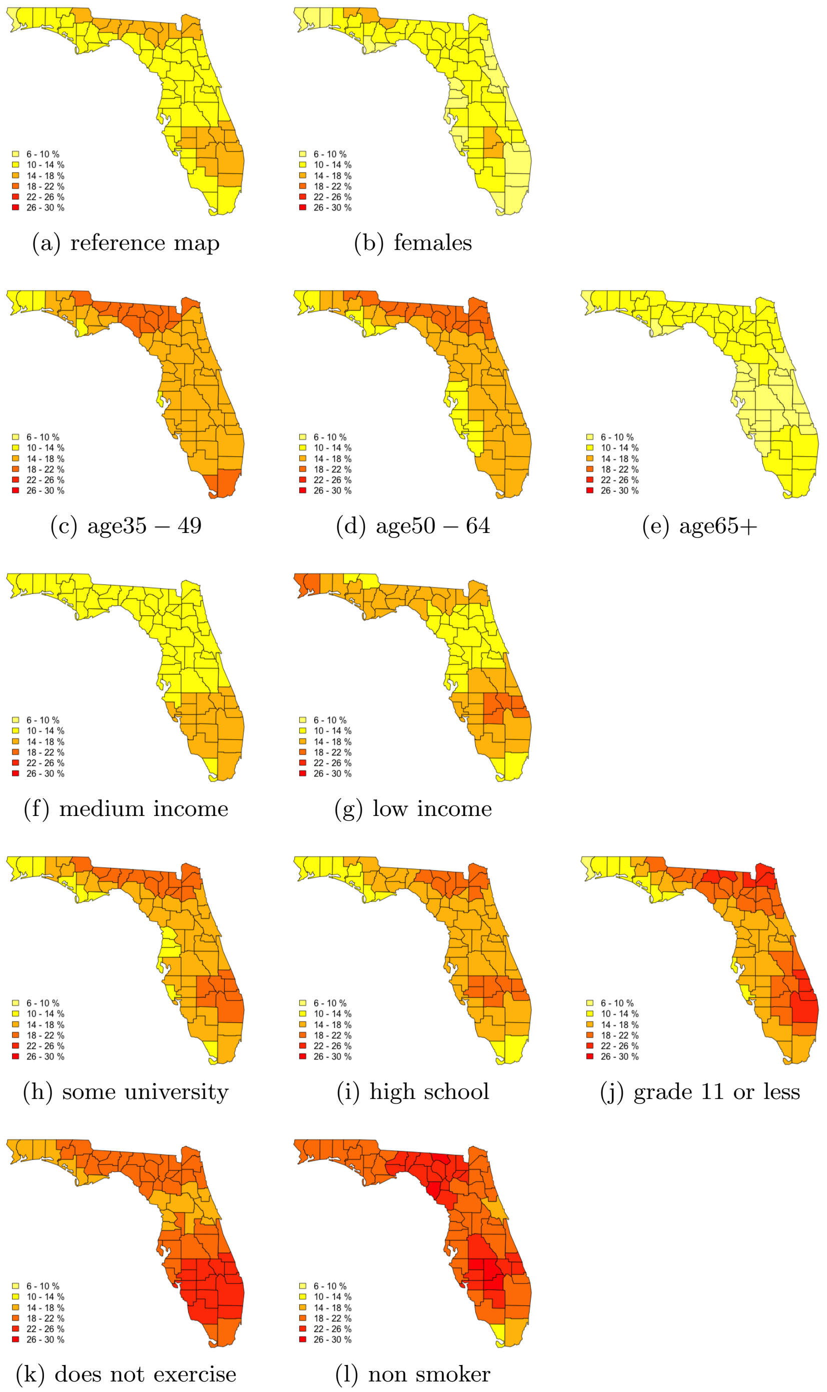
where $\sum_{i=1}^{6} a_k(s)$ are the spatially varying coefficients for the variables age, income, physical activity, education, sex and smoke.

To understand the spatial variations of these coefficients, maps are produced which show the probabilities of obesity for each category at each centroid while keeping the other variables constant at the reference level. These plots can be compared to the reference map (map the reference categories for all the variables), to compare between categories. In addition, each individual map for each category can show how the obesity probabilities and coefficients are varying by space. A flat map of the same probabilities indicates that the coefficients did not vary spatially.

One can clearly see when comparing each category to the reference map in Figure 5.4, which categories give higher probabilities of obesity and in which counties. For instance for the female category, we see that the probabilities are lower than the reference map (which contains the the male category) in almost all the counties. The highest probabilities were found for the non-smoker and does not exercise categories, which is expected as these categories had the highest odds ratios in the parametric results of Table B.4. In the parametric results the black race category also had a high odds ratio of being obese compared to whites; however, the race variable was found to have spatially constant coefficients according to the test as shown in Table 5.2, and therefore there is no map produced for the race categories.

The high obesity rates in the does not exercise and non-smoker maps also appear to be spatially varying as we see different probabilities of obesity in different counties. The highest probabilities of obesity (over 20%) appear in some counties in the North and South of Florida. More spatial variation can be seen in the grade 11 or less category, which contains a county in the Northwest (Escambia county) with a predicted probability of obesity as low as 6-10% for this category, while also having obesity probabilities of 22-26% in the Southeastern and Northeastern counties. Other categories have less spatial variation, such as the age category 65 and over which have predicted probabilities of obesity between 6-14% in all the counties. Low income appears to be more spatially varying than medium income as we see several different probabilities of obesity in different counties for those with low income, with higher probabilities of obesity in the Southeastern and Northwestern counties compared to the other counties. For medium income, the probability of obesity was divided in two regions, with the Northern counties all having probabilities between 10-14% , and the Southern counties between 14-18%.

To check whether there is any spatial auto-correlation in the final model, a variogram of the model residuals was produced as shown in Figure 5.5 using the `geoR` package in `R`. Variograms plot the variance of the pairwise residual differences from two spatial locations (the longitude and latitude in this case) versus the distance between these two spatial coordinates. A



(a) reference map

(b) females

(c) age35 − 49

(d) age50 − 64

(e) age65+

(f) medium income

(g) low income

(h) some university

(i) high school

(j) grade 11 or less

(k) does not exercise

(l) non smoker

Figure 5.4: Maps of the probabilities of obesity for each category of the variables found to be significantly spatially varying. The reference map shows the probabilities when the reference categories males, age 18 − 34, high income, university graduate and above education, does exercise, and current smoker are selected.

variogram which is flat (a constant horizontal straight line) indicates that the residuals are uncorrelated, while residuals which are spatially correlated will result in a variogram which increases sharply before plateauing (Wood, 2006a). The figure below shows a flat variogram, which indicates that the final spatial varying coefficient model found has no spatial auto-correlation.

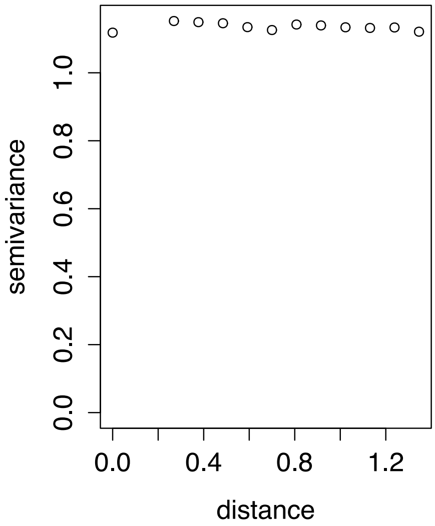


Figure 5.5: Variogram of residuals from the final spatial VCM.

In addition to testing for spatial auto-correlation, further analysis was conducted to check whether there are any boundary effects in the selected final spatial varying coefficient model. The results of this analysis are shown in Appendix B.3 and indicate that there are in fact no or very little boundary effects. This is expected as the method used is a very flexible model, however this is provided that enough knots are used in the model.

Chapter 6

Discussion and conclusions

The approach presented and discussed provides an alternative method for the study of health surveillance data for temporal and spatial trends. Using varying coefficient models provides a sufficiently dynamic approach for studying trends while taking into consideration the relationships between the changes in the effects of the independent variables in the model. The traditional methods used for the analysis of public health surveillance data (such as BRFS data), while important for studying the trends of outcomes as well as for understanding the impact of certain health policies on these outcomes, do not look at the time or spatial related effects. Therefore, in the traditional methods, a different question is being addressed compared to that from using the varying coefficient model approach. The application of the varying coefficient model approach to BRFS data produces results which can directly show the improving or deteriorating situations in certain subgroups of the population with respect to the outcome of interest. In both applications, using the smoking status outcome in Italy and the obesity outcome in the U.S.A., it was found that certain categories had increasing odds ratio trends while others decreased, therefore highlighting which subgroups in the population to target in health policy interventions. The same is true for the spatial analysis, in which one can see where and in which subgroups interventions need to be made. The next step would be constructing a spatial-temporal model to not only see where interventions are required, but how this is changing over time and in which subgroups of the population.

The comparison of the five techniques to fit the models in the Italian results chapter, was mainly conducted between estimation methods that required the selection of the number of knots to fit the smooth functions (such as in polynomial spline estimation), and methods which did not require this step since a high number of knots are added with a penalty to control the smoothness. Comparisons were also made between two functions used in the R software mgcv package, one which is designed for large datasets. The highest computation times were found when knot selection was required. When using P-splines or smoothing splines, knot selection was not required and computation times were improved especially when using the bam function designed for large datasets compared to the gam function. The selection of P-splines as the recommended method was due to the flexibility this method gives the user in selecting the degree of the spline and the degree of the penalty separately, therefore giving more control over the nature of the smoothing functions. Local regression was not used in the analysis and was not recommended, as this method does not allow for a separate level of smoothing for each coefficient. Therefore, this method assumes that all the coefficients have the same level of smoothness which may not be the case. A two-step local regression procedure was developed to provide two levels of smoothness (less smooth and more smooth) (Fan and Zhang, 2000), however this still does not have a separate level of smoothness for each coefficient as is the case with estimation using splines.

There are some limitations to using the varying coefficient model approach with non-parametric techniques. For one, when studying temporal trends, the larger the observation period the more likely the method used to construct the varying coefficient models will capture changes in the coefficients even if the change is not very large. This was evident in the U.S. data analysis in which for the 168 months of observation, the method required almost all the independent variables to have varying coefficients even though many of the coefficient plots showed relevantly constant trends. This is a problem from a computational point of view, as the more there are variables with varying coefficients required in the model, the greater the computation time. In fact, the final model for the U.S. time varying coefficient model could not be completed due to the large computation time and memory required. Increases in the period of observation also means that more knots are required in the model which also increases computation time. In fact, for the U.S.A. analysis three knots were taken for each year of observation, as placing the full number of knots (one for each month for 168 months) would have been computational infeasible. The recommendation is therefore to use a limited number of variables when the sample size becomes significantly large. Despite these issues, the recommended method of using the P-splines with the `bam` function, still allowed for fitting a varying coefficient model to a large dataset from the U.S. BRFSS (over two million observations) with eleven variables having significant varying coefficients in the final model, and with an acceptable computation time (1113.8 minutes or approximately 19 hours). With the growing importance of big data and the fact that BRFS systems (both in Italy and the U.S.A.) are still ongoing and therefore the sample size is continuously increasing, computation time becomes an important issue and it is important to have a method that is practical and feasible to use in future applications.

Another important limitation is regarding the method used for the spatial varying coefficient models. The method used with taking the tensor product of the variables which represent the longitude and latitude of the geographical centroids, do not take into account the boundary effects. Bayesian methods are usually used for studying spatial varying coefficient models and are able to account for boundary effects. However, the increased computation time and the assumptions required for selecting the prior distributions are some limitations of using Bayesian methods. The method used here for the spatial varying coefficient models is still a very flexible model which can take into account the spatial local variations (given that enough knots are provided), and give us a good idea of the changes in the probabilities of the outcome of interest across space. Even if boundary effects could be present in the model, their contribution in predicting the outcome probabilities should not be of greater importance than the contributions of the independent variables and their varying coefficients which are present in the model (as was found in further analysis in Appendix B.3). Another issue, is the use of geographical centroids for representing space. The use of population centroids may be more meaningful for the analysis of health data as it takes the coordinates of highest population locations.

Finally, as the outcome variables used in both applications were binary variables and the measures were not aggregated before analysis, model diagnostics can be difficult to obtain. The usual plots of fitted values versus the residuals are not meaningful in this case (Wood, 2006a). However, other techniques can be used in this case, for instance comparing the residuals of the final model to the residuals of the final model using simulated data (independent observations) (Wood, 2006a). This was conducted for both time varying coefficient models from the Italian and the U.S.A. data, and in both bases the results have shown that the residuals may in fact be independent. For the spatial varying coefficient model, a variogram was produced which showed that the residuals are spatially independent. Therefore, as expected this provides possible evidence that the data are independent since a new random sample is taken each month for these BRFS data.

In conclusion, the varying coefficient approach can be a useful tool for obtaining more insight from the BRFS data. It provides a novel technique for studying temporal and spatial trends in the public health field with this type of data, and can be easily adapted to any health outcome of interest and with simple to follow steps. The computation times are also reasonable, although with increasing periods of observation the computation time will naturally increase. Finally, the method can produce results in forms that are easily understood by health practitioners or health policy makers in order to influence decision makers for health policy interventions.

Appendix A

Appendices: Italian Data

A.1 Parametric Results: Italian Analysis

Table A.1: Descriptives of variables used for analysis using the Italian PASSI data.

Variable (abbreviation)		No.	Percent
Response Variable			
Smoking status (smoke)	Smoker	41654	28.1
	Non smoker	106612	71.9
Socio-economic and Demographic Variables			
Age (age)	18-29	26634	18.0
	30-39	31351	21.1
	40-49	35119	23.7
	50-59	28398	19.2
	60-69	26764	18.1
Sex (sex)	Female	75293	50.8
	Male	72973	49.2
Marital Status (mstatus)	Married	90918	61.3
	Single	45603	30.8
	Widowed/Div./Sep.*	11745	7.9
Education (edu)	University	19272	13.0

Continued on next page

Table A.1 – Continued from previous page

Variable (abbreviation)		**No.**	**Percent**
	High school	65347	44.1
	Middle school	46575	31.4
	Primary school or less	17072	11.5
Income level (inc)	Low	17912	12.1
	Medium	58448	39.4
	High	71906	48.5
Work status (work)	Yes	87124	58.8
	No	61142	41.2
Region (region)	North	74687	50.4
	Centre	36275	24.5
	South	37304	25.2
Citizenship (citizen)	Italian	142845	96.3
	Non-italian	5421	3.7
Lifestyle Variables and health risk variables			
Physical activity (phy)	Active	49051	33.1
	Partially active	56799	38.3
	Sedentary	42416	28.6
Alcohol consumption (alco)	Non-drinker	90575	61.1
	Low risk drinker	42428	28.6
	High risk drinker	15263	10.3
Depression (depress)	Depressed	9843	6.6
	Not depressed	138423	93.4

*Widowed, divorced, or separated.

Table A.2: Prevalence of smoking status by socio-demographic and health risk factors as well as unadjusted and adjusted odds ratios using the Italian PASSI data.

	Non smoker	**Smoker**	**Unadjusted**	**Adjusted**
Variable	**No. (%)**	**No. (%)**	**OR (95% C.I.)**	**OR (95%C.I.)**
Socio-economic and Demographic Variables				
Age				
18-29	17513 (65.8)	9121 (34.2)	Referent group	Referent group
30-39	21443 (53.7)	9908 (31.6)	0.89 (0.86-0.92)***	0.98 (0.94-1.01)
40-49	24819 (70.7)	10300 (29.3)	0.80 (0.77-0.82)***	0.87 (0.83-0.91)***
50-59	20840 (73.4)	7558 (26.6)	0.70 (0.67-0.72)***	0.78 (0.74-0.82)***
60-69	21997 (82.2)	4767 (17.8)	0.42 (0.40-0.43)***	0.51 (0.48-0.54)***
Sex				
Female	57410 (76.2)	17883 (23.8)	Referent group	Referent group
Male	49202 (67.4)	23771 (32.6)	1.60 (1.52-1.59)***	1.58 (1.54-1.62)***
Marital Status				
Married	69313 (76.2)	21605 (23.8)	Referent group	Referent group
Single	29672 (65.1)	15931 (34.9)	1.72 (1.68-1.77)***	1.47 (1.42-1.52)***
Widowed/Div.	7627 (64.9)	4118 (35.1)	1.73 (1.66-1.80)***	1.84 (1.77-1.92)***
Education				
University	14957 (77.6)	4315 (22.4)	Referent group	Referent group
High school	46828 (71.7)	18519 (28.3)	1.37 (1.32-1.42)***	1.35 (1.30-1.41)***
Middle school	31313 (67.2)	15262 (32.8)	1.69 (1.62-1.76)***	1.80 (1.72-1.87)***
Primary or less	13514 (79.2)	3558 (20.8)	0.91 (0.87-0.96)***	1.41 (1.33-1.49)***
Income				
Low	11280 (63.0)	6632 (37.0)	1.79 (1.73-1.85)***	1.75 (1.68-1.82)***
Medium	41205 (70.5)	17243 (29.5)	1.27 (1.24-1.31)***	1.28 (1.24-1.31)***
High	54127 (75.3)	17779 (24.7)	Referent group	Referent group
Work status				
Work	59878 (68.7)	27246 (31.3)	Referent group	Referent group
Do not work	46734 (76.4)	14408 (23.6)	0.68 (0.66-0.69)***	0.75 (0.73-0.77)***
Region				
North	54510 (73.0)	20177 (27.0)	Referent group	Referent group
Centre	25415 (70.1)	10860 (29.9)	1.15 (1.12-1.19)***	1.21 (1.17-1.24)***
South	26687 (71.5)	10617 (28.5)	1.10 (1.05-1.11)***	1.11 (1.07-1.14)***
Citizenship				
Italian	102846 (72.0)	39999 (28.0)	0.89 (0.84-0.94))***	0.98 (0.92-1.05)
Non-italian	3766 (69.5)	1655 (30.5)	Referent group	Referent group

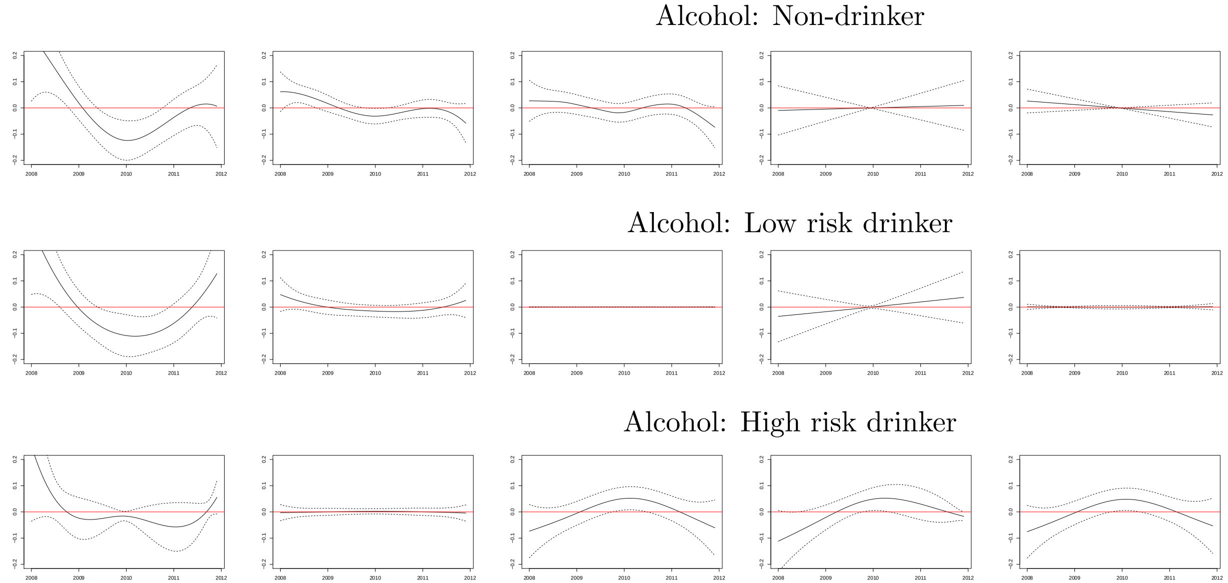
Continued on next page

Table A.2 – Continued from previous page

Variable	**No. (%)**	**No. (%)**	**OR** (95% **C.I.)**	**OR** (95%**C.I.)**
Lifestyle and Health Risk Variables				
Physical activity				
Active	34829 (71.0)	14222 (29.0)	Referent group	Referent group
Partially active	41835 (73.7)	14964 (26.3)	0.88 (0.85-0.90)***	0.95 (0.92-0.98)***
Sedentary	29948 (70.6)	12468 (29.4)	1.02 (0.99-1.05)	1.15 (1.12-1.19)***
Alcohol				
Non-drinker	68275 (75.4)	22300 (24.6)	Referent group	Referent group
Low risk drinker	29377 (69.2)	13051 (30.8)	1.36 (1.33-1.40)***	1.51 (1.47-1.55)***
High risk drinker	8960 (58.7)	6303 (41.3)	2.15 (2.08-2.23)***	2.31 (2.22-2.40)***
Depression				
Depressed	6335 (64.4)	3508 (35.6)	1.46 (1.39-1.52)***	1.43 (1.36-1.49)***
Not depressed	100277 (72.4)	38146 (27.6)	Referent group	Referent group

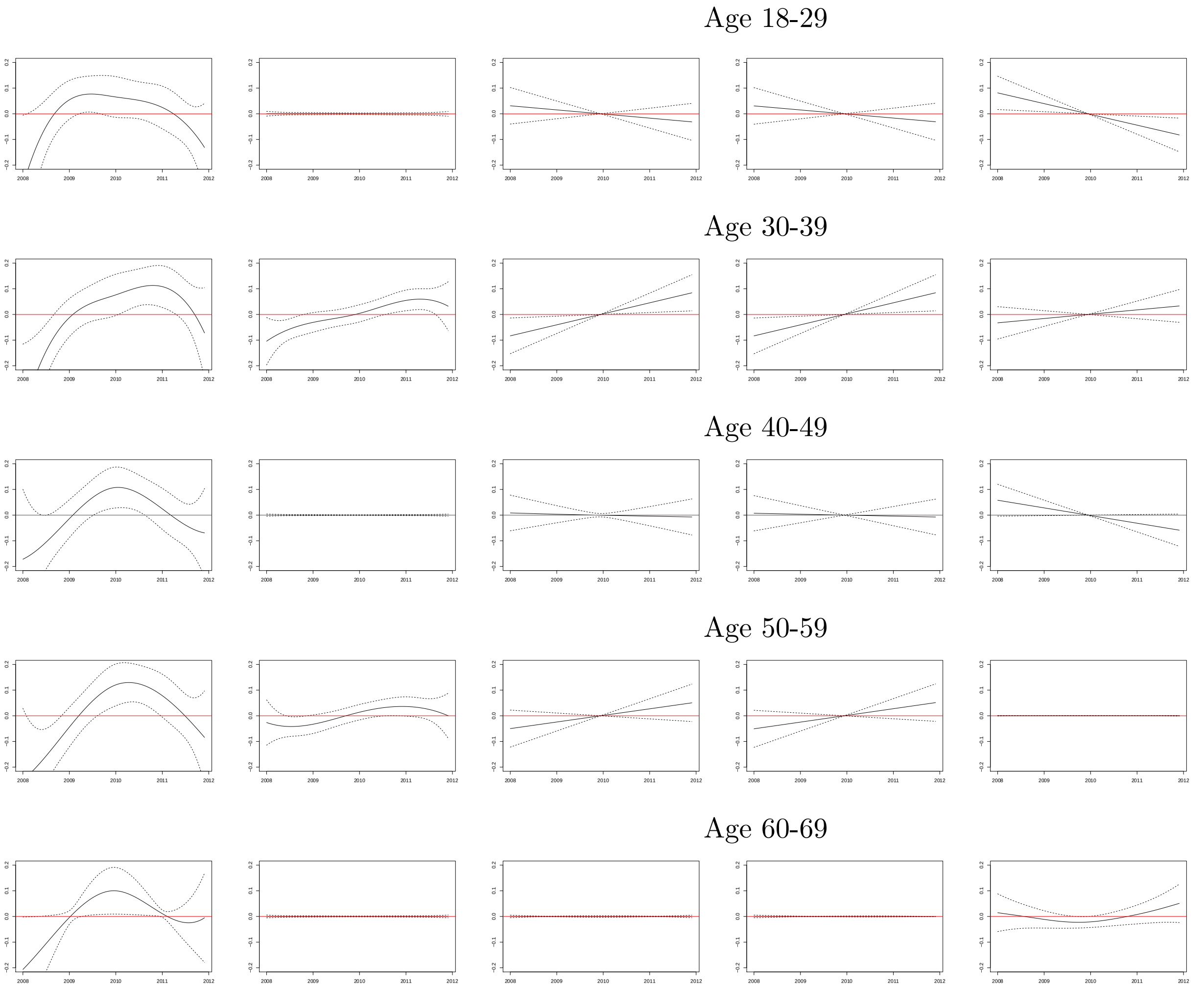
Significance codes: *** 0.001, ** 0.01, * 0.05. Each variable gave a p-value of < 0.001 in the χ^2 test of independence. OR - odds ratios with 95% confidence intervals.

A.2 Comparison of Time Varying Coefficient Plots for the Five Methods: Italian Analysis



(a) Method I (b) Method II (c) Method III (d) Method IV (e) Method V

Figure A.1: Comparison of alcohol categories coefficient plots from the smoking status time VCM between the five methods.



Age 18-29

Age 30-39

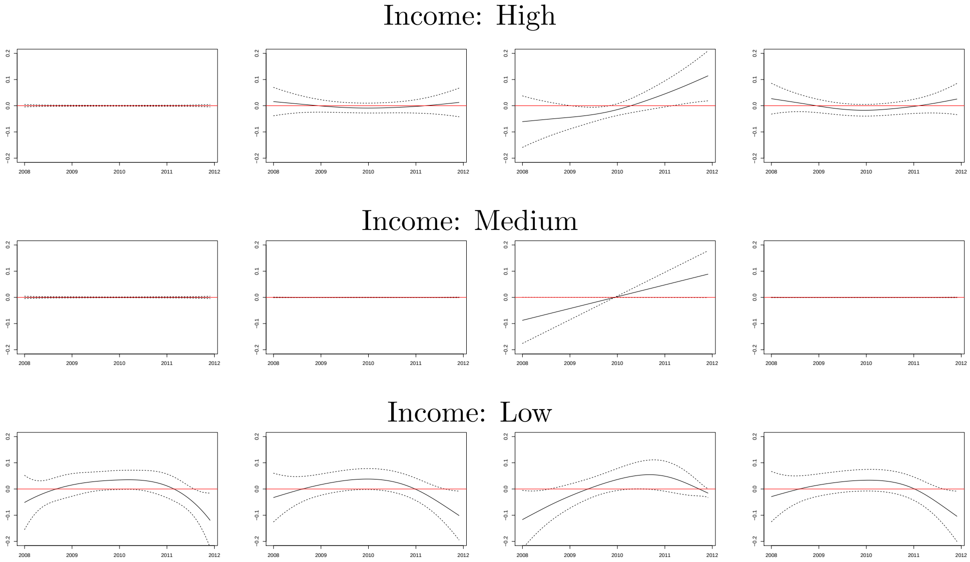
Age 40-49

Age 50-59

Age 60-69

(a) Method I (b) Method II (c) Method III (d) Method IV (e) Method V

Figure A.2: Comparison of age categories coefficient plots from the smoking status time VCM between the five methods.

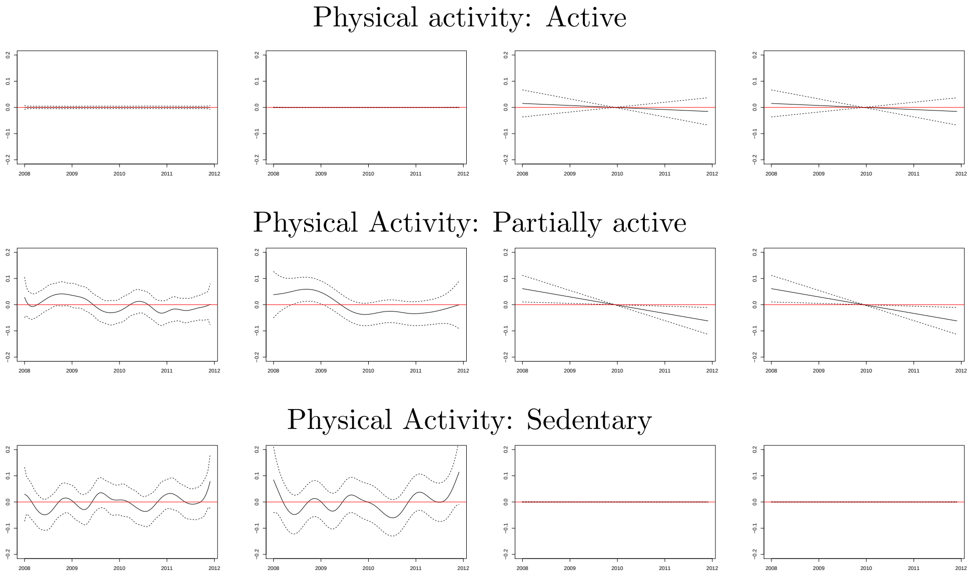

Income: High

Income: Medium

Income: Low

(a) Method II (b) Method III (c) Method IV (d) Method V

Figure A.3: Comparison of income categories coefficient plots from the smoking status time VCM between four methods (Method I did not include an income variable).


Physical activity: Active

Physical Activity: Partially active

Physical Activity: Sedentary

(a) Method II (b) Method III (c) Method IV (d) Method V

Figure A.4: Comparison of physical activity coefficient plots from the smoking status time VCM between four methods (Method I did not include a physical activity variable).

Appendix B

Appendices: U.S.A. Data

B.1 Parametric Results: USA Analysis

Table B.1: Descriptives of variables used for obesity time VCM for the USA from 1993-2009 (Part I).

Variable (abbreviation)		**No.**	**Percent**
Response Variable			
Obesity (obese)	Obese	487111	23.6
	Not obese	1578578	76.4
Socio-economic and Demographic Variables			
Age (age)	18-34	416991	20.2
	35-49	591648	28.6
	50-64	554967	26.9
	65+	502083	24.3
Sex (sex)	Female	1232820	59.7
	Male	832869	40.3
Marital Status (mstatus)	Married/couple	1196085	57.9
	Widowed/Div./Sep.*	577875	28.0
	Never married	291729	14.1
Education (edu)	University+	650556	31.5
	Some university	558948	27.1
	High school	643981	31.2
	< grade 11	212204	10.3

Continued on next page

Table B.1 – Continued from previous page

Variable (abbreviation)		No.	Percent
Income level (income)	Low	142280	6.9
	Medium	1186708	57.4
	High	736701	35.7
Work status (work)	Works	1197726	58.0
	Retired	470179	22.8
	Does not work	397784	19.2
Region (region)	West	459546	22.2
	North East + DC	498748	24.1
	Mid West	478134	23.2
	South	629261	30.5
Race (race)	White	1691971	81.9
	Black	152806	7.4
	Hispanic	109351	5.3
	Other	111561	5.4
Lifestyle and health risk variables			
Physical exercise (phy)	Exercises	1526432	73.9
	Does not exercise	539257	26.1
Fruit and	5+ servings	510563	24.7
vegetable consumption (fruit)	< 5 servings	1555126	75.3
Smoking status (smoke)	Current smoker	423058	20.5
	Non smoker	1642631	79.5
Other variables			
Health care (hplan)	Has health plan	1827584	88.5
access	Does not have health plan	238105	11.5
General health (genhealth)	Good to excellent	1717326	83.1
status	Poor to fair	1717326	16.9

*Widowed, divorced, or separated.

Table B.2: Descriptives of variables used for the obesity spatial VCM for Florida state in 2010 (Part II).

Variable (abbreviation)		No.	Percent
Response Variable			
Obesity (obese)	Obese	9431	29.4
	Not obese	22679	70.6
Socio-economic and Demographic Variables			
Age (age)	18-34	2846	8.9
	35-49	5782	18.0
	50-64	10237	31.9
	65+	13245	41.2
Sex (sex)	Female	19849	61.8
	Male	12261	38.2
Marital Status (mstatus)	Married/couple	17983	56.0
	Widowed/Div./Sep.*	11387	35.5
	Never married	2740	8.5
Education (edu)	University+	9392	29.2
	Some university	9132	28.4
	High school	10422	32.5
	< grade 11	3164	9.9
Income level (income)	Low	4034	12.6
	Medium	19758	61.5
	High	8318	25.9
Work status (work)	Works	12263	38.2
	Retired	11714	36.5
	Does not work	8133	25.3
Race (race)	White	26861	83.7
	Black	2432	7.6
	Hispanic	1717	5.3
	Other	1100	3.4

Continued on next page

Table B.2 – Continued from previous page

Variable (abbreviation)		**No.**	**Percent**
Lifestyle and health risk variables			
Physical exercise (phy)	Exercises	22808	71.0
	Does not exercise	9302	29.0
Smoking status (smoke)	Current smoker	5854	18.2
	Non smoker	26256	81.8
Other variables			
Health care (hplan)	Has health plan	27832	86.7
access	Does not have health plan	4278	13.3
General health (genhealth)	Good to excellent	24510	76.3
status	Poor to fair	7600	23.7

*Widowed, divorced, or separated.

Table B.3: Prevalence of obesity by socio-demographic and health risk factors as well as unadjusted and adjusted odds ratios for the USA between 1993-2009 (Part I).

	Not obese	**Obese**	**Unadjusted**	**Adjusted**
Variable	**No. (%)**	**No. (%)**	**OR** (95% **CI**)	**OR** (95% **CI**)
Age				
18-34	340758 (81.7)	76233 (18.3)	Referent group	Referent group
35-49	446888 (75.5)	144760 (24.5)	1.45 (1.43-1.46)***	1.48 (1.47-1.50)***
50-64	394038 (71.0)	160929 (29.0)	1.82 (1.81-1.84)***	1.81 (1.79-1.83)***
65+	396894 (79.0)	105189 (21.0)	1.18 (1.17-1.20)***	1.03 (1.02-1.04)***
Sex				
Female	943740 (76.6)	289080 (23.4)	Referent group	Referent group
Male	634838 (76.2)	198031 (23.8)	1.02 (1.01-1.03)***	1.07 (1.07-1.08)***
Marital Status				
Married/couple	917082 (76.7)	279003 (23.3)	Referent group	Referent group
Widowed/Div.	435676 (75.4)	142199 (24.6)	1.07 (1.06-1.08)***	0.97 (0.96-0.98)***
Never married	225820 (77.4)	65909 (22.6)	0.96 (0.95-0.97)***	1.02 (1.01-1.03)***
Education				
University+	529781 (81.4)	120775 (18.6)	Referent group	Referent group
Some university	420410 (75.2)	138538 (24.8)	1.45 (1.43-1.46)***	1.38 (1.37-1.39)***
High school	478012 (74.2)	165969 (25.8)	1.52 (1.51-1.54)***	1.36 (1.34-1.37)***
<grade 11	150375 (70.9)	61829 (29.1)	1.80 (1.78-1.82)***	1.35 (1.34-1.37)***
Income				
High	583771 (79.2)	152930 (20.8)	Referent group	Referent group
Medium	898974 (75.8)	287734 (24.2)	1.22 (1.21-1.23)***	1.11 (1.10-1.12)***
Low	95833 (67.4)	46447 (32.6)	1.85 (1.83-1.87)***	1.43 (1.41-1.45)***
Work status				
Works	921920 (77.0)	275806 (23.0)	Referent group	Referent group
Retired	367471 (78.2)	102708 (21.8)	0.93 (0.93-0.94)***	0.90 (0.89-0.91)***
Does not work	289187 (72.7)	108597 (27.3)	1.26 (1.25-1.27)***	0.99 (0.98-0.99)**
Region				
West	360699 (78.5)	98847 (21.5)	Referent group	Referent group
NorthE+DC	389526 (78.1)	109222 (21.9)	1.02 (1.01-1.03)***	0.98 (0.97-0.99)***
Midwest	358436 (75.0)	119698 (25.0)	1.22 (1.21-1.23)***	1.17 (1.16-1.18)***
South	469917 (74.7)	159344 (25.3)	1.24 (1.23-1.25)***	1.07 (1.06-1.08)***
Race				
White	1313672 (77.6)	378299 (22.4)	Referent group	Referent group
Black	97857 (64.0)	54949 (36.0)	1.95 (1.93-1.97)***	1.71 (1.69-1.73)***
Hispanic	80638 (73.7)	28713 (26.3)	1.24 (1.22-1.25) ***	1.06 (1.04-1.07)***
Other	86411 (77.5)	25150 (22.5)	1.01 (0.99-1.03)	0.99 (0.97-1.00)
Physical activity				
Exercises	1208995 (79.2)	317437 (20.8)	Referent group	Referent group
No exercise	369583 (68.5)	169674 (31.5)	1.75 (1.74-1.76)***	1.48 (1.47-1.49)***

Continued on next page

Table B.3 – Continued from previous page

Variable	**No. (%)**	**No. (%)**	**OR** (95% **CI**)	**OR** (95% **CI**)
Fruit & vegetable consumption				
5+	405500 (79.4)	105063 (20.6)	Referent group	Referent group
< 5	1173078 (75.4)	382048 (24.6)	1.26 (1.25-1.27)***	1.16 (1.15-1.17)***
Smoking status				
Non smoker	1240742 (75.5)	401889 (24.5)	Referent group	Referent group
Current smoker	337836 (79.9)	85222 (20.1)	0.78 (0.77-0.79)***	0.61 (0.60-0.61)***
Health care access				
No health plan	177042 (74.4)	61063 (25.6)	Referent group	Referent group
Has health plan	1401536 (76.7)	426048 (23.3)	0.88 (0.87-0.89)***	1.04 (1.03-1.05)***
General health status				
Good/excellent	1356546 (79.0)	360780 (21.0)	Referent group	Referent group
Poor/fair	222032 (63.7)	126331 (36.3)	2.14 (2.12-2.16)***	1.86 (1.84-1.87)***

Significance codes: *** 0.001, ** 0.01, * 0.05. Each variable gave a p-value of < 0.001 in the χ^2 test of independence. OR - odds ratios with 95% confidence intervals.

Table B.4: Prevalence of obesity by socio-demographic and health risk factors as well as unadjusted and adjusted odds ratios for Florida in 2010 (Part II).

Variable	**Not obese** **No. (%)**	**Obese** **No. (%)**	**Unadjusted** **OR (95% CI)**	**Adjusted** **OR (95% CI)**
Age				
18-34	2059 (72.3)	787 (27.7)	Referent group	Referent group
35-49	3863 (66.8)	1919 (33.2)	1.30 (1.18-1.44)***	1.31 (1.18-1.45)***
50-64	6774 (66.2)	3463 (33.8)	1.34 (1.22-1.47)***	1.29 (1.16-1.43)***
65+	9983 (75.4)	3262 (24.6)	0.85 (0.78-0.94)***	0.79 (0.70-0.89)***
Sex				
Female	14178 (71.4)	5671 (28.6)	0.90 (0.86-0.95)***	0.83 (0.79-0.88)***
Male	8501 (69.3)	3760 (30.7)	Referent group	Referent group
Marital Status				
Married/couple	12753 (70.9)	5230 (29.1)	Referent group	Referent group
Widowed/Div.	8082 (71.0)	3305 (29.0)	1.00 (0.95-1.05)	0.96 (0.91-1.02)
Never married	1844 (67.3)	896 (32.7)	1.19 (1.09-1.29)***	1.00 (0.91-1.10)
Education				
University+	7175 (76.4)	2217 (23.6)	Referent group	Referent group
Some university	6386 (69.9)	2746 (30.1)	1.39 (1.30-1.49)***	1.36 (1.27-1.46)***
High school	7147 (68.6)	3275 (31.4)	1.48 (1.39-1.58)***	1.37 (1.27-1.47)***
< grade 11	1971 (62.3)	1193 (37.7)	1.96 (1.80-2.12)***	1.56 (1.41-1.72)***
Income				
High	6236 (75.0)	2082 (25.0)	Referent group	Referent group
Medium	13946 (70.6)	5812 (29.4)	1.25 (1.18-1.32)***	1.02 (0.96-1.09)
Low	2497 (61.9)	1537 (38.1)	1.84 (1.70-2.00)***	1.17 (1.06-1.30)**
Work status				
Works	8544 (69.7)	3719 (30.3)	Referent group	Referent group
Retired	8781 (75.0)	2933 (25.0)	0.77 (0.72-0.81)***	0.91 (0.84-0.98)*
Does not work	5354 (65.8)	2779 (34.2)	1.19 (1.12-1.27)***	0.97 (0.91-1.04)
Race				
White	19447 (72.4)	7414 (27.6)	Referent group	Referent group
Black	1290 (53.0)	1142 (47.0)	2.32 (2.13-2.52)***	1.95 (1.78-2.13)***
Hispanic	1157 (67.4)	560 (32.6)	1.27 (1.14-1.41) ***	1.02 (0.91-1.14)
Other	785 (71.4)	315 (28.6)	1.05 (0.92-1.20)	0.92 (0.80-1.06)
Physical activity				
Exercises	17002 (74.5)	5806 (25.5)	Referent group	Referent group
No exercise	5677 (61.0)	3625 (39.0)	1.87 (1.78-1.97)***	1.67 (1.58-1.77)***
Smoking status				
Non smoker	18258 (69.5)	7998 (30.5)	1.35 (1.27-1.44)***	1.88 (1.75-2.02)***
Current smoker	4421 (75.5)	1433 (24.5)	Referent group	Referent group

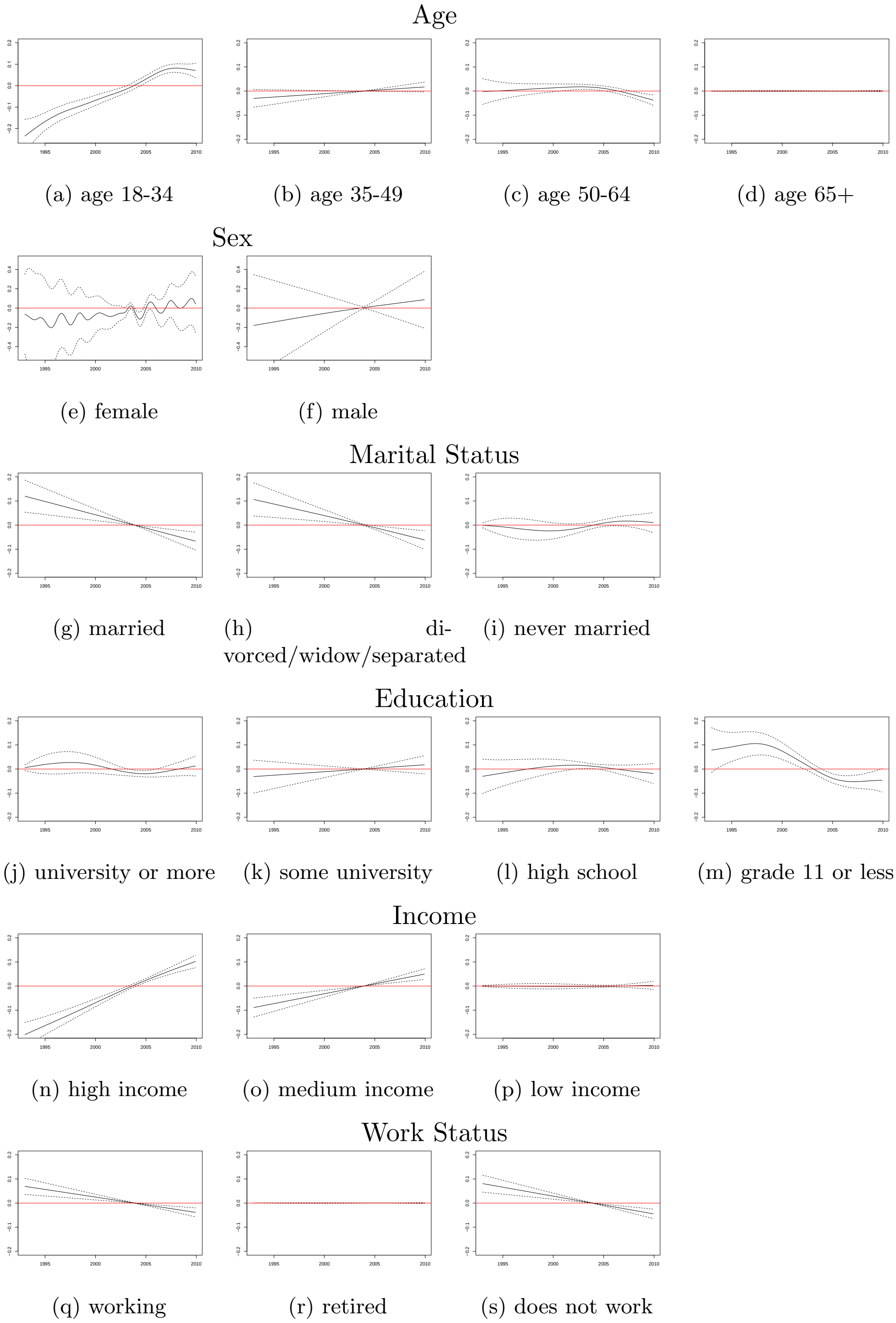
Continued on next page

Table B.4 – Continued from previous page

Variable	**No.** (%)	**No.** (%)	**OR** (95% **CI**)	**OR** (95% **CI**)
Health care access				
No health plan	2865 (67.0)	1413 (33.0)	Referent group	Referent group
Has health plan	19814 (71.2)	8018 (28.8)	0.82 (0.77-0.88)***	1.07 (0.99-1.16)
General health status				
Good/excellent	18133 (74.0)	6377 (26.0)	Referent group	Referent group
Poor/fair	4546 (59.8)	3054 (40.2)	1.91 (1.81-2.02)***	1.67 (1.57-1.77)***

Significance codes: *** 0.001, ** 0.01, * 0.05. Each variable gave a p-value of < 0.001 in the χ^2 test of independence. OR - odds ratios with 95% confidence intervals.

B.2 Time Varying Coefficient Plots: USA analysis


Age

(a) age 18-34 (b) age 35-49 (c) age 50-64 (d) age 65+

Sex

(e) female (f) male

Marital Status

(g) married (h) divorced/widow/separated (i) never married

Education

(j) university or more (k) some university (l) high school (m) grade 11 or less

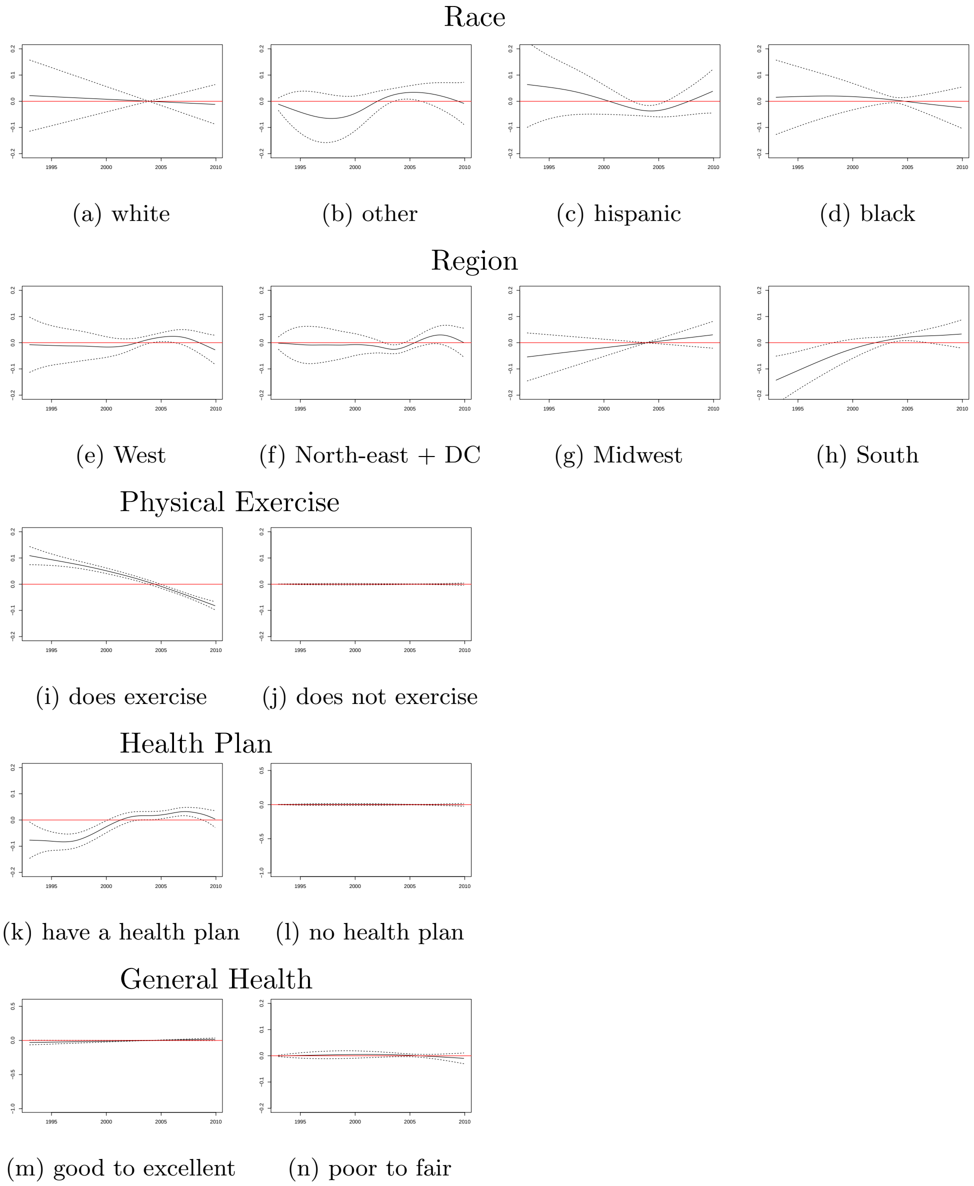
Income

(n) high income (o) medium income (p) low income

Work Status

(q) working (r) retired (s) does not work

Figure B.1: Time varying coefficient plots for age, sex, marital status, education, income, and work variables of the obesity VCM (Part I analysis).


Race

(a) white (b) other (c) hispanic (d) black

Region

(e) West (f) North-east + DC (g) Midwest (h) South

Physical Exercise

(i) does exercise (j) does not exercise

Health Plan

(k) have a health plan (l) no health plan

General Health

(m) good to excellent (n) poor to fair

Figure B.2: Time varying coefficient plots for race, region, physical exercise, health plan, and general health variables of the obesity VCM (Part I analysis).

B.3 Boundary effect analysis for the spatial VCM

To check whether there are boundary effects in the spatial varying coefficient model, two methods were used. Method I tests whether the probabilities of the border counties (i.e. counties that are on the border) predicted from the final spatial varying coefficient model,

$$\begin{aligned}\text{logit}(OBS) = b_0 + \sum_{j=1}^{p} b_j Z_j + a_0(s) + a_1(s)age + a_2(s)income \qquad \text{(B.1)}\\ + a_3(s)phy + a_4(s)edu + a_5(s)sex + a_6(s)smoke,\end{aligned}$$

which was fit with all the data, lies in the 95% confidence interval of the probabilities of the border counties from the same model fit using only inland county data (i.e. counties that are not on the border). For Method II, a new model was fit which includes a dummy variable for the counties (border or inland county) in the same final spatial varying coefficient model. In addition, interaction terms with this dummy variable and the variables which had spatially varying coefficients were added. This model can be written as:

$$\begin{aligned}\text{logit}(OBS) = b_0 + \sum_{j=1}^{p} b_j Z_j + a_0(s) + a_1(s)age + a_2(s)income \qquad \text{(B.2)}\\ + a_3(s)phy + a_4(s)edu + a_5(s)sex + a_6(s)smoke\\ + \alpha_1 cnty + \alpha_2 cnty{:}age + \alpha_3 cnty{:}income + \alpha_4 cnty{:}phy\\ + \alpha_5 cnty{:}edu + \alpha_6 cnty{:}sex + \alpha_7 cnty{:}smoke,\end{aligned}$$

where `cnty` is the county dummy variable. The results of this model show that the p-values of this dummy variable and all the interaction terms were not significant except for the interaction with female category and nonsmoker category. The following step was to check the probabilities, in which we see whether the probabilities of the border counties predicted from Model B.1 lies in the 95% confidence interval of the probabilities of the border counties from Model B.2. The results of these coverage probabilities for both methods are found in Table B.5.

The results shown from Method I have reasonably high coverage probabilities except for the two age categories 35-49 and 50-64. The data used to fit the model for this method uses only inland county data, which pro-

Table B.5: Checking boundary effects of final spatial VCM, with reported coverage probability for each category with spatially varying coefficients using two Methods.

		Method I	**Method II**
Variable	**Categories**	**Cov. prob. (%)**	**Cov. prob. (%)**
Reference categories		90.7	100.0
Sex	female	93.0	97.7
Age	35-49	58.1	100.0
	50-64	51.1	97.7
	65+	88.4	100.0
Education	university	93.0	100.0
	high school	100.0	100.0
	< grade 11	100.0	100.0
Income	medium	97.7	100.0
	low	95.3	100.0
Physical Activity	no exercise	90.7	100.0
Smoking status	non-smokers	86.1	100.0

duced a much smaller sample size of 11,388 observations compared to the total sample size of 32,110 and the border county data of 20,722. Therefore, there may be some lack of precision of the estimates due to this reduced sample size. However, we still see that most of the categories have high coverage probabilities. In Method II, we see very high coverage probabilities, in fact only two categories did not give a 100% coverage probability. This indicates that the county dummy variable and their interaction terms are not changing the probability estimates and are therefore not important predictors in the model, which was also indicated by the non-significant p-values of the these variables in the model. These results indicate that there is little boundary effect in the final spatial varying coefficient model shown in Model B.1.

Bibliography

Ahluwalia, I.B., Mack, K.A., and Mokdad, A. 2005. Report from the CDC. Changes in selected chronic disease-related risks and health conditions for nonpregnant women 18-44 years old BRFSS. *Journal of Women's Health*, **14**(5), 382–386.

Ashford, A., Kiros, G.E., and López, I.A. 2010. Trends and Correlates of Breast Cancer Screening among Florida Women: Analysis of 2001 and 2008 BRFSS Data. *Florida Public Health Review*, **7**, 17–25.

Assunçao, R.M. 2003. Space varying coefficient models for small area data. *Environmetrics*, **14**(5), 453–473.

Augustin, N.H., Musio, M., von Wilpert, K., Kublin, E., Wood, S.N., and Schumacher, M. 2009. Modeling spatiotemporal forest health monitoring data. *Journal of the American Statistical Association*, **104**(487), 899–911.

Augustin, N.H., Trenkel, V.M., Wood, S.N., and Lorance, P. 2013. Space-time modelling of blue ling for fisheries stock management. *Environmetrics*, **24**(2), 109–119.

Baldissera, S., Campostrini, S., Binkin, N., Minardi, V., Minelli, G., Ferrante, G., Salmaso, S., and the PASSI Coordinating Group. 2011. Peer Reviewed: Features and Initial Assessment of the Italian Behavioral Risk Factor Surveillance System (PASSI), 2007-2008. *Preventing Chronic Disease*, **8**(1).

Biller, C., and Fahrmeir, L. 2001. Bayesian varying-coefficient models using adaptive regression splines. *Statistical Modelling*, **1**(3), 195–211.

Binkin, N., Gigantesco, A., Ferrante, G., and Baldissera, S. 2010. Depressive symptoms among adults 18–69 years in Italy: results from the Italian behavioural risk factor surveillance system, 2007. *International Journal of Public Health*, **55**(5), 479–488.

Brezger, A., and Lang, S. 2006. Generalized structured additive regression based on Bayesian P-splines. *Computational Statistics & Data Analysis*, **50**(4), 967–991.

Brunsdon, C., Fotheringham, A.S., and Charlton, M.E. 1996. Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis*, **28**(4), 281–298.

Cai, Z., Fan, J., and Li, R. 1999. *Generalized Varying-coefficient models*. Available at: http://escholarship.org/uc/item/7n0494s4 , Department of Statistics, UCLA.

Cai, Z., Fan, J., and Li, R. 2000. Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association*, 888–902.

Campostrini, S., and McQueen, D. 2011. White paper on surveillance and health promotion. *International Union for Health Promotion and Education.*

Campostrini, S., Holtzman, D., Mcqueen, D.V., and Boaretto, E. 2006. Evaluating the effectiveness of health promotion policy: changes in the law on drinking and driving in California. *Health Promotion International*, **21**(2), 130–135.

Campostrini, S., McQueen, D.V., and Abel, T. 2011. Social determinants and surveillance in the new Millennium. *International Journal of Public Health*, **56**(4), 357–358.

Center for Disease Control and Prevention (CDC). 2014. *About the Behavioral Risk Factor Surveillance System (BRFSS).* Available at: http://www.cdc.gov/brfss/about/about_brfss.htm. Retrieved on: 5 January 2014.

Cheng, M.Y., Zhang, W., and Chen, L.H. 2009. Statistical estimation in generalized multiparameter likelihood models. *Journal of the American Statistical Association*, **104**(487), 1179–1191.

Chiang, C.T., Rice, J.A., and Wu, C.O. 2001. Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *Journal of the American Statistical Association*, **96**(454), 605–619.

Cleveland, W.S., Grosse, E., and Shyu, W.M. 1992. Local regression models. *Statistical Models in S*, 309–376.

Congdon, P. 2003. Modelling spatially varying impacts of socioeconomic predictors on mortality outcomes. *Journal of Geographical Systems*, **5**(2), 161–184.

Deaton, A. 1985. Panel data from time series of cross-sections. *Journal of Econometrics*, **30**(1-2), 109–126.

Declich, S., and Carter, A.O. 1994. Public health surveillance: historical origins, methods and evaluation. *Bulletin of the World Health Organization*, **72**(2), 285.

Diggle, P. 1990. *Time series: a Biostatistical Introduction.* Vol. 5. Oxford University Press.

Diggle, P. 2002. *Analysis of Longitudinal Data.* Vol. 25. Oxford University Press, USA.

Eilers, P.H.C., and Marx, B.D. 1996. Flexible smoothing with B-splines and penalties. *Statistical Science*, 89–102.

Eilers, P.H.C., and Marx, B.D. 2002. Generalized linear additive smooth structures. *Journal of Computational and Graphical Statistics*, **11**(4), 758–783.

Eilers, P.H.C., Heim, S., and Marx, B.D. 2005. *Varying coefficient tensor models for brain imaging.* Tech. rept. Discussion paper/Sonderforschungsbereich 386 der Ludwig-Maximilians-Universität München.

Fahimi, M., Link, M., Schwartz, D.A., Levy, P., and Mokdad, A. 2008. Tracking chronic disease and risk behavior prevalence as survey participation declines: statistics from the Behavioral Risk Factor Surveillance System and other national surveys. *Preventing Chronic Disease*, **5**(3), A80.

Fahrmeir, L., and Lang, S. 2001. Bayesian inference for generalized additive mixed models based on Markov random field priors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **50**(2), 201–220.

Fahrmeir, L., Lang, S., Wolff, J., and Bender, S. 2000. Semiparametric Bayesian time-space analysis of unemployment duration. **211**(sfb386). Available at: http://epub.ub.uni-muenchen.de/1601/.

Fahrmeir, L., Kneib, T., and Lang, S. 2004. Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica*, **14**(3), 731–762.

Fan, A.Z. 2013. Trends in Cigarette Smoking Rates and Quit Attempts Among Adults With and Without Diagnosed Diabetes, United States, 2001–2010. *Preventing Chronic Disease*, **10**.

Fan, J., and Huang, T. 2005. Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli*, **11**(6), 1031–1057.

Fan, J., and Zhang, J.T. 2000. Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **62**(2), 303–322.

Fan, J., and Zhang, W. 1999. Statistical estimation in varying coefficient models. *The Annals of Statistics*, **27**(5), 1491–1518.

Fan, J., and Zhang, W. 2008. Statistical methods with varying coefficient models. *Statistics and its Interface*, **1**(1), 179.

Flegal, K.M., Carroll, M.D., Ogden, C.L., and Johnson, C.L. 2002. Prevalence and trends in obesity among US adults, 1999-2000. *The Journal of the American Medical Association*, **288**(14), 1723–1727.

Gamerman, D., Moreira, A.R.B., and Rue, H. 2003. Space-varying regression models: specifications and simulation. *Computational Statistics & Data Analysis*, **42**(3), 513–533.

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

Hastie, T., Tibshirani, R., and Friedman, J. 2009. *The Elements of Statistical Learning*. 2nd edn. Springer New York.

Heim, S., Fahrmeir, L., Eilers, P.H.C., and Marx, B.D. 2007. 3D space-varying coefficient models with application to diffusion tensor imaging. *Computational Statistics & Data Analysis*, **51**(12), 6212–6228.

Hoover, D.R., Rice, J.A., Wu, C.O., and Yang, L.P. 1998. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, **85**(4), 809–822.

Hu, S.S., Balluz, L., Battaglia, M.P., and Frankel, M.R. 2011. Improving public health surveillance using a dual-frame survey of landline and cell phone numbers. *American Journal of Epidemiology*, **173**(6), 703–711.

Huang, J.Z., and Shen, H. 2004. Functional Coefficient Regression Models for Non-linear Time Series: A Polynomial Spline Approach. *Scandinavian Journal of Statistics*, **31**(4), 515–534.

Huang, J.Z., Wu, C.O., and Zhou, L. 2002. Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, **89**(1), 111–128.

Huang, J.Z., Wu, C.O., and Zhou, L. 2004. Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica*, **14**(3), 763–788.

Jia, H., and Lubetkin, E.I. 2009. Time trends and seasonal patterns of health-related quality of life among US adults. *Public Health Reports*, **124**(5), 692.

Kauermann, G., and Tutz, G. 1999. On model diagnostics using varying coefficient models. *Biometrika*, **86**(1), 119–128.

Lee, L.M., and Thacker, S.B. 2010. *Principles and Practice of Public Health Surveillance*. Oxford University Press, USA.

Li, R., and Liang, H. 2008. Variable selection in semiparametric regression modeling. *Annals of Statistics*, **36**(1), 261.

Ma, Z., Kuller, L.H., Fisher, M.A., and Ostroff, S.M. 2013. Peer Reviewed: Use of Interrupted Time-Series Method to Evaluate the Impact of Cigarette Excise Tax Increases in Pennsylvania, 2000–2009. *Preventing Chronic Disease*, **10**.

Marx, B.D. 2010. P-spline varying coefficient models for complex data. *Statistical Modelling and Regression Structures*, 19–43.

Minardi, V., Campostrini, S., Carrozzi, G., Minelli, G., and Salmaso, S. 2011. Social determinants effects from the Italian risk factor surveillance system PASSI. *International Journal of Public Health*, **56**(4), 359–366.

Mokdad, A.H. 2009. The behavioral risk factors surveillance system: Past, present, and future. *Annual Review of Public Health*, **30**, 43–54.

Muller, W.G. (ed). 2007. *Collecting Spatial Data: Optimum Design of Experiments for Random Fields*. Springer. Chap. Fundamentals of Spatial Statistics, pages 11–37.

NCEAS. 2013. *Geospatial Space Studies: US County Boundary File with Centroids and US BEnsus FIPS Codes*. Available at: www.nceas.ucsb.edu/scicomp/casestudies. Retrieved on: 15 August 2013.

PASSI - Coordinating technical group of the behavioural risk factor system. 2013. *PASSI (Progressi Delle Aziende Sanitarie per la Salute in Italia) - The Italian behavioral risk factor surveillance system*. Available at: http://www.epicentro.iss.it/passi/en/english.asp. Retrieved on: 27 June 2013.

Pierannunzi, C. 2012. Methodologic changes in the Behavioral Risk Factor Surveillance System in 2011 and potential effects on prevalence estimates. *Morbidity and Mortality Weekly Report - Center for Disease Control*, **61**, 410–413.

R Core Team. 2012. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ruppert, D., Wand, M.P., and Carroll, R.J. 2003. *Semiparametric Regression*. Vol. 12. Cambridge University Press.

Serdula, M.K., Gillespie, C., Kettel-Khan, L., Farris, R., Seymour, J., and Denny, C. 2004. Trends in fruit and vegetable consumption among adults in the United States: behavioral risk factor surveillance system, 1994-2000. *American Journal of Public Health*, **94**(6), 1014–1018.

Shi, Z., Taylor, A.W., Goldney, R., Winefield, H., Gill, T.K., Tuckerman, J., and Wittert, G. 2011. The use of a surveillance system to measure changes in mental health in Australian adults during the global financial crisis. *International Journal of Public Health*, **56**(4), 367–372.

Simpson, M.E., Serdula, M., Galuska, D.A., Gillespie, C., Donehoo, R., Macera, C., and Mack, K. 2003. Walking trends among US adults: the behavioral risk factor surveillance system, 1987–2000. *American Journal of Preventive Medicine*, **25**(2), 95–100.

Taylor, A.W., Campostrini, S., and Beilby, J. 2013. Demographic Trends in Alcohol Use: The Value of a Surveillance System. *American Journal of Health Behavior*, **37**(5), 641–653.

Troost, J.P., Rafferty, A.P., Luo, Z., and Reeves, M.J. 2012. Temporal and Regional Trends in the Prevalence of Healthy Lifestyle Characteristics: United States, 1994–2007. *American journal of public health*, **102**(7), 1392–1398.

US Department of Health and Human Services. 2013. *Prior HHS Poverty Guidelines.* Available at: http://aspe.hhs.gov/poverty/figures-fed-reg.cfm. Retrieved on: 5 August 2013.

Wang, L., Li, H., and Huang, J.Z. 2008. Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association*, **103**(484), 1556–1569.

Wasserman, L. 2006. *All of Nonparametric Statistics.* Springer New York.

West, M., and Harrison, J. 1997. *Bayesian Forecasting and Dynamic Models.* Springer Verlag.

Wheeler, D.C., and Páez, A. 2010. Geographically weighted regression. *Pages 461–486 of: Handbook of Applied Spatial Analysis.* Springer.

Wood, S.N. 2006a. *Generalized Additive Models: an Introduction with R.* Vol. 66. Chapman & Hall.

Wood, S.N. 2006b. Low-Rank Scale-Invariant Tensor Product Smooths for Generalized Additive Mixed Models. *Biometrics*, **62**(4), 1025–1036.

Wood, S.N. 2007. *mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation.* R package version 1.7-22.

Wood, S.N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**(1), 3–36.

Wood, S.N., Scheipl, F., and Faraway, J.J. 2013. Straightforward intermediate rank tensor product smoothing in mixed models. *Statistics and Computing*, 1–20.

Wu, C.O., and Chiang, C.T. 2000. Kernel smoothing on varying coefficient models with longitudinal dependent variable. *Statistica Sinica*, **10**(2), 433–456.

Wu, C.O., Yu, K.F., and Chiang, C.T. 2000. A two-step smoothing method for varying-coefficient models with repeated measurements. *Annals of the Institute of Statistical Mathematics*, **52**(3), 519–543.

Xia, Y., Zhang, W., and Tong, H. 2004. Efficient estimation for semivarying-coefficient models. *Biometrika*, **91**(3), 661–681.

Young, L.J., Gotway, C.A., Yang, J., Kearney, G., and DuClos, C. 2008. Assessing the association between environmental impacts and health outcomes: A case study from Florida. *Statistics in Medicine*, **27**(20), 3998–4015.

Zack, M.M., Moriarty, D.G., Stroup, D.F., Ford, E.S., and Mokdad, A.H. 2004. Worsening trends in adult health-related quality of life and self-rated health-United States, 1993-2001. *Public Health Reports*, **119**(5), 493.

Zhang, W., Lee, S.Y., and Song, X. 2002. Local polynomial fitting in semivarying coefficient model. *Journal of Multivariate Analysis*, **82**(1), 166–188.

Shireen Assaf

CURRICULUM VITAE

Contact Information

University of Padova
Department of Statistics
via Cesare Battisti, 241-243
35121 Padova. Italy.

e-mail: assaf@stat.unipd.it, shireen122@gmail.com

Current Position

Since January 2011; (expected completion: March 2014)
PhD Student in Statistical Sciences, University of Padova.
Thesis title: Behaviour Risk Factor Surveillance Data using Varying Coefficient Models.
Supervisor: Prof. Stefano Campostrini
Co-supervisor: Prof. Carlo Gaetan.

Research interests

- Applied statistics for health and public health interventions through the analysis of survey data.
- Study of trends in health data.
- Multivariate statistics.
- Spatial statistics.

Education

September 2005 – June 2007
Master of Science degree in Population Health .
American University of Beirut, Faculty of Health Sciences
Title of dissertation: "Consanguinity and Reproductive wastage in the Palestinian Territories."
Supervisor: Prof. Marwan Khawaja
Final mark: 90%

September 1997 – June 2001
Bachelor of Science degree in Environmental & Occupational Health.
University of Washington, School of Public Health
Final mark: Cumulative Grade Point Average 3.5/4 - Departmental Grade Point Average 3.8/4.

Visiting periods

July 2013 – October 2013
Center for Disease Control and Prevention,
Atlanta, GA, U.S.A.
Supervisor: Dr. Carol Gotway Crawford

Work experience

March 2009 – January 2011
UNESCO project - Palestinian Women's Research and Documentation Center.
Research Coordinator.

April 2007 – September 2008
Health, Social Research and Development Consultant - Various assignments with several NGOs and the American University of Beirut.
Independent Consultant.

November 2007 – February 2008
CISP - Comitato Internazionale per lo Sviluppo dei Popoli.
Public Health Expert.

July 2006 – September 2006
World Health Organization Yemen Office.
WHO Intern.

November 2004 – September 2005
Palestine Red Crescent Society.
Project Coordinator.

September 2003 – June 2004
United States Environmental Protection Agency.
Environmental Careers Organization Paid Intern.

February 2002 – May 2003
Peace Corps Morocco.
Health and Sanitation Volunteer.

September 2001 – January 2002
University of Washington Field Research & Consultation Group.
Research Assistant II.

January 1998 – June 2001
University of Washington Chemistry Study Center.
Undergraduate Teacher's Assistant.

Awards and Scholarship

2011-2013
Scholarship from the University of Padua for PhD graduate studies.

2005-2007
Full Scholarship from the Ford Foundation for M.S. graduate studies.

2001
Outstanding Student Award in the School of Public Health and Community Medicine from the University of Washington.

2001
Cind Treser Scholarship from the Washington State Environmental Health Association.

2000 and 2001
University of Washington Quarterly Deans List.

Computer skills

- R, Stata, and SPSS.
- LaTex
- Word, Excel, Powerpoint

Language skills

English: native; Arabic: fluent; Italian: basic (written/spoken).

Publications

Articles in journals

Assaf, S., Chaban, S., (2013). Domestic violence against single never-married women in the Occupied Palestinian Territory. *Violence Against Women Journal.* **Vol. 19 No.3**, 422–441.

Khawaja, M., Assaf, S., Yamout, R. (2011). Predictors of displacement behaviour during the 2006 Lebanon war. *Global Public Health Journal.* **Vol. 6, No. 5**, 488–504.

Khawaja, M., Assaf, S., Jarallah, Y. (2009). The transition to lower fertility in the West Bank and Gaza Strip: evidence from recent surveys. *Journal of Population Research.* **Vol. 26, No. 2**, 153–174.

Assaf, S., Khawaja, M., DeJong, J., Mahfoud, Z., Younis, K. (2009). Consanguinity and reproductive wastage in the Palestinian Territories. *Paediatric and Perinatal Epidemiology.* **Vol. 23, No. 2**, 107–115.

Assaf, S., Khawaja, M. (2009). Consanguinity trends and correlates in the Palestinian Territories. *Journal of Biosocial Science.* **Vol. 41, No. 1**, 107–124.

Working papers

Khawaja, M., Jurdi, R., Assaf, S., Yeretzian, J., (2009). Unmet need for the utilization of womens labor: findings from three impoverished communities in outer Beirut, Lebanon. *Economic Research Forum Working papers series.* **No. 494**

Conference presentations

Assaf, S., Campostrini, S., (2013). Surveillance data analysis using varying coefficient models. (oral presentation) *DAGStat 2013 Conference*, Freiburg, Germany, 18 - 22 March 2013.

Assaf, S., Campostrini, S., (2013). Behaviour risk factor surveillance data analysis using time-varying coefficient models: application to smokers in Italy. (poster presentation) *The 8th World Alliance for Risk Factor Surveillance Global Conference*, Beijing, China, 29 November - 31 October.

References

Prof. Stefano Campostrini
University of Ca Foscari
Venice, Italy
Phone: +39 041 234 736
e-mail: stefano.campostrini@unive.it

Dr. Carol Gotway Crawford
Center for Disease Control and Prevention
Atlanta, GA, U.S.A.
Phone: +14043889618
e-mail: cdg7@cdc.gov

Prof. Marwan Khawaja
UN ESCWA
Beirut, Lebanon
Phone: +961-1-978365
e-mail: khawaja@un.org