STATISTICAL COMPUTATION AND INFERENCE FOR FUNCTIONAL DATA ANALYSIS

A Thesis
Presented to
The Academic Faculty

by

Huijing Jiang

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
The H. Milton Stewart School of Industrial and Systems Engineering

Georgia Institute of Technology
December 2010

STATISTICAL COMPUTATION AND INFERENCE FOR FUNCTIONAL DATA ANALYSIS

Approved by:

Professor Nicoleta Serban, Advisor
The H. Milton Stewart School of
Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Yasuo Amemiya
Statistical Analysis and Forecasting
Group
IBM T.J. Watson Research Center

Professor Alexander Gray
College of Computing
Georgia Institute of Technology

Professor Paul Kvam
The H. Milton Stewart School of
Industrial and Systems Engineering
Georgia Institute of Technology

Professor Ming Yuan
The H. Milton Stewart School of
Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: September 22nd, 2010

To my parents,

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my advisor, Dr. Nicoleta Serban. It is her who encouraged me to pursue a doctoral degree in statistics at the very beginning. I am truly honored to be the first PhD student she solely surprised. Her guidance and inspirations have not only greatly contributed to the completion of this thesis but have also been invaluable insights to every aspects of my graduate life. She has not only been my academic advisor but also an elder sister and a friend to me.

I would like to thank my thesis committee Dr. Yasuo Amemiya, Dr. Alex Gray, Dr. Paul Kvam and Dr. Ming Yuan for their valuable comments and suggestions. In particular, I would like to thank Dr. Yasuo Amemiya for his supervision and insightful discussions on statistical research during my internship at IBM T.J. Watson Research Lab. I would like to thank Dr. Alex Gray and his PhD student Nikolaos Vasiloglou II for their helpful suggestions and for their support in developing the C++ library for my research projects.

My thanks also go to Dr. John Zhang from School of Chemistry and Biochemistry for leading me to the road toward a scientific researcher; to Dr. William Rouse at Tennenbaum Institute for providing me a two-year fellowship and inspiring me of my research direction; to Dr. Jim Berger for providing me with financial support for my visit at Statistical and Applied Mathematical Sciences Institute (SAMSI) and to Dr. Nell Sedransk for being my mentor at SAMSI.

I would also say thank you to my friends, Dr. Ozgun Caliskan Demirag, Dr. Xinwei Deng, Huizhi Xie, Lingyan Ruan, Heeyoung Kim, Sungil Kim, Shaudi Hosseini and Peng Tang. Everyone of you make my life Georgia Tech so enjoyable. I am also

thankful to my former college roommates Jing Li and Shanshan Li for sharing the joys and difficulties of this experience with me.

Last but not the least, I would like to give my heartfelt appreciation and gratitude to my parents (my father Jinrong Jiang and my mother Jvfen Yan) for their constant love and support throughout these years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

My doctoral research dissertation focuses on two aspects of functional data analysis (FDA):

- FDA under spatial interdependence (Chapters 1 and 2)
- FDA for multi-level data (Chapters 3 and 4)

Functional data analysis (FDA) is a branch of statistics that offers exploratory and inferential methodology for random functions varying over a continuum. The continuum is often time, but may also be spatial location, wavelength, probability among others. In functional data analysis, data are discrete observations sampled from the random functions. The field of functional data analysis has already provided a series of competitive approaches, but they are generally limited to the assumption of independence between functionals. This assumption is rather restrictive in many research fields such as biological sciences (e.g. fMRI, microarray), climatology science (e.g. rainfall measurements across different stations), industrial engineering (e.g. performance analysis of spatially-distributed enterprises), public health (e.g. disease outbreak monitoring), and many others. The first part of my dissertation (Chapter 1 and Chapter 2) focuses on developing modelling and inference procedure for functional data under spatial dependence. The methodology introduced in this part is motivated by a research study on inequities in accessibility to financial services.

In the first chapter, I present a novel model-based method for clustering random time curves or functions which are spatially interdependent. A cluster consists of time functions which are similar in shape. The time functions are decomposed into spatial global and time-dependent cluster effects using a semi-parametric model. We also assume that the clustering membership is a realization from a Markov random

field. Under these model assumptions, we borrow information across curves from nearby locations resulting in enhanced estimation accuracy of the cluster effects and of the cluster membership. In a simulation study, we assess the estimation accuracy of our clustering algorithm under a series of settings: small number of time points, high noise level and varying dependence structures. Over all simulation settings, the spatial-functional clustering method outperfoms existing model-based clustering methods. In the case study presented in this chapter, we focus on estimates and classifies service accessibility patterns varying over a large geographic area (California and Georgia) and over a period of 15 years. The focus of this study is on financial services but it generally applies to any other service operation.

Chapter 2 introduces an association analysis of space-time varying processes, which is rigorous, computational feasible and implementable with standard software. We introduce general measures to model different aspects of the temporal and spatial association between processes varying in space and time. Using a nonparametric spatiotemporal model, we show that the proposed association estimators are asymptotically unbiased and consistent. We complement the point association estimates with simultaneous confidence bands to assess the uncertainty in the point estimates. In a simulation study, we evaluate the accuracy of the association estimates with respect to the sample size as well as the coverage of the confidence bands. In the case study in this chapter, we investigate the association between service accessibility and income level. The primary objective of this association analysis is to assess whether there are significant changes in the income-driven equity of financial service accessibility over time and to identify potential under-served markets.

In the second part of the thesis (Chapters 3 and 4), I discuss novel statistical methodology for analyzing multilevel functional data including a clustering method based on a functional ANOVA model and a spatio-temporal model for functional data with a nested hierarchical structure.

In Chapter 3, I introduce and compare a series of clustering approaches for multilevel functional data. For brevity, I present the clustering methods for two-level data: multiple samples of random functions, each sample corresponding to a case and each random function within a sample/case corresponding to a measurement type. A cluster consists of cases which have similar within-case means (level-1 clustering) or similar between-case means (level-2 clustering). Our primary focus is to evaluate a model-based clustering to more straightforward hard clustering methods. The clustering model is based on a multilevel functional principal component analysis. In a simulation study, we assess the estimation accuracy of our clustering algorithm under a series of settings: small vs. moderate number of time points, high noise level and small number of measurement types. We demonstrate the applicability of the clustering analysis to a real data set consisting of time-varying sales for multiple products sold by a large retailer in the U.S.

The last chapter briefly presents my ongoing research project on developing a statistical model for estimating temporal and spatial associations of a series of time-varying variables with an intrinsic nested hierarchical structure. This work has a great potential in many real applications where the data are areal data collected from different data sources and over geographic regions of different spatial resolution.

CHAPTER I

CLUSTERING RANDOM CURVES UNDER SPATIAL INTERDEPENDENCE

1.1 Introduction

Due to an increasing number of applications with a very large number of random (time) functions, data reduction methods such as clustering play an important role in Functional Data Analysis (FDA). The literature on functional data clustering is divided into hard and model-based methods. Examples of hard clustering methods are Hastie et al. (2000); Bar-Joseph et al. (2002); and Serban (2008). Examples of model-based clustering are James and Sugar (2003); Fraley and Raftery (2002); and Wakefield et al.(2002). Although there are many competitive approaches to clustering functional data, they are generally limited to the assumption of independence between curves. However, there are many case studies including our motivating application where this assumption does not hold - the service accessibility functions are spatially interdependent since each function corresponds to a census tract. It is important to account for interdependence not only to enhance the estimation accuracy of the cluster patterns and of the cluster membership by borrowing information across dependent curves but also to allow estimation of the underlying dependence. Recent research in clustering functional data has considered spatial interdependence in the cluster membership (Blekas et al., 2007, Shi and Wang, 2008) or within-cluster dependence (Booth et al., 2008).

In contrast to the existing approaches for clustering functional data under spatial interdependence, the method introduced in this chapter models the spatial interdependence in the joint distribution of the functional data and the cluster membership,

and therefore, it allows for both within- and between-cluster interdependence. It is important to assume both within- and between-spatial interdependence because the spatial interdependence extends beyond the cluster membership. Moreover, the clustering method is based on a semi-parametric model that allows estimation of both global temporal-spatial effects as well as cluster effects. Lastly, we propose a computationally efficient method by employing a low-rank approximation to reduce the size of the dependence matrix. Clustering is commonly used as a tool for summarizing a large number (more than 1000) of functional profiles, and therefore, computational efficiency is crucial in clustering spatially interdependent functional data since the dependence matrix tends to be very large.

We cluster multiple time functions observed with error:

$$Y_{ij} = f_{s_j}(t_{ij}) + \sigma_\varepsilon \varepsilon_{ij},\ j = 1, \ldots, S, \tag{1}$$

where $f_{s_j}(t)$ is the time function corresponding to location s_j and $(t_{1,j}, \ldots, t_{T,j})$ are the observed time points for this time function. Additionally, s_j for $j = 1, \ldots, S$ are spatial units from a d-dimensional spatial domain. In the model-based clustering framework, the complete data are $(Y_j, Z_j),\ j = 1, \ldots, S$ where $Y_j = (Y_{1j}, \ldots, Y_{Tj})$ and Z_j's are missing latent variables defining the cluster membership. We refer to our modeling procedure as the *Functional-Spatial Clustering Model (FSCM)*.

A first contribution of the clustering method in this chapter is decomposing the time functions $f_{s_j}(t)$ for $j = 1, \ldots, S$ into global and cluster effects which summarize a large number of spatially-dependent curves to meaningful summaries including spatial trends and summary temporal patterns. The decomposition is

$$f_{s_j}(t) = \mu(t) + \tau(s_j) + \mu_{0,j} + \mu_{Z(s_j)}(t). \tag{2}$$

The global temporal effect $\mu(t)$ is separable from the global-spatial effect $\tau(s)$ to simplify the interpretation of the spatial-temporal global pattern. The cluster effects are conditional on the cluster membership $Z_j = Z(s_j),\ j = 1, \ldots, S$. The cluster

trends are spatially-scaled deviations from the overall trend $\mu(t)$. The spatial-global effect $\tau(s)$ accounts for spatial variations. The parameters $\mu_{0,j}$ for $j = 1, \ldots, S$ are curve-specific deviations from the spatial global pattern; $\mu_k(t), k = 1, \ldots, C$ are time-dependent cluster effects and we constrain their sum to be equal to zero - $\sum_{k=1}^{C} \mu_k(t) = 0$. These offset parameters ensure clustering by shape regardless of scale. In Section 1.2.1, we expand on the spatial-functional clustering model.

The second contribution of this chapter is that our model assumes (Y_j, Z_j) for $j = 1, \ldots, S$ spatially correlated; that is, both the latent variables $Z_j,\ j = 1, \ldots, S$ and the conditional observations $Y_j|Z_j,\ j = 1, \ldots, S$ are spatially dependent, which further specifies the spatial dependence structure for the joint data $(Y_j, Z_j),\ j = 1, \ldots, S$. Under the spatial dependence of the joint data, we borrow information across curves corresponding to nearby locations yet maintaining local resolution. In Section 1.2.2, we describe the distribution assumptions and provide insights into the computational and estimation challenges under these assumptions.

The third contribution of this chapter is an estimation procedure which allows clustering a large number of time functions (S large). There are two computational challenges that we address in Section 1.3.

To assess the advantages and limitations of the spatial-functional clustering method, we illustrate our methodology with simulated data (Section 3.6). In the simulation study, we investigate a range of model scenarios with varying noise levels and spatial correlation structures to compare the spatial-functional clustering introduced in this chapter to existing model-based clustering methods.

One motivating application of our method in this chapter is to analyze the accessibility and the equitable distribution of financial services. Research in service accessibility has emerged as economic and social equity advocates recognized that where people live influences their opportunities for economic development, access to quality healthcare, and political participation (Blackwell and Treuhaft, 2008; Frumkin

et al., 2004; Lee and Rubin, 2007; Morland et al., 2002). In this chapter, we leverage new statistical methods for estimating and describing service accessibility trends that can be used to inform about potential business opportunities as well as about the extent of service distribution inequities. Inequity in service accessibility results in site configurations with significant concentrations of service sites in some geographic areas (served markets) and virtually no service sites in others (unserved or under-served markets), even though current and potential customers are present in both.

Many existing studies have analyzed the accessibility and the equitable distribution of various services but they are limited to small geographic areas such as towns and only one year of data (Graves, 2003; Larson, 2003; Moore et al., 2006; Morland et al., 2002; Small and McDermott, 2006; Talen, 2001; Talen and Anselin, 1998). One primary challenge in analyzing service accessibility over a large geographic area with inference at a high resolution level and a long period of time is estimation and characterization of a large number of time-varying accessibility curves/functions. For example, in this chapter we measure service accessibility in California and Georgia at the census tract level over 15 years; this results in 7115 accessibility functions in California and 1624 in Georgia. To prevail over this challenge, we propose using a clustering method to reduce the information content of geographically and temporally varying data to meaningful global spatial-temporal trends as well as summary local temporal trends that reveal the prevalent changes in service accessibility in a given geographic space. The focus of this study is on the distribution of financial services but the service accessibility framework applies generally to other services both public (e.g. education) and private (e.g. food stores).

1.2 Functional Spatial Clustering Model

In model-based clustering, the underlying assumption is that the complete data are bivariate variables (Y_j, Z_j) for $j = 1, \ldots, S$ where $Y_j = \{Y_{ij}\}_{i=1,\ldots,T}$ are realizations

from a multivariate distribution with mean vector μ_j and covariance Σ_j, and the cluster membership Z_j is a latent variable (Fraley and Raftery, 2002). A common estimation method for model-based clustering is the Estimation-Maximization algorithm where at the Estimation step, we impute or predict the cluster membership $Z = (Z_1, \ldots, Z_S)$ and at the Maximization step, we estimate the parameters specifying the conditional distribution of $Y_j|Z,\ j = 1, \ldots, S$. Therefore, we need to specify the conditional distribution $Y_j|Z,\ j = 1, \ldots, S$ and the distribution of the latent variable Z which in turn, specifies the distribution of the complete data. *In the existing approaches to model-based clustering, $Y_j|Z,\ j = 1, \ldots, S$ are assumed conditionally independent and the clustering membership $Z_j,\ j = 1, \ldots, S$ are also assumed independent. In this chapter, we relax these assumptions to spatial dependence.* In Section 1.2.1, we introduce a functional model for the conditional distribution of $Y_j|Z$, and in Section 1.2.2, we describe a locally-dependent Markov model for the latent variable $Z = \{Z_j\}_{j=1,\ldots,S}$.

1.2.1 Functional Model for Conditional Distribution

Given the cluster membership $\{Z_1 = z_1, \ldots, Z_S = z_S\}$ with $z_1, \ldots, z_S \in \{1, \ldots, C\}$ (C is the number of clusters), we assume that $Y_{ij}|Z_j$ follows a multivariate distribution with a functional representation

$$Y_{ij}|(Z_j = k) = \mu(t_i) + \tau(s_j) + \mu_{0,j} + \mu_k(t_i) + \sigma_\varepsilon^2 \varepsilon_{ij} \tag{3}$$

where $\mu(t)$ is the global mean, $\tau(s)$ is the smooth spatial variation, $\mu_{0,j}$ is a curve-specific offset parameter and $\mu_k(t)$ is the cluster trend. The errors are assumed independent and identically distributed. In our estimation algorithm, we first estimate $\mu(t_i)$ and then estimate the other model components from the demeaned data, $Y_{ij} - \mu(t_i)$. Therefore, without loss of generality, we take $\mu(t) = 0$.

The spatial dependence in our model comes from two sources: 1. The spatial

level-shifts $\tau(s_j)$, which account for global spatial variations regardless of the clustering membership; and 2. The cluster memberships Z_i which are spatially dependent - locations close together are more likely to be in the same cluster. We model the temporal functionality in our data through the global effect $\mu(t)$, which is the overall time trend; and the cluster shape functions $\mu_k(t)$. In this chapter, we assume that the classes of functions where $\mu_k(t)$ and $\tau(s)$ lie are (separable) Hilbert spaces, $\mu_k(t) \in \mathcal{H}^k,\ k = 1, \ldots, C$, and respectively, $\tau(s) \in \mathcal{H}^S$. Following the functional representation in a reproducing-kernel Hilbert space (Wahba, 1990), we decompose the cluster patterns and the spatial trend according to

$$\mu_k(t) = \sum_{\nu=1}^{p} \phi_{T,\nu}(t)\beta_{k,\nu} + \sum_{i=1}^{T} K_T(t, t_i)u_{k,i} \text{ and } \tau(s) = \sum_{\nu=1}^{q} \phi_{S,\nu}(s)\alpha_\nu + \sum_{j=1}^{S} K_S(s, s_j)w_j \quad (4)$$

where the p basis functions $\{\phi_{T,1}, \ldots, \phi_{T,p}\}$ span $\mathcal{H}_0^k$ and $K_T(t, t_i)$ is the reproducing kernel for the space $\mathcal{H}_1^k$, and therefore, the Hilbert space $\mathcal{H}^k$ decomposes into $\mathcal{H}^k = \mathcal{H}_0^k \oplus \mathcal{H}_1^k$ with $\mathcal{H}_1^k \perp \mathcal{H}_0^k$. Similarly, we represent $\mathcal{H}^S = \mathcal{H}_0^S \oplus \mathcal{H}_1^S$ where $\{\phi_{S,1}, \ldots, \phi_{S,q}\}$ span $\mathcal{H}_0^S$ and $K_S(s, s_i)$ is the reproducing kernel for $\mathcal{H}_1^S$. Under this model formulation, the estimation of the offset parameters $\mu_{0,j}$ and of the functions $\mu_k(t),\ k = 1, \ldots, C$ and $\tau(s)$ is equivalent to solving a penalized least squares problem

$$\frac{1}{S}\frac{1}{T}\sum_{i=1}^{T}\sum_{j=1}^{S} \|Y_{ij} - \tau(s_j) - \mu_{0,j} - \mu_k(t_i)\|^2 + \sum_{k=1}^{C} \lambda_k \|P_\perp \mu_k\| + \lambda_S \|P_\perp \tau_s\| \quad (5)$$

where $\|P_\perp \mu_k\| = u_k K_T u_k'$ with $K_T = \{K_T(t_{i_1}, t_{i_2})\}_{i_1, i_2 = 1, \ldots, T}$ penalizing the shape of the kth cluster pattern and $\|P_\perp \tau_s\| = w K_S w'$ with $K_S = \{K_S(s_{i_1}, s_{i_2})\}_{i_1, i_2 = 1, \ldots, S}$ penalizing the shape of the spatial global effect. λ_k is the smoothing parameter for $\mu_k(t)$, which controls the trade-off between goodness-of-fit and the departure of the estimate from the space $\mathcal{H}_0^k$. λ_s is the smoothing parameter for $\tau(s)$.

1.2.2 Markov Model for the Cluster Membership

We assume that the clustering configuration $Z = (Z_1 = Z(s_1), \ldots, Z_S = Z(s_S))$ is a realization of a locally-dependent Markov random field (MRF) which is a stochastic

process with the Markov property. Under the Markov assumption, the probability that the s_ith spatial unit belongs to kth cluster depends on the states of its nearest neighbors where a state is defined by the cluster membership but it is conditionally independent of any other spatial units. Following the current literature on MRF modelling, we model the probability mass function $P(Z_j = z_{s_j}) = p(z_{s_j})$ with Gibbs distribution. This distribution originates from statistical physics where it is used to model the states of atoms and molecules, and later on, was adopted by statisticians to model Markov Random Fields. The probability mass function for Gibbs distribution is defined as

$$p(z_{s_j} = k) = \pi_{s_j,k} = \frac{1}{N_{s_j}(\psi)} \exp(U_{s_j,k}(\psi)) \tag{6}$$

where $U_{s_j,k}(\psi) = \sum_{s_i \in \partial s_j} \psi I(z_{s_i} = k)$ is called the energy function. Large values of $U_{s_j}(\psi)$ correspond to spatial patterns with large spatially connected sub-areas belonging to the same cluster. Small values of $U_{s_j}(\psi)$ correspond to spatial patterns that do not display any sort of spatial organization. $N_{s_j}(\psi)$ is a normalizing constant called the *partition function* and ∂s_j is a prescribed neighborhood of the s_jth spatial unit for which we apply $k-$nearest neighbors to define the neighborhood structure. The probability mass function depends on ψ called the *interaction parameter*. The larger ψ is, the more extensive the spatial dependence of cluster membership Z is. The value $\psi = 0$ corresponds to the uniform distribution on the configuration space.

One difficulty in this formulation is that the normalizing constant depends on the scale parameter ψ. Because of this dependence, we do not have a closed form expression for the likelihood function when ψ is a parameter. In the HMRF literature (Besag, 1986, Archer and Titterington, 2002), this difficulty is overcome by assuming local dependence on each spatial unit s_i, i.e., s_i only depends on its neighbors ∂s_i. Thus the joint distribution of $Z_1, \ldots, Z_S$ can be approximated by a pseudo-likelihood

function,

$$f(z_1, \ldots, z_S) \approx \prod_{j=1}^{S} f(z_{s_j} | z_{\partial s_j}; \psi). \tag{7}$$

In addition to the difficulty of estimating ψ, computational challenges arise in recovering the cluster membership $Z_1, \ldots, Z_S$ because of the spatial dependence between the $Y_j|Z$. To the best of our knowledge, in all relevant work, the $Y_j|Z$ are assumed conditionally independent for computational feasibility although this is one of the most contested assumptions (see Besag, 1986 and the following discussions; Archer and Titterington 2002; and the references therein). Because we allow for a global-spatial effect, our model relaxes this assumption to spatial dependence.

1.3 Computational Challenges

Following the functional model in (3) and the global and cluster effects decomposition in (4), we rewrite the model in a vector-matrix form

$$Y|Z = \mathbf{I}\mu_0 + X_T\beta + B_T u + X_S\alpha + B_S w + \sigma_\varepsilon^2 \varepsilon \tag{8}$$

• Y is a vector consisting of all observations stacked in the following order $Y = (Y_1', \ldots, Y_S')' = (Y_{11}, \ldots, Y_{T1}; Y_{12}, \ldots, Y_{T2}; \ldots Y_{1S}, \ldots, Y_{TS})'$;

• $\mu_0 = (\mu_{0,1}, \ldots, \mu_{0,S})'$; and $\beta = (\beta_1', \ldots, \beta_K')'$ with $\beta_k = (\beta_{k1}, \ldots, \beta_{kp})'$;

• $u = (u_1', \ldots, u_K')'$ with $u_k = (u_{k,1}, \ldots, u_{k,T})'$; $\alpha = (\alpha_1, \ldots, \alpha_q)'$ and $w = (w_1, \ldots, w_S)'$.

The design matrices are $\mathbf{I} = (I_S \otimes \mathbf{1}_T)$, $X_T = E \otimes \Phi_T$ with $\Phi_T = \{\phi_{T,\nu}(t_i)\}_{i=1,\ldots,T;\nu=1,\ldots,p}$ for the cluster effects, and $X_S = \Phi_S \otimes \mathbf{1}_T$, with $\Phi_S = \{\phi_{S,\nu}(s_j)\}_{j=1,\ldots,S;\nu=1,\ldots,q}$ for the spatial-global trend. The kernel matrices are $B_T = E \otimes K_T$ and $B_S = K_S \otimes \mathbf{1}_T$. Moreover, $E = \{\delta(z_j = k)\}_{j=1,\ldots,S;k=1,\ldots,C}$ is an indicator matrix.

Under this model formulation, solving the penalized least squares problem in (5) is equivalent to estimating the model parameters μ_0, β, u and α by minimizing

$$\frac{1}{S}\frac{1}{T}\|Y - \mathbf{I}\mu_0 - X_T\beta - B_T u - X_S\alpha - B_S w\|^2 + \sum_{k=1}^{C} \lambda_k u_k' K_T u_k + \lambda_s w' K_S w \tag{9}$$

One computational challenge in fitting the functional model is that the estimation algorithm involves operations with the kernel matrix K_S which takes a large amount of CPU time and memory storage. A second challenge is the selection of the smoothing parameters $\lambda_k,\ k = 1, \ldots, C$ and λ_S. A common method for obtaining smoothing parameters is by Generalized Cross Validation (GCV) (Wahba 1990). However, applying GCV to our model involves a high-dimensional optimization problem, which is computational infeasible. We address these two computational challenges in this section.

1.3.1 Low Rank Approximation of K_S

In our model fitting, we need to perform linear algebra operations with a high-dimensional full rank (S-by-S) kernel matrix K_S which requires operations of computational complexity $O(S^3)$. To reduce the computational cost, we propose reducing the dimensionality of the kernel matrix K_S by a low-rank (S-by-J, $J \ll S$) approximation $\tilde{K}_S$. We minimize the penalized least squares problem in (9) using the approximated $\tilde{K}_S$ instead of the full rank kernel matrix K_S. The optimal choice of $\tilde{K}_S$ will result in minimum possible smoothing perturbation.

There are several approaches for finding a low rank approximation $\tilde{K}_S$. One possibility is to define $\tilde{K}_S$ on a coarse grid with locations $\kappa = (\kappa_1, \ldots, \kappa_J)$ (known as 'knots') superimposed on the domain of the original locations $\mathbf{s} = (s_1, \ldots, s_S)$. The resulting low-rank approximation of K_S is $\tilde{K}_S = \{K_S(s_i, \kappa_j)\}_{i=1,\ldots,S;j=1,\ldots,J}$. Another alternative is using the empirical orthogonal functions (Wikle and Cressie, 1999).

In this chapter, we use the method by Wood (2003) where we define $\tilde{K}_S = K_S U_J$ and $\hat{K}_S = (U_J \tilde{K}_S)'$ where the columns of U_J are the J eigenvectors corresponding to the largest J absolute eigenvalues from the eigen-decomposition of $K_S = UDU'$ where D is a diagonal matrix of eigenvalues of K_S and U is a matrix whose ith column is the eigenvector corresponding to $d_{i,i}$, the ith diagonal element of D.

Under low-rank approximation, the penalized least square problem becomes

$$\frac{1}{S}\frac{1}{T}\|Y - \mathbf{I}\mu_0 - X_T\beta - B_T u - X_S\alpha - \tilde{B}_S w_J\|^2 + \sum_{k=1}^{C} \lambda_k u_k' K_T u_k + \lambda_s w_J' \hat{K}_S w_J \quad (10)$$

where $\tilde{B}_S = \tilde{K}_S \otimes \mathbf{1}_T$ and w_J is a vector of length J related to w by $w = U_J w_J$. Because of the low-rank approximation, we will obtain an approximated estimator for the spatial effect $\tau(s)$. This approximated estimator is accurate when $K_S w - \tilde{K}_S w_J$ in the model fit and $w' K_S w - w_J \hat{K}_S w_J$ in the penalty term $\|P_\perp \tau_s\|$ are small. Wood (2003) proved that constructing $\tilde{K}_S$ using the truncated eigenvectors minimizes the change of the model fit and the shape of the smooth function simultaneously.

1.3.2 Smoothing Parameters

To obtain a solution of the penalized least squares problem in (9), we write the proposed model in an equivalent mixed effects model (Ruppert, Wand and Carroll, 2003). We first re-scale our data and rewrite the model as

$$Y|Z = X\theta + Bb + \varepsilon \quad (11)$$

where $\theta = (\mu_0', \beta', \alpha')'$ and $b = (u', w_J')'$. The design matrices are $X = \begin{bmatrix} \mathbf{I} & X_T & X_S \end{bmatrix}$, and $B = \begin{bmatrix} B_T & \tilde{B}_S \end{bmatrix}$. Then the solution of minimizing the penalized likelihood (10) is the solution of the system of linear equations

$$X'X\theta + X'Bb = X'Y$$
$$B'X\theta + (B'B + V)b = B'Y$$

where $V = diag(ST\lambda_1 K_T, \ldots, ST\lambda_C K_T, ST\lambda_s \hat{K}_S)$.

It follows that the spline smoothing model is equivalent to a linear mixed model. If we apply Cholesky factorization, $K_T = H_T H_T'$ and $\hat{K}_S = H_S H_S'$, the model (11) becomes

$$Y|Z = X\theta + H\gamma + \varepsilon \quad (12)$$

where θ is a vector of fixed effects and γ is a vector of random effects and distributed as $N(0, \Gamma)$ with $\Gamma = diag(\sigma_1^2 \mathbf{I}_T, \ldots, \sigma_C^2 \mathbf{I}_T, \sigma_s^2 \mathbf{I}_J)$ and the design matrix of the random effects is $H = \begin{bmatrix} E \otimes H_T & \tilde{H}_S \otimes \mathbf{1}_T \end{bmatrix}$ with $\tilde{H}_S = \tilde{K}_S (H_S')^{-1}$. The solution to the above penalized smoothing model is equivalent to solving a normal equation

$$\begin{bmatrix} X'X & X'H \\ H'X & H'H + \sigma_\varepsilon^2 \Gamma^{-1} \end{bmatrix} \begin{bmatrix} \theta \\ \gamma \end{bmatrix} = \begin{bmatrix} X'Y \\ H'Y \end{bmatrix}$$

where the smoothing penalties are $\lambda_k = \frac{\sigma_\epsilon^2}{TS\sigma_k^2}$ and $\lambda_s = \frac{\sigma_\epsilon^2}{TS\sigma_s^2}$.

1.4 Model Estimation and Selection

1.4.1 Estimation Algorithm

A modified EM algorithm is applied to estimate the model parameters $\Theta = (\mu_0, \beta_1, \ldots, \beta_C, \alpha, \sigma_\varepsilon^2, \psi)$ and predict the cluster membership by maximizing the the complete likelihood,

$$L(\Theta) = f(Y, \gamma, Z; \mu_0, \theta, \sigma_\varepsilon^2, \sigma_s^2, \sigma_1^2, \ldots, \sigma_C^2, \psi) =$$

$$f(Y|\gamma, Z; \mu_0, \theta, \sigma_\varepsilon^2) \cdot f(\gamma|Z; \sigma_s^2, \sigma_1^2, \ldots, \sigma_C^2) \cdot f(z_1, \ldots, z_S; \psi). \tag{13}$$

where the joint likelihood $f(z_1, \ldots, z_S; \psi)$ is approximated by the psuedo-likehood 7.

The E-step of the EM algorithm consists of finding the conditional expectation of (13) given the observed data Y and the current parameter estimates. The M-step updates the parameters Θ by maximizing the expectation of the likelihood function (13). The constraint on the cluster fixed effects ($\beta_1 + \ldots + \beta_C = 0$) ensures identifiability of the model parameters. The details of the E and M iterative steps in the EM estimation algorithm can be found in the Appendix A.

1.4.2 Select The Number of Clusters

One difficulty in unsupervised classification or clustering is that the number of clusters is unknown. The problem of identifying the number of clusters is equivalent to a model

selection problem. AIC is a common likelihood-based model selection criterion. When the model under consideration contains random effects, it is not straightforward what likelihood function to use in defining AIC. Vaida and Blanchard (2005) discussed this issue by defining two variations of AIC - marginal AIC (mAIC) and conditional AIC (cAIC) for mixed-effects model selection. Following their arguments, if only the fixed effects contain information about the number of clusters, mAIC should be used. For $mAIC = -2logf(y|\hat{\theta}) + 2(\#\text{ of parameters})$, $f(y|\hat{\theta})$ is the marginal likelihood corresponding to the model

$$Y = X\theta + \epsilon, \epsilon \sim N(0, H\Gamma H' + \sigma_\varepsilon^2 I_{ST})$$

and the number of parameters is the sum of the fixed effects and the number of parameters specifying the random effects. On the other hand, if the model selection involves both the fixed effects and random effects, conditional AIC (cAIC) should be used. For $cAIC = -2logf(y|\hat{\theta}, \gamma) + 2(\#\text{ of parameters})$, $f(y|\hat{\theta}, \gamma)$ is the conditional likelihood corresponding to the model

$$Y = X\theta + H\gamma + \varepsilon, \varepsilon \sim N(0, \sigma_\varepsilon^2 I_{ST})$$

and the number of parameters is the effective degrees of freedom.

Our model formulation is different from Vaida and Blanchard (2005) in that there are two multivariate random variables to condition on Y, the latent variable Z and the random effects γ and therefore, the conditional AIC does not apply. If we were to use the marginal likelihood as defined above, we would integrate out γ which incorporates information about the clustering. Instead we consider the joint likelihood in (13) to select the number of clusters. Following the notation by Vaida and Blanchard (2005), we define the AIC variant with joint likelihood as $jAIC = -2logf(y, \gamma, Z) + 2df$ where df is a function of C.

1.5 Simulation

In this simulation study, our primary objective is to assess the prediction accuracy of cluster membership under a series of spatial interdependence structures and varying noise levels. We compare our method with two other existing model-based clustering methods: *Mclust* (Fraley and Raftery, 2002) and *Fclust* (James and Sugar, 2003)

In our simulation model, the cluster shapes $\mu_{z_j}(t)$ and spatial scaling function $\tau(s)$ are generated using different basis functions and kernels from the ones in our estimation procedure described in Section 1.2.1. The objective is to assess that the proposed estimation method provides accurate estimates of the clustering membership, cluster patterns and spatial dependence under different model specifications.

1.5.1 Simulation Set-up

We generated a synthetic data with six clusters of curves from the functional model

$$Y_{ij}(t_{ij}) = \tau(s_j) + \mu_{z_j}(t_{ij}) + \sigma_\varepsilon^2 \epsilon_{ij}, \text{ with } t_{ij} = (i-1)/(T-1),\ i = 1, \ldots, T,\ j = 1, \ldots, S \quad (14)$$

Functional Model for Conditional Distribution. We simulate the time-dependent cluster patterns from

$$\mu_{z_j}(t_{ij}) = \sum_{\nu=1}^{5} (\theta_{z_j,\nu} + \gamma_{j,\nu}) b_\nu(t_{ij})$$

where $z_j \in \{1, \ldots, 6\}$ is the cluster membership of the jth curve ($C = 6$ clusters), $b_\nu(t), \nu = 1, 2, \ldots, 5$ are cosine basis functions and $\theta_{z_j} = (\theta_{z_j 1}, \ldots, \theta_{z_j 5})$ are the first five coefficients obtained from the Fourier decomposition of the functional patterns in Figure 1 (a) to (f) (shown in red). We add the random disturbances $\gamma_j \sim N(0, \sigma_\gamma^2 I), \sigma_\gamma^2 = 0.1$ to the coefficients θ_{z_j} to slightly distort the functional patterns. In this study, we consider $T = 10$. We simulate the global spatial pattern from

$$\tau(s_j) = \sum_{\nu=1}^{q} \alpha_\nu \phi_{S,\nu}(s_j) + \sum_{i=1}^{S} \gamma_i K_S(\|s_i - s_j\|)$$

where $\phi_{S,\nu}(s_j)$ is defined in Section 1.2.1, $\alpha_\nu = (0.1, 0.1)'$ and $\gamma_i \sim N(0, \sigma_s^2)$ with $\sigma_s^2 = 1$ or 10. We simulate the spatial dependence using the Matérn covariance matrix

$$K_S(\|s_i - s_j\|) = \sigma_s^2 \frac{1}{\Gamma(\rho)} \left(\frac{\phi \|s_i - s_j\|}{2} \right)^\rho 2B_\rho(\phi \|s_i - s_j\|).$$

Matérn class of functions provides correlation surfaces with a wide range of smoothness levels controlled through the parameter ρ (see Matérn, 1986). The range parameter ϕ defines the extent of spatial dependence. B_ρ is the modified Bessel function of the second kind of order $\rho > 0$. In our simulation, we use Matérn covariance matrix of order $\nu = \frac{2}{3}$ and range parameter $\phi = 0.1 \max_{i,j} \|s_i - s_j\| = 14.67$.

We generate $\mu_{z_j}(t)$ using a Fourier basis and $\tau(s)$ using Matérn kernel and estimate them using the decomposition (4) with a thin plate splines basis.

Markov Model for the Cluster Membership. The spatial units $s_1, \ldots, s_S$ with $S = 8454$ are the centroids of the census tracts in five southeastern state - Florida, Georgia, South Carolina, North Carolina and Tennessee. The cluster membership $(z_{s_1}, \ldots, z_{s_S})$ is generated from the Gibbs distribution described in Section 1.2.2 with $\psi = 0.5$ or $\psi = 0.9$. In Figure 2, we present the maps of the spatial distribution of Z for $\psi = 0.5$ and $\psi = 0.9$. Following HMRF methodology, the cluster membership of a site s_j is sampled from a multinomial distribution with proportion parameters $(\pi_{s_j,1}, \ldots, \pi_{s_j,C})$ where $\pi_{s_j,k}$ is the Gibbs distribution defined in Equation (6).

Simulation settings. We investigate the estimation accuracy of the cluster membership and of the dependence structure by varying three model parameters:

1. Spatial Dependence of Z controlled by the hyperparameter ψ. The larger ψ is, the more extensive the spatial dependence of cluster membership Z is.

2. Conditional Spatial Dependence controlled by σ_s^2. In our study, $cov(Y_i|Z) = \sigma_s^2 K_S K_S' + \sigma_\varepsilon^2 I_S$ where $Y_i = (Y_{i1}, \ldots, Y_{iS})'$ is the vector of S locations at ith time point. Given $\sigma_\varepsilon^2 I_S$, the Conditional Spatial Dependence is strong when σ_s^2 is large.

3. Noise level controlled by σ_ε^2.

Table 1: Model Settings: Spatial Dependence(Left), Conditional Spatial Dependence(Middle), Noise Level(Right)

	$\psi = 0.5$	
	$\sigma_s^2 = 1$	$\sigma_s^2 = 10$
$\sigma_\varepsilon^2 = 10$	Weak; Weak; High	Weak; Strong; High
$\sigma_\varepsilon^2 = 50$	Weak; Weak; Low	Weak; Strong; Low
	$\psi = 0.9$	
	$\sigma_s^2 = 1$	$\sigma_s^2 = 10$
$\sigma_\varepsilon^2 = 10$	Strong; Weak; High	Strong; Strong; High
$\sigma_\varepsilon^2 = 50$	Strong; Weak; Low	Strong; Strong; Low

Table 1 lists eight scenarios derived from combining the three factors above.

Number of clusters. We applied jAIC to all eight settings in Table 1 with C ranging from 1 to 10 clusters; for all cases, the minimum value for jAIC is attained at $C = 6$ clusters as initially simulated, validating jAIC as a criterion for selecting the number of clusters.

1.5.2 The Accuracy of the Cluster Membership Estimation

In our synthetic example, because we have the true clustering membership, we can assess the accuracy of the clustering prediction for the method introduced in this chapter and other existing methods using a clustering/classification error. We measure the clustering error using the Rand index (Rand, 1971), which is the fraction of all misclustered pairs of curves. Let $\mathcal{C} = \{f_1, \ldots, f_S\}$ denote the set of true curves, $\hat{\mathcal{C}} = \{\hat{f}_1, \ldots, \hat{f}_S\}$ denote the set of estimated curves, and T and $\hat{T}$ denote the true and estimated clustering maps, respectively. Rand index is defined by

$$\mathcal{R}(\mathcal{C}, \hat{\mathcal{C}}) = \frac{\sum_{r<s} I(T_k(f_r, f_s) \neq \hat{T}_k(f_r, f_s))}{\binom{N}{2}}.$$

Therefore, the Rand index is low when there are only few misclustered curves. We compute the Rand index for FSCM, Mclust, Fclust.

Table 2 compares the Rand index under a series of simulation settings described in the previous section. We summarize our findings as follows:

Table 2: Rand index for the clustering membership, FSCM, Mclust, Fclust

	$\psi = 0$	
	$\sigma_s^2 = 1$	$\sigma_s^2 = 10$
$\sigma_\varepsilon^2 = 10$	$0.45\%, , 11.5\%$	$7.63\%, , 11.9\%$
$\sigma_\varepsilon^2 = 50$	$11.1\%, , 14.4\%$	$14.6\%, , 17.3\%$
	$\psi = 0.5$	
	$\sigma_s^2 = 1$	$\sigma_s^2 = 10$
$\sigma_\varepsilon^2 = 10$	$0.34\%, 0.35\%, 11.7\%$	$0.39\%, 0.46\%, 9.95\%$
$\sigma_\varepsilon^2 = 50$	$8.13\%, 13.56\%, 14.5\%$	$8.21\%, 14.5\%, 18.3\%$
	$\psi = 0.9$	
	$\sigma_s^2 = 1$	$\sigma_s^2 = 10$
$\sigma_\varepsilon^2 = 10$	$0.24\%, 0.381\%, 7.05\%$	$0.21\%, 0.47\%, 8.42\%$
$\sigma_\varepsilon^2 = 50$	$6.83\%, 12.55\%, 14.36\%$	$5.17\%, 9.63\%, 17.2\%$

1. FSCM outperforms both Mclust and Fclust under all experimental settings. We find that the cluster membership prediction accuracy improves significantly.

- Under strong spatial correlation in the cluster membership Z controlled by ψ.
- Under strong conditional spatial correlation controlled by σ_s^2 (keep σ_ε^2 fixed).

2. When the noise level (controlled by σ_ε^2) is high, we observe a less accurate clustering estimation over all three methods. However, the clustering error for FSCM is lower under strong dependence implying that *assuming spatial dependence enhances the prediction accuracy of the cluster membership by borrowing information across curves in the nearby locations.*

Spatial Dependence of the Clustering Membership. The spatial dependence of the cluster membership $Z = (Z_1, \ldots, Z_S)$ is modelled from a Markov Random Field where $Z = (Z_1, \ldots, Z_S)$ follows the Gibbs distribution. The hyperparameter ψ in Gibbs distribution determines the extension of the spatial correlation. In our simulation study, we examine the cluster estimation accuracy for two different values of this hyperparameter ($\psi = 0.5$ and $\psi = 0.9$). Table 3 lists the estimated value of ψ for all eight settings. The results show that the estimated values are close to the true values.

Table 3: Simulation study: estimation of the Gibbs parameter ψ

	$\psi = 0$		$\psi = 0.5$		$\psi = 0.9$	
	$\sigma_s^2 = 1$	$\sigma_s^2 = 10$	$\sigma_s^2 = 1$	$\sigma_s^2 = 10$	$\sigma_s^2 = 1$	$\sigma_s^2 = 10$
$\sigma_\varepsilon^2 = 10$	-0.011	0.21	0.43	0.43	0.85	0.85
$\sigma_\varepsilon^2 = 50$	-0.028	0.22	0.46	0.51	0.87	0.88

The Estimation Accuracy of Functional-Spatial Pattern. In Figure 1, we present the true cluster patterns along with the estimated cluster patterns under one simulation setting (large noise level and strong dependence). We also compare the simulated $\tau(s)$ to the estimated spatial-global pattern. Our method accurately recovers the simulated time functions and spatial patterns. In Figure 3 and 4, we also provide the cluster trends and the curves assigned to each of the six clusters for the two comparative methods, Mclust and Fclust. The cluster patterns identified using Mclust are similar to the patterns estimated using the method introduced in this chapter with the exception that for Mclust, Clusters 1 and 6 have similar patterns. On the other hand, Fclust fails to identify the cluster trends.

1.6 Classification of Service Accessibility

1.6.1 Preliminaries

Data Source. The location data for financial services were acquired from the Federal Deposit Insurance Corporation (FDIC). In our study we use data starting from 1994 to 2009. We geocoded the site addresses of the FDIC-insured service providers using ArcView - a GIS software provided by ESRI.

Accessibility Measure. One of the main challenges in measuring service accessibility is defining the distance of the residents of a community or small geographic area to the sites in a service network: given the space occupied by a community U and the service locations in the network, $\mathcal{S} = \{s_1, \ldots, s_n\}$, define this distance as $d(U, \mathcal{S})$. In this research, we quantify $d(U, \mathcal{S})$ using a sampling procedure: 1. Sample

the geographic space of the neighborhood U and obtain the neighborhood locations: $u_1, \ldots, u_B$ (B is the number of samples); and 2. Compute $d(U, \mathcal{S})$ as a summary of the street-network distances between the sample locations in U and all neighboring sites in the network $\mathcal{S}$: $d(u_b, s_i)$ for $b = 1, \ldots, m$ and $i = 1, \ldots, n$. In contrast to the existing methods for computing $d(U, \mathcal{S})$ (Talen 1996, 1998, 2001; Lovett et al. 2002), this sampling technique assumes that neighborhoods occupy uneven geographic areas varying in size and shape.

In this chapter, we measure the accessibility from a community U to the network $\mathcal{S}$ as a summary of the street-network distances $\{d(u_b, s_i)\}_{b=1,\ldots,B;i=1,\ldots,n}$ by modifying the accessibility measures discussed in Talen and Anselin (1998). Specifically, we use the travel cost to measure how much a person in a given neighborhood U is required to travel to a service site and compute the distance of neighborhood to a service network as the average travel cost across individual sampling locations in year t,

$$Y(U,t) = \frac{1}{B}\sum_{b=1}^{B}\left(\frac{1}{n}\sum_{i=1}^{n}\left[d(u_b, s_i)I(t_i \leq t)I(d(u_b, s_i) \leq \epsilon)\right]\right) \tag{15}$$

where $I(t_i \leq t) = 1$ if the site s_i has been opened before or at time t, and zero otherwise; and $I(d(u_b, s_i) \leq \epsilon) = 1$ if the distance between the sampling location u_b is within an ϵ distance from the site s_i. The threshold ϵ measures the maximum proximity value for accessibility. In our implementation, we averaged the street-network distance to the closest three sites.

Dividing the geographic space into contiguous spatial units $U_s,\ s = 1, \ldots, S$, where each spatial unit corresponds to a neighborhood (e.g. census tract), the accessibility measures vary across the geographic space: $Y(U_s, t) = Y_s(t)$.

We apply the clustering method introduced in this chapter to service accessibility curves, $Y_s(t)\ s = 1, \ldots, S$, separately for two states, California and Georgia. Using the $jAIC$ criteria for selecting the number of clusters, we identify 9 clusters for California. For Georgia, the AIC values decrease smoothly without a significant change up to maximum of ten clusters. By inspecting the clustering trends for various numbers of

clusters, we concluded that seven clusters capture the prevalent accessibility patterns in Georgia.

In the next section, we discuss a series of plots which summarize the spatial and temporal accessibility patterns. In this chapter, we included the global patterns along with the cluster trends and their mapping to the geographic space. Additional figures are included in the Appendix B - the distribution of the accessibility curves by cluster and the clustering provided by the two comparison methods - Fclust and Mclust. When interpreting the figures in this chapter, one has to bear in mind that *large values for the accessibility measure (high travel cost) correspond to low access to financial service.*

1.6.2 Discussion

Using demographics and neighborhood classifications provided by ESRI - Sourcebook America, we describe the demographic and economic profile of each service accessibility cluster for California and Georgia. We contrast the trends across clusters and comment on their ethnic composition as it is one of the most cited characteristics in service distribution inequities. We also point out potential business opportunities for service providers.

Global time-varying trends. We first highlight that the global accessibility to financial services hasn't changed significantly for California but it has increased slightly for Georgia from 1994 to 2006 but deteriorate again after 2007 possibly due to the financial crisis (Figures 5 (a) and 6 (a)). The average global accessibility over the period under study (1994-2009) is twice higher in California than in Georgia.

California. Cluster 3 consists of more than 83% of the total number of communities in California. For this cluster, the financial service accessibility cluster trend is approximately flat. Therefore, more than 83% of the communities in California have experience insignificant change in their access to financial services in the past

15 years although some are under-served whereas many others are over-served. By simply overlapping the green color in Figure 5 (d) (cluster 3) to red color in Figure 5 (c) (low accessibility), a financial service provider may identify many of these under-served communities; augmented attention to these communities will not only enhance the equity of financial service distribution but will also open new investment opportunities to service providers.

The communities that form clusters 1, 4, 8 and 9 feature very low accessibility to financial services. Cluster 9 have experienced a sharp increase in accessibility (decrease in the travel cost) with a peak in 1997 followed by a sharp decrease. On the other hand, cluster 8 has experienced and increase in accessibility until 1997 followed by no significant change afterwards. The census tracts in cluster 8 are near National Forests (Klamath, Modoc and San Bernardino) and some regions in Los Angeles. Many of the communities in cluster 9 are located near the National Parks of California (Death Valley, Mojave and surrounding regions of National Forests) and some regions in Los Angels. The change in service accessibility in 1997 may have been caused by the tropical storms that hit northern California from late December 1996 to early January 1997. The Klamath National Forest experienced its worst flood since 1974. This may have been accompanied by the 1997 Asian economic crisis in Los Angels. The changes in financial service accessibility are much smoother for cluster 1 and 4 than those for cluster 8 and 9.

The ethnic/racial composition of the communities in clusters 1 and 4 are similar to the overall demographic trends in California; on the other hand, communities in cluster 8 have predominantly white population whereas many communities in cluster 9 have large Hispanic population. Moreover, the average income level is significantly higher for communities in clusters 1 and 4 as compared to those in clusters 8 and 9. We therefore conclude that the demographic profile is not a significant driving factor for changes in financial service accessibility for the communities in these four

clusters. Particulary, since the communities in clusters 1 and 4 are under-served with a decreasing access to financial services, they offer notable business opportunities for service providers.

The communities in clusters 5 and 6 feature medium accessibility to financial services along with high income level, medium-high population density and ethnic/racial diversity similar to the overall diversity trend in California. The communities in these two clusters have very similar demographic profiles. Following the ACORN classification provided by ESRI (http://www.caci.co.uk/acorn/), the demographic composition in these 17 communities extends to prosperous baby boomers, thriving immigrants, southeastern families, successful suburbanites among others. For these communities, the access to financial services has increased over the past 15 years - a more sluggish increase for cluster 5 although with higher population density than cluster 6. The communities in these two clusters are examples of thriving communities for which the increase in access to financial services has matched the economic opportunities.

The communities in clusters 2 and 7 feature medium-low accessibility to financial services, medium income level and medium population density. Communities in cluster 2 have higher hispanic population than the overall percentage in California whereas some communities in cluster 7 have higher white population percentage. Although the overall access to financial services is lower for cluster 2 than for cluster 7, the time-varying trend in cluster 2 is a slow decrease in accessibility with a slight increase in the most recent years whereas the time-varying trend in cluster 7 is a sharp increase in accessibility. Therefore, although the communities in these two clusters present similar economic potential they have experienced different access to financial services over the past 15 years - a lower access of the communities in cluster 2 also with a higher hispanic population. Similar to clusters 1 and 4, the communities in cluster 2 are potential opportunity markets for financial service providers.

Georgia. Cluster 2 consists of more than 65% of the communities in Georgia. This cluster consists of rural and urban communities, and varying demographic profiles. Similarly to California, a large number of communities in Georgia have experienced insignificant change in accessibility to financial services.

The communities in clusters 1, 6 and 7 generally have very low accessibility to financial services (under-served). Cluster 1 and 7 feature low density population and low income whereas cluster 6 features medium density population, medium income and predominant white population. The accessibility has increased for cluster 6 but decreased in cluster 1; the accessibility for communities in cluster 7 has experienced multiple changes over the past 15 years ending in an accessibility slightly higher than in 1994.

The communities in clusters 3, 4 and 5 have medium accessibility to financial services although the demographic profiles is medium income and medium-high density population. Over the past 15 years, for all three clusters, the access to financial services has increased with a sharper increase for cluster 5. Following the ACORN classification provided by ESRI, the demographic composition in these communities extends to middle class, young frequent movers and rural industrial workers. All three clusters have a higher percentage of white population as compared to the overall ethnic/racial composition in Georgia; communities in cluster 5 have predominant white population.

1.7 Conclusions

The spatial-functional clustering method in this chapter is a means for summarizing the spatial global and time-dependent cluster effects of a large number of spatially-dependent functionals. An alternative method to summarizing the global and local trends in a spatial-temporal process is to fit a spatial-temporal model and analyze the spatial changes over multiple maps that vary with time and the temporal changes that

vary with space. However, this approach is tedious as it requires contrasting multiple maps and multiple time profiles. On the other hand, the proposed clustering method offers readily interpretable summaries of the temporal changes in the spatial-temporal process.

From a methodological point of view, one important aspect of our clustering model is that it allows for spatial dependence in the complete data (Y, Z). Under this assumption, the spatial and temporal trends are accurately estimated by borrowing information across curves in the nearby locations. In our simulation study, we found that accounting for spatial dependence results in enhanced prediction accuracy of the cluster membership under sparse temporal grid and under low noise level, a difficult statistical setting.

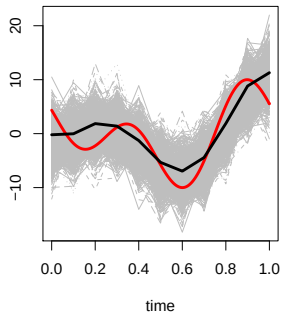
Another aspect of our clustering model is that it is computationally efficient. In the estimation model, we overcame two computational challenges in clustering a large number of spatially-dependent curves: we employ a low-rank approximation of the large kernel matrix and we allow automatic estimation of a large number of penalty parameters by estimating an equivalent linear mixed-effects models.

In our motivating application, describing financial service accessibility in California and Georgia, we find that over a period of 15 years, there have been a small number of communities in California that have increasing access to financial services (about 12%) matching the increase in their economic potential. For most of the other communities, there have been an insignificant or downward change. That is, the inequities in 1994 have perpetuated into 2009; in fact, these inequities may have accentuated due to a significant shift in demographics throughout California. One the other hand, Georgia have faced more changes than California but for the worse. A large percentage of communities have lower accessibility to financial services in 2009 than in 1994; for 65% of these communities the decrease is low in magnitude and for a small number of communities (3%), the decrease is significant. Noteworthy, most

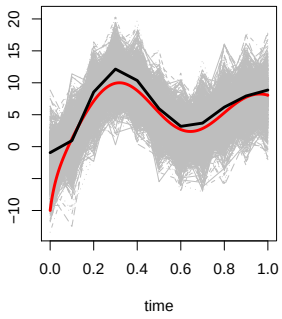
of the communities in Georgia that have experienced an increase in financial service accessibility (22%) have a higher percentage of white population.

We applied two other clustering methods to the service accessibility curves, Mclust and Fclust. These two comparative approaches are commonly employed in clustering multivariate data. The clustering provided by these two methods assign the flat accessibility curves throughout all clusters; therefore, most of the cluster patterns are flat without significant differences between clusters. On the other hand, FCSM assigns most constant curves in one cluster and discovers many more meaningful patterns than the comparative methods.

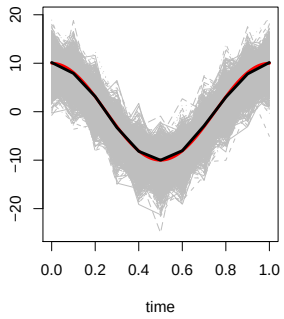
Other potential applications of the clustering approach in this chapter extend to marker segmentation by clustering demand measured at varying site locations as well as performance analysis of a service enterprise that is spatially distributed where the performance may be measure as the sales divided by the site size.



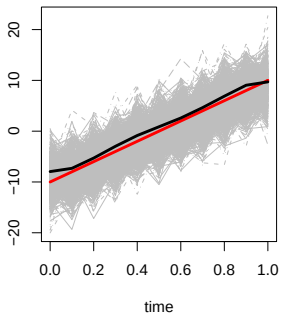
(a) 1st Cluster



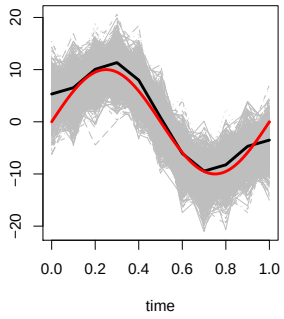
(b) 2nd Cluster



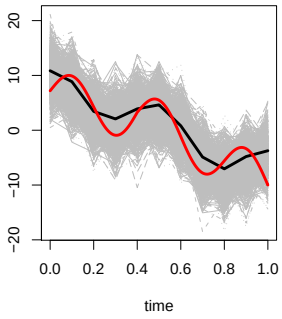
(c) 3rd Cluster



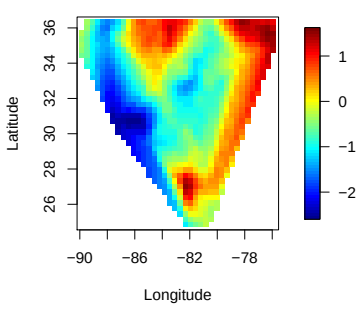
(d) 4th Cluster



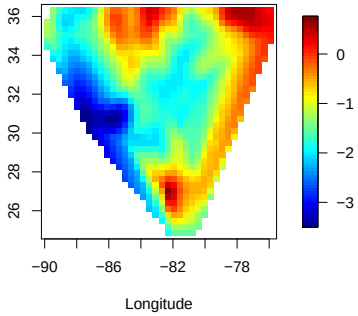
(e) 5rd Cluster



(f) 6th Cluster

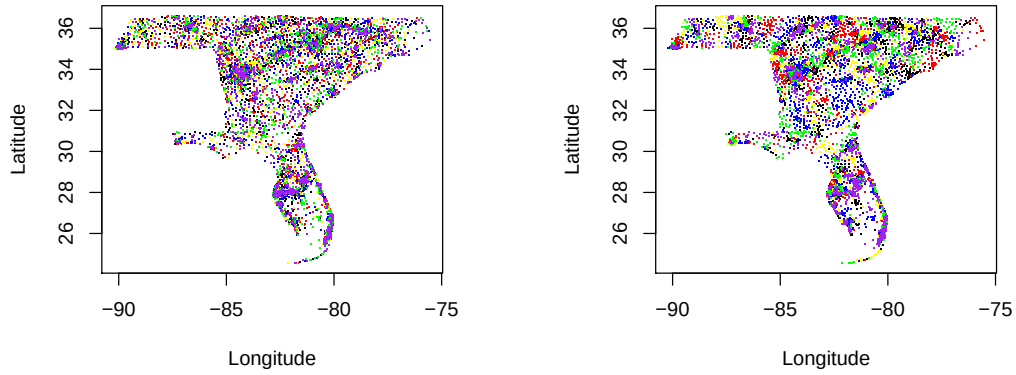


(g) True Spatial Pattern



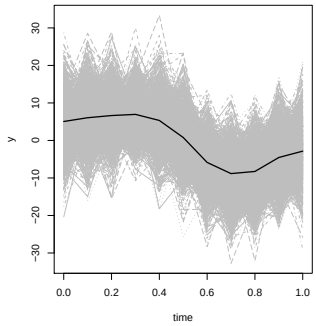
(h) Estimated Spatial Pattern

Figure 1: Simulation setting: $\sigma_s^2 = 1, \sigma_\varepsilon^2 = 10, \psi = 0.9$. (a) to (f) are the functional pattern where the red line the true cluster pattern $\mu_k(t)$; the grey lines are simulated data, $Y(t)$ according to the equation 14 and the black line is the cluster pattern $\hat{\mu}_k(t)$ estimated using FSCM. (g) and (h) are the true and simulated (using our method) spatial effects.

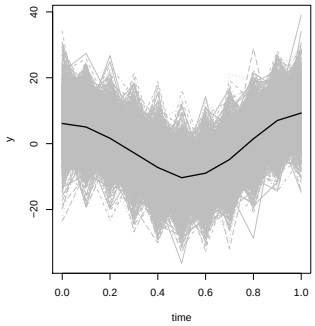


(a) $\psi = 0.5$ (b) $\psi = 0.9$

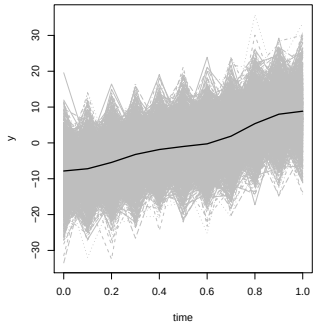
Figure 2: The distribution of the cluster membership generated from Gibbs distribution with $\psi = 0.5$ and $\psi = 0.9$



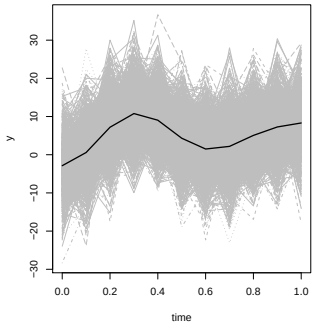
(a) 1st Cluster



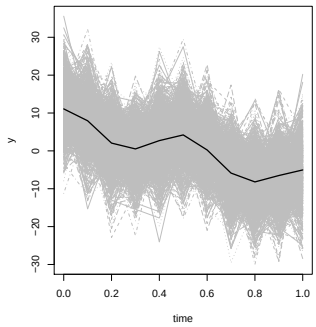
(b) 2nd Cluster



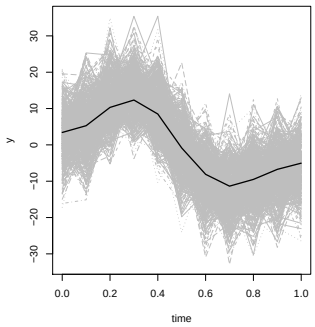
(c) 3rd Cluster



(d) 4th Cluster

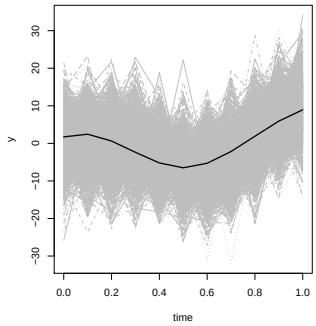


(e) 5th Cluster

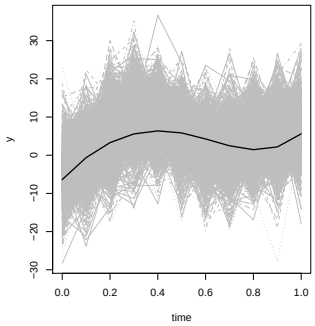


(f) 6th Cluster

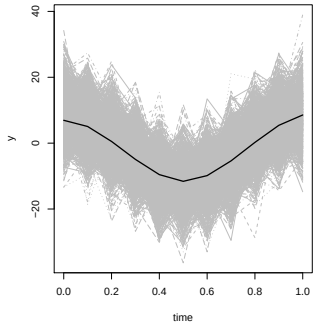
Figure 3: Simulation study: cluster pattern estimated by 'mclust'



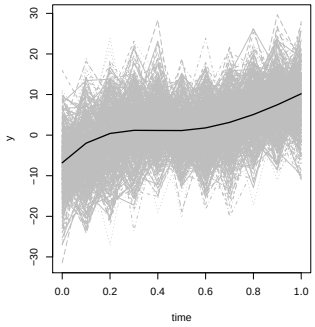
(a) 1st Cluster



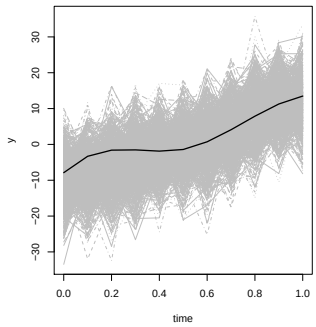
(b) 2nd Cluster



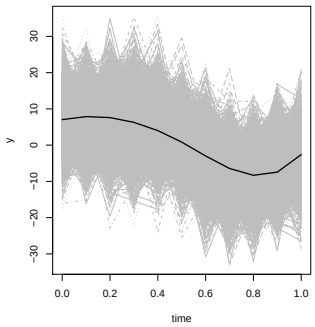
(c) 3rd Cluster



(d) 4th Cluster

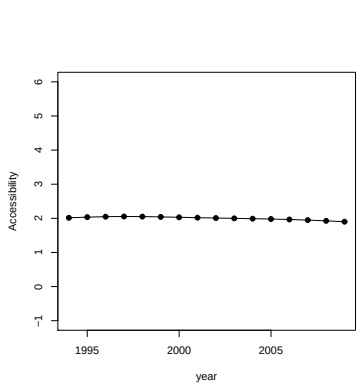


(e) 5th Cluster

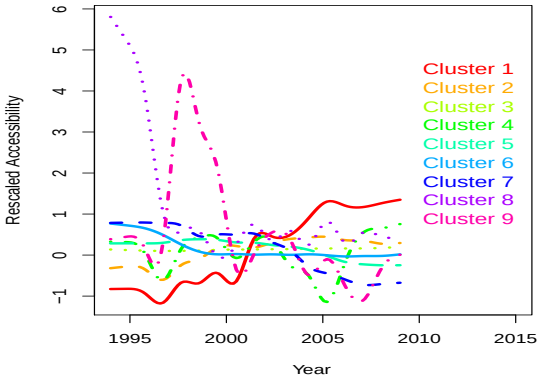


(f) 6th Cluster

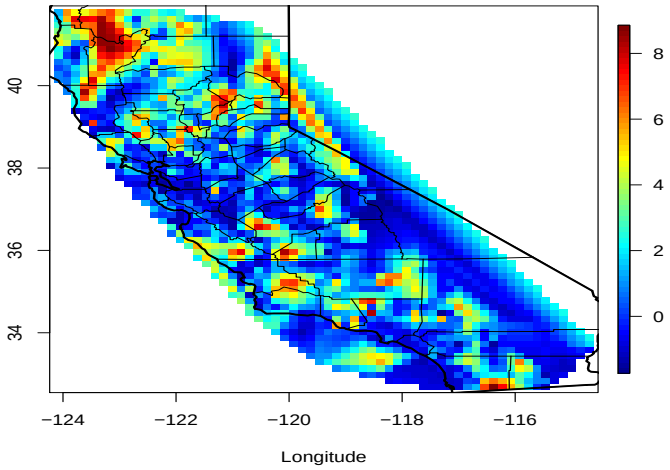
Figure 4: Simulation study: cluster pattern estimated by 'fclust'



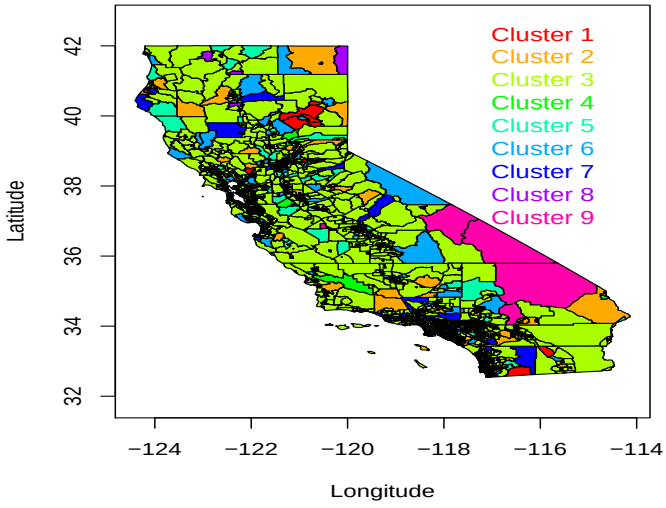
(a) Global Time Trend



(b) Cluster Trends

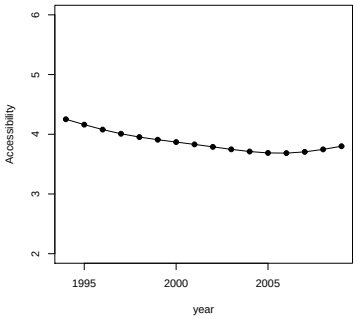


(c) Global Spatial Trend

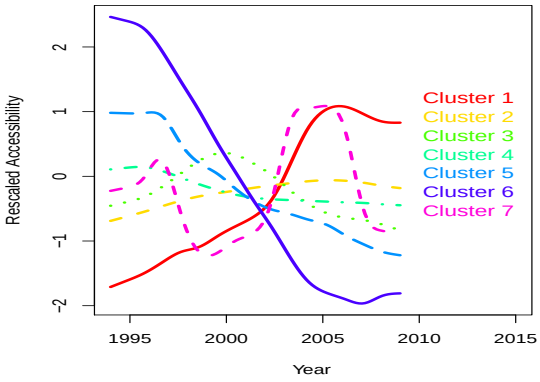


(d) Cluster Map

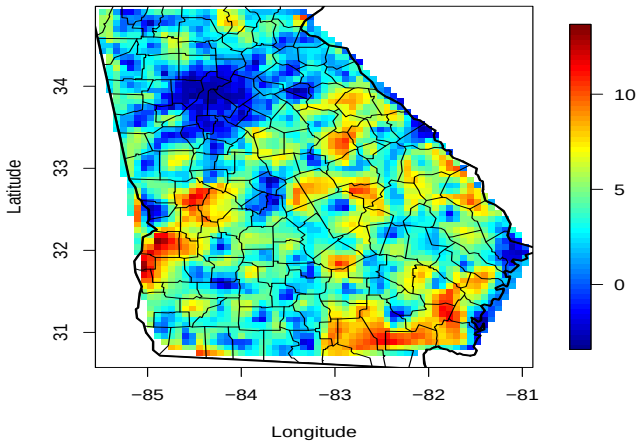
Figure 5: **California**: Temporal and spatial trends for the travel cost used to measure the accessibility of communities to financial services.



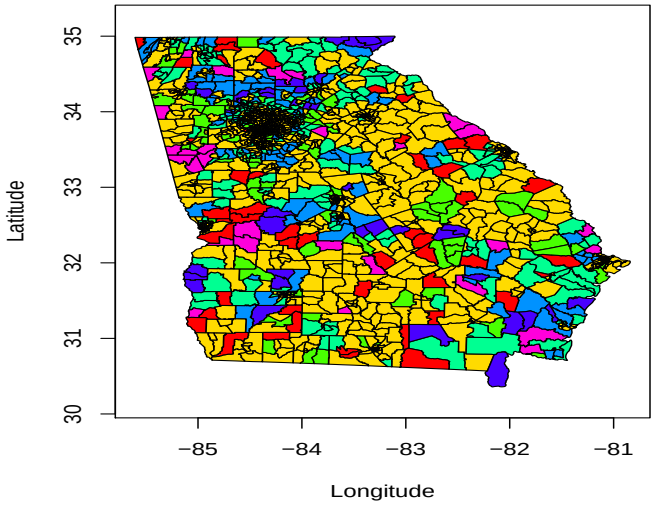
(a) Global Time Trend



(b) Cluster Trends



(c) Global Spatial Trend



(d) Cluster Map

Figure 6: **Georgia**: Temporal and spatial trends for the travel cost used to measure the accessibility of communities to financial services.

CHAPTER II

ASSOCIATION ANALYSIS OF SPACE-TIME VARYING PROCESSES: A FUNCTIONAL APPROACH

2.1 Introduction

In the existing literature, association analysis between two processes extends to time-varying random functions (Heckman and Zamar, 2000; Wang et al, 2000), spatial processes (Lee, 2001; Maruca and Jacquez, 2002; Huang and Zhang, 2006) and time-dependent processes observed longitudinally (He et a., 2004; Dubin and Müller, 2005; Zhou et al., 2008).

In this chapter, we discuss an association analysis for space-time varying processes. We introduce methods for estimating the temporal association at varying spatial locations (*space-varying association*) and the spatial association at varying time points (*time-varying association*) between two processes observed with measurement error. For example, one ad-hoc approach for estimating temporal association is to view the two processes as time-varying functions ($X(s,t) = X_s(t)$ and $Y(s,t) = Y_s(t)$) and estimate the association using a functional data analysis (FDA) approach. Using FDA, we would first smooth out the time-varying functions $X_s(t)$ and $Y_s(t)$, and for each spatial location s, estimate the association at s as the correlation or standardized inner product of the detrended smooth curves - $\rho(s) = \text{cor}\left\{\hat{X}_s(t) - \hat{\mu}_X(t), \hat{Y}_s(t) - \hat{\mu}_Y(t)\right\}$. A similar approach may be employed for spatial association varying in time - $\rho(t) = \text{cor}\left\{\hat{X}_t(s) - \hat{\tau}_X(s), \hat{Y}_t(s) - \hat{\tau}_Y(s)\right\}$.

One primary limitation of the approach described above is that the smoothing step ignores the spatial (temporal) dependence; i.e. $X_s(t)$ are spatially interdependent time-varying functions and $X_t(s)$ are temporally interdependent space-varying

functions. Dependence across functions induces data redundancy which translates into a smaller effective number of degrees of freedom than that number of observations. This will lead to under-estimation of the variability in the data, which in turn, will result in under-smooth association estimates.

In the existing research, the standard approach for estimating the association between two space-time varying processes is to isolate coupled modes of variability between time series or between multivariate spatial processes. Techniques such as combined PCA, maximum covariance analysis (MCA) or canonical correlation analysis (CCA) are common practice (see Bretherton et al., 1992; Storch and Zwiers, 1999; Salim et al 2005 and the references therein). In these methods, two space-time varying processes $X_{ij} = X(s_j, t_i)$ and $Y_{ij} = Y(s_j, t_i)$ observed discretely at m time points and n spatial locations are decomposed as

$$X_{ij} = \sum_{k=1}^{K} u_{kj}\alpha_{ki} \text{ and } Y_{ij} = \sum_{k=1}^{K} v_{kj}\beta_{ki}$$

using single-value decomposition, for example. Further, $U_k = (u_{k1}, \ldots, u_{kn})$ and $V_k = (v_{k1}, \ldots, v_{kn})$ are used to explore the spatial association whereas $A_k = (\alpha_{k1}, \ldots, \alpha_{km})$ and $B_k = (\beta_{k1}, \ldots, \beta_{km})$ are used to explore the temporal association. Time-varying association is measured by the maximum correlation coefficients between the vectors A_k and Y_i called left heterogeneous association and between the vectors X_i and B_k called right heterogeneous association. Similarly, for space-varying association. Therefore, the output consists of multiple right and left time-varying patterns describing the spatial association and multiple right and left space-varying patterns describing the temporal association. Although these exploratory tools are common practice, they have a series of limitations:

• The output consists of multiple time-varying and space-varying association patterns; their interpretation could be tedious and challenging when a large number of canonical components are needed to explain the total variability. Moreover, the number of components/patterns needs to be optimally selected to fully describe the

association between the two space-time varying processes.

• If the processes are slowly varying in space and/or time, we would expect that the spatial components U_k, V_k and the temporal components A_k, B_k have some degree of smoothness. An alternative approach is to take into account the time-functionality by allowing $\alpha_{ki} = \alpha_k(t_i)$ and $\beta_{ki} = \beta_k(t_i)$ to be time-varying functions and to take into account the space-functionality by allowing $u_{kj} = u_k(s_j)$ and $v_{kj} = v_k(s_j)$ to be space-varying functions. Salim et al. (2005) address this challenge using a regularized version of the maximum covariance analysis. To model the spatial dependence, they assume a non-stationary conditional autoregressive model to model the spatial dependence ignoring the time-dependence.

• When the number of time points is much smaller than the number of space points, the existing methods may fail to provide space-varying association estimates due to the computational instability under the setting of large dimensionality but small sample size. Similarly, when the number of space points is much larger than the number of time points, these methods may fail to estimate the time-varying association.

• Finally, the existing association methods for space- and time-varying processes are exploratory in nature without a theoretical and inferential foundation.

The primary contribution of this chapter is an association analysis for space-time varying processes that overcomes these limitations. Our proposed association analysis is based on a spatiotemporal model taking into account the space- and time-functionality in the data. The association measures are rigorously defined, computational feasible, implementable with standard software and they allow estimation of both contemporaneous and lagged association. Finally, we accompany these measures with asymptotic properties and confidence band estimates which may be used to assess the accuracy of the association estimates.

Denote $X(s,t)$ and $Y(s,t)$ with $s \in \mathcal{S}$ and $t \in \mathcal{T}$ two space-time processes observed

with error. We assume that $\mathcal{T}$ and $\mathcal{S}$ are two different Lebesgue measurable domains with no comparable coordinate units. To simplify our presentation, we will introduce the association analysis assuming $\mathcal{S}$ is a geographic space and $\mathcal{T}$ is a time domain but the method applies to general $\mathcal{S}$ and $\mathcal{T}$.

For studying the association between two processes $X(s,t)$ and $Y(s,t)$, we define two inner products: $<,>_{\mathcal{S}}= \int dw_s$ is the inner product between two functions over the space domain $\mathcal{S}$ with respect to a measure $dw_s = w_s(s)ds$ where $w_s(s)$ is a nonnegative weight function with $< 1,1 >_{\mathcal{S}}= \int w_s(s)ds = 1$. Similarly, define $<,>_{\mathcal{T}}= \int dw_t = \int w_t(t)dt$ is the inner product between two functions over the temporal domain $\mathcal{T}$. A simple choice for the weight functions is $w_t(t) = \frac{1}{Length(\mathcal{T})} I_{[\mathcal{T}]}$ and $w_s(s) = \frac{1}{Area(\mathcal{S})} I_{[\mathcal{S}]}$. We focus on global and local association measures:

- The *global temporal association* at lag η is defined as

$$\rho_{gT,\eta} = \rho_T(\mu_x(t), \mu_y(t+\eta)) = \frac{< \mu_x(t), \mu_y(t+\eta) >_{\mathcal{T}}}{\|\mu_x(t)\|_{\mathcal{T}} \|\mu_y(t+\eta)\|_{\mathcal{T}}}$$

where $\mu_x(t)$ and $\mu_y(t)$ are global smoothed temporal trends of the two processes.

- The *global spatial association* at lag δ is defined as

$$\rho_{gS,\delta} = \rho_S(\tau_x(s), \tau_y(s+\delta)) = \frac{< \tau_x(s), \tau_y(s+\delta) >_{\mathcal{S}}}{\|\tau_x(s)\|_{\mathcal{S}} \|\tau_y(s+\delta)\|_{\mathcal{S}}}$$

where $\tau_x(s)$ and $\tau_y(s)$ are global smoothed spatial trends of the two processes.

- The *time-varying association* at temporal lag η is defined as

$$\rho(t, t+\eta) =< f_x^*(s,t), f_y^*(s,t+\eta) >_{\mathcal{S}} . \tag{16}$$

- The *spatial-varying association* at spatial lag δ is defined as

$$\rho(s, s+\delta) =< f_x^*(s,t), f_y^*(s,t+\eta) >_{\mathcal{T}} \tag{17}$$

where $f_x^*(s,t)$ is standardized smoothed version of $X(s,t)$ such that $\|f_x^*(s,t)\|_{\mathcal{T}} = \|f_x^*(s,t)\|_{\mathcal{S}} = 1$ and $< f_x^*, 1 >_{\mathcal{T}}=< f_x^*, 1 >_{\mathcal{S}}= 0$; similarly for $f_y^*(s,t)$. We expand on these association measures in Section 2.3.

The time-varying correlation measure reveals the spatial association of the two processes X and Y at varying time points. On the other hand, the space-varying association measure reveals the temporal correlation of the two processes at varying locations in the space domain.

We estimate the time-varying and space-varying correlation measures by first projecting the two processes $X(s,t)$ and $Y(s,t)$ in $L^2(\mathcal{S} \times \mathcal{T})$ into a finite space of dimensionality $n_S \times m_T$. Specifically, we decompose the two processes using a tensor product decomposition

$$\begin{cases} X(s,t) = \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} \gamma_{k,l}^{X} \phi_l(t) \psi_k(s) + \epsilon_x(s,t) \\ Y(s,t) = \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} \gamma_{k,l}^{X} \phi_l(t) \psi_k(s) + \epsilon_y(s,t). \end{cases} \tag{18}$$

where $\{\phi_l(t),\ l = 0, 1, \ldots\}$ and $\{\psi_k(s),\ k = 0, 1, \ldots\}$ are basis of functions in $L^2(\mathcal{T})$ and respectively, in $L^2(\mathcal{S})$. The error terms $\epsilon_x(s,t)$ and $\epsilon_y(s,t)$ are assumed independent. The space-time tensor product decomposition has been previously reviewed by Kyriakidis and Journel (1999) and applied to various case studies (Wood, 2006 and Clarke et al., 2006). Generally, we cannot estimate more parameters than the number of degrees of freedom, thus we actually estimate the projection onto the first $m_T \leq m$ temporal and $n_S \leq n$ spatial basis of functions (m is the number of observation time points and n is the number of the observation space points). The smoothness of the fitted spatial-temporal surface depends on n_S and m_T.

There are two alternatives to control the smoothness of the estimated surfaces. One alternative is to optimally select $n_S << n$ and $m_T << m$ and use ordinary least squares to estimate the model coefficients. This approach introduces modeling bias and it requires solving a two-dimensional optimization problem since we need to select m_T and n_S simultaneously. A second alternative is to use n_S and m_T large enough to reduce the modeling bias but control smoothness by penalizing the model coefficients γ's. Ruppert (2002) empirically suggests that after the minimum necessary number of spline basis functions (n_S and m_T in our notation) has been reached,

the modelling bias is quite small and it can be ignored. Because we use penalized instead of ordinary least squares procedure to estimate the model coefficients, this second approach introduces shrinkage bias. Li and Ruppert (2008) derive theoretical results for the shrinkage bias for penalized splines estimation under the assumption of ignorable modeling bias. The existing theoretical results apply to one-dimensional functions only. In this chapter, we provide asymptotic properties for the shrinkage bias of the model coefficients under the tensor product spatiotemporal model using penalized splines with different spatial and temporal smoothing parameters (Wood, 2006).

Using the finite tensor product decomposition, we can equivalently express the association measures in terms of the coefficients γ_{kl}^X and γ_{kl}^Y for $k = 1, \ldots, n_s$ and $l = 1, \ldots, m_T$. Using the asymptotic properties of the estimated model coefficients, we show that the estimators of the time- and space-varying association measures are asymptotically unbiased and consistent under large m and n (in-fill asymptotics).

Leveraging the new statistical methodology introduced in this chapter, we study the accessibility to financial services for the state of Georgia in the U.S. Historically, income level has been one of the main drivers of inequities in service accessibility. In response to discriminatory practices against low-income neighborhoods, a practice known as redlining, the U.S. regulatory body amended financial service providers through the Community Reinvestment Act (1977) to meet the needs of all demographic groups, including low-income population. To investigate whether the inequities in financial service accessibility have weaken over time, we estimate the time-varying association between service accessibility measured by an utilization-adjusted travel cost and the income level. A decrease towards zero of the time-varying association will indicate decreased income-based inequities in financial service accessibility over time. On the other hand, service providers are particularly interested in identifying service delivery markets in which the accessibility and the income level

move in the opposite direction - either an area of economic growth but with reduced service accessibility or an area of economic decline but potentially overserved. The map of space-varying association allows identifying such markets since it estimates the temporal association at various geographic locations.

The article is organized as follows. Section 4.2 introduces the nonparametric model decomposition for the spatio-temporal processes and the asymptotic properties of this model. Section 2.3 provides the estimation approach for the association measures as well as their asymptotic properties. We illustrate our association measures with simulated data in Section 3.6 to investigate the accuracy of the association estimators. In Section 2.5 we apply the association analysis introduced in this chapter to study the association between service accessibility and income level. Section 2.6 concludes the chapter. Some technical details and proofs are deferred to the Appendix.

2.2 General Model

In this section we introduce a nonparametric modeling procedure for spatiotemporal processes and we discuss its asymptotic properties. This model is further used in deriving the association measures discussed in Section 2.3.

We observe a realization of the process $Y(s,t)$ at discrete spatial and temporal points: $Y(s_j,t_i)=Y_{ij},$ with $s_j \in \mathcal{S}$ and $t_i \in \mathcal{T}$ for $j=1,\ldots,n,\ i=1,\ldots,m.$ We model the spatiotemporal process $Y(s,t)$ using the following tensor product decomposition

$$Y_{ij} = f(s_j,t_i)+\epsilon_{ij},\ j=1,\ldots,n,\ i=1,\ldots,m \tag{19}$$

$$f(s,t) = \sum_{k=0}^{n_S}\sum_{l=0}^{m_T}\gamma_{k,l}\phi_l(t)\psi_k(s) \tag{20}$$

where $\{\phi_0(t),\phi_1(t),\ldots\}$ and $\{\psi_0(s),\psi_1(s),\ldots\}$ are basis of functions in $L^2(\mathcal{T})$ and respectively, in $L^2(\mathcal{S})$. The error terms $\epsilon_x(s,t)$ and $\epsilon_y(s,t)$ are assumed independent. One has to bear in mind, that *although the tensor-product basis functions are separable, the decomposition of* $Y(s,t)$ *is not.*

The smoothness level of the decomposition in (20) is controlled by n_S and m_T. Various methods have been proposed for selection of the optimal smoothness level including cross-validation to minimize the mean square error (the bias and variance trade-off) and penalization of the coefficients (Rice, 2004; Ramsay and Silverman, 2005; and the references therein). In this chapter, we pursue the later method since it is less computational expensive; that is, we adopt a penalized regression approach allowing for separate smoothing penalties for the spatial and temporal dimensions since space and time are incomparable units. Note that the use of the tensor product basis of functions facilitates the separation of the smoothing penalties in time and space dimensions.

2.2.1 Penalized Regression

Let $f_{s|t_i}(s)$ $(=f(s,t_i))$ be a function of s with t_i held constant, and $f_{t|s_j}(t)$ $(=f(s_j,t))$ a function of t with s_j held constant defined as follows

$$\begin{aligned} f_{s|t_i}(s) &= \sum_{k=0}^{n_S} \alpha_{k,i}\psi_k(s), \alpha_{k,i} = \sum_{l=0}^{m_T} \gamma_{kl}\phi_l(t_i) = \Phi_i'\gamma_k \\ f_{t|s_j}(t) &= \sum_{l=0}^{m_T} \beta_{l,j}\phi_l(t), \beta_{l,j} = \sum_{k=0}^{n_S} \gamma_{kl}\psi_k(s_j) = \Psi_j'\gamma_l \end{aligned}$$

where $\Phi_i' = (\phi_0(t_i), \phi_1(t_i), \ldots, \phi_{m_T}(t_i))'$ and $\Psi_j' = (\psi_0(s_j), \psi_1(s_j), \ldots, \psi_{n_S}(s_j))'$; $\gamma_l = (\gamma_{0,l}, \ldots, \gamma_{n_S,l})'$ and $\gamma_k = (\gamma_{k,0}, \ldots, \gamma_{k,m_T})'$.

Following Wood (2006), a natural way of measuring wiggliness of $f(s,t)$ is to use the additive penalty

$$J(f) = \lambda_s \int_t J_s(f_{s|t})dt + \lambda_t \int_s J_t(f_{t|s})ds$$

where the penalty functions $J_s(f_{s|t_i})$ and $J_t(f_{t|s_j})$ control the smoothness of the conditional functions $f_{t|s_j}(t)$ and respectively, $f_{s|t_i}(s)$; and λ_s and λ_t are smoothing parameters. This penalty can be further approximated by

$$J(f_{st}) \approx \lambda_s \sum_{i=1}^{m} h_{t_i} J_s(f_{s|t_i}) + \lambda_t \sum_{j=1}^{n} h_{s_j} J_t(f_{t|s_j})$$

with h_t and h_s are constants of proportionality related to the spacing of the t_i and s_j. Denote $\alpha_i = (\alpha_{1,i}, \ldots, \alpha_{n_S,i})'$ and $\beta_j = (\beta_{1,j}, \ldots, \beta_{m_T,j})'$. It follows that $\alpha_i = (\Phi_i \otimes I_{n_S})\gamma = \tilde{\Phi}_i\gamma$ and $\beta_j = (I_{m_T} \otimes \Psi_j)\gamma = \tilde{\Psi}_j\gamma$. Using the approximation above, we can re-write the penalty function as

$$J(f_{st}) \approx \lambda_s h_t \sum_{i=1}^{m} \gamma'\tilde{\Phi}_i B_s \tilde{\Phi}_i \gamma + \lambda_t h_s \sum_{j=1}^{n} \gamma'\tilde{\Psi}_j B_t \tilde{\Psi}_j \gamma = \lambda_s \gamma' \tilde{B}_s \gamma + \lambda_t \gamma' \tilde{B}_t \gamma.$$

where B_s and B_t are penalty matrices depending on the definition of the penalty functions $J_s(\cdot)$ and $J_t(\cdot)$. Incorporating h_t and h_s into the smoothing parameters, we estimate γ by solving the following objective function

$$\min_{\gamma} \sum_{i=1}^{m}\sum_{j=1}^{n} \|Y_{ij} - B_{ij}\gamma\|^2 + \gamma'(\lambda_s \tilde{B}_s + \lambda_t \gamma' \tilde{B}_t)\gamma \text{ where } B_{ij} = \Psi_j \otimes \Phi_i. \tag{21}$$

2.2.2 Estimation

In this section, we discuss various alternatives to estimation of the nonparametric model discussed in the previous section. Under penalized regression, the coefficients γ are unknown deterministic parameters and are estimated by minimizing the penalized least squares sum in equation (21).

An emerging popular approach is to cast the penalized regression model into a mixed effects model. Under the mixed effects model, $f(s,t)$ becomes a random function. The main advantage of estimating the penalized coefficients under the equivalent mixed effects model is that the smoothing parameters are automatically updated using the estimates of the variance components of the random effects and random errors. Although the representation of the penalized regression as a mixed model is well established (see e.g., Ruppert et al., 2003), most of the literature in this area has relied on the use of function bases which naturally separate into some components identifiable as fixed effects, and others as random effects.

However, the tensor product model (20) doesn't follow this particular separation

into fixed and random effects. Recently, a reparamerization strategy has been suggested (Fahrmeir et al., 2004; Wood, 2006) to decompose the vector of regression coefficients γ into an unpenalized part (the fixed effects) and a penalized part (the random effects). In our model notation, this reparameterization strategy assumes that the modified penalty matrix $\lambda_s\tilde{B}_s + \lambda_t\tilde{B}_t$ is rank deficient. When the penalty matrix is full rank, this reparametrization reduces to a random coefficient model because all the γs are effectively penalized, thus random effects.

Apart from the frequentist perspective, the penalized regression approach is equivalent to a Bayesian model where γ is assumed random with prior distribution specified by the penalty function (Silverman, 1985; Wahba, 1990; Fahrmeir et al, 2004). Under the Bayesian framework and the assumption of normality, the posterior likelihood is equivalent to the the penalized least squares objective function in (21). The conceptual difference is that the solution to the objective function in (21), the penalized estimator $\hat{\gamma} = (B'B + \lambda_s\tilde{B}_s + \lambda_t\tilde{B}_t)^{-1}B'Y$, is the Best Linear Unbiased Predictor (BLUP) of γ under the penalized regression and the MAP (maximum a posteriori) estimator under the Bayesian framework (Silverman, 1985) when γ are assumed to have an 'almost' normal prior distribution

$$f(\gamma) \sim \exp[-\frac{1}{2}\gamma'(\tilde{B}_s/\tau_s^2 + \tilde{B}_t/\tau_t^2)^{-1}\gamma],$$

where τ_s and τ_t are parameters controlling the dispersion of the prior. It follows that the negative log-posterior distribution is

$$-\log f(\gamma|Y) \propto (Y - B\gamma)'(Y - B\gamma) + \gamma'(\sigma^2/\tau_s^2\tilde{B}_s + \sigma^2/\tau_t^2\tilde{B}_t)^{-1}\gamma.$$

Compared with the penalized objective function (21), when the smoothing parameter $\lambda_s = \sigma^2/\tau_s^2$ and $\lambda_t = \sigma^2/\tau_t^2$, the penalized estimator is equivalent to the MAP estimator. We will use this model equivalence in deriving approximate credible intervals for the association measures in Section 2.3.3.

2.2.3 Asymptotics

In our theoretical study, we assume that the model (20) is the true model, i.e., n_S and m_T are the intrinsic dimensionality of the processes and there is no model bias; or although the model is not true, i.e., we estimate infinite dimensional functions using finite number of basis functions n_S and m_T, the modeling bias is ignorable when n_S and m_T is large enough. Therefore, in this section, we only asymptotic properties for the *shrinkage bias* of the penalized estimator of γ under the following assumptions.

A.1: The smoothing bases for the temporal and spatial domains are knots-based

$$\phi_l(t_i) = K_T(t_i - \kappa_l^T), \psi_k(s_j) = K_S(s_j - \kappa_k^S)$$

where K_T and K_S are temporal and respectively, spatial kernels, and $\kappa_1^T, \ldots, \kappa_{m_T}^T$ and $\kappa_1^S, \ldots, \kappa_{n_S}^S$ are the knots spanning the temporal and spatial domains.

A.2: The number of time points and the number of the spatial points are large: $m \longrightarrow \infty$ and $n \longrightarrow \infty$ under regularly observed space and time domains or the maximum distance between any two design points tends to zero under irregular sampled domains (in-fill asymptotics).

A.3: The distance between any two temporal-knots and any two spatial-knots are bounded from above: $|\kappa_l^T - \kappa_{l'}^T| > d_T$ and $\|\kappa_k^S - \kappa_{k'}^S\| > d_S$; where d_T and d_S are away from zero.

Under assumption A.1, the number of temporal and spatial knots control the roughness of the fit of the model, and therefore, they need to be selected optimally. One method to overcome the knots selection problem is to impose constrains on the effects $\alpha_{k,i}$ and $\beta_{l,j}$,

$$\sum_{i=1}^{m}\sum_{k=1}^{n_S}\alpha_{i,k}^2 < C_1, \quad \sum_{j=1}^{n}\sum_{l=1}^{m_T}\beta_{j,l}^2 < C_2 \quad \text{for some choice of } C_1 \text{ and } C_2.$$

The penalized least squares problem then becomes

$$\min_{\gamma} \sum_{i=1}^{m}\sum_{j=1}^{n} \|Y_{ij} - (\Psi_i \otimes \Phi_i)\gamma\|^2 + \lambda_s \sum_{i=1}^{m} \gamma' \tilde{\Phi}_i' \tilde{\Phi}_i \gamma + \lambda_t \sum_{j=1}^{n} \gamma' \tilde{\Psi}_j' \tilde{\Psi}_j \gamma$$

and therefore, the penalty matrices B_s and B_t are identity matrices.

THEOREM 1. *Under assumptions A.1-A.3, we have*

(a.) The penalized estimator $\hat{\gamma} = (B'B + \lambda_s \tilde{B}_s + \lambda_t \tilde{B}_t)^{-1} B'Y$ *is biased with bias*

$$\mathbb{B}(\hat{\gamma}) = -[I + (\lambda_s \tilde{B}_s^{-1} + \lambda_t \tilde{B}_t^{-1})^{-1}]^{-1}\gamma.$$

- *As the temporal sample size* $m \to \infty$, $\mathbb{B}(\hat{\gamma}) \to -\lambda_s \tilde{B}_s^{-1}\gamma$.
- *As the spatial sample size* $n \to \infty$, $\mathbb{B}(\hat{\gamma}) \to -\lambda_t \tilde{B}_t^{-1}\gamma$
- *As both* $m \to \infty$ *and* $n \to \infty$, $\mathbb{B}(\hat{\gamma}) \to 0$

(b.) The variance of the penalized estimator is

$$\mathbb{V}[\hat{\gamma}] = \sigma_\epsilon^2 (B'B + \lambda_s \tilde{B}_s + \lambda_t \tilde{B}_t)^{-1} B'B (B'B + \lambda_s \tilde{B}_s + \lambda_t \tilde{B}_t)^{-1}.$$

- *As the temporal sample size* $m \to \infty$, $\mathbb{V}(\hat{\gamma}) \to \sigma_\epsilon^2 [B'B + 2\lambda_s \tilde{B}_s + \lambda_s^2 (\tilde{B}_s \tilde{B})_t^{-1}]^{-1}$.
- *As the spatial sample size* $n \to \infty$, $\mathbb{V}(\hat{\gamma}) \to \sigma_\epsilon^2 [B'B + 2\lambda_t \tilde{B}_t + \lambda_t^2 (\tilde{B}_t \tilde{B})_s^{-1}]^{-1}$.
- *As both* $m \to \infty$ *and* $n \to \infty$,$\mathbb{V}(\hat{\gamma}) \to \sigma_\epsilon^2 (B'B)^{-1}$.

According to Theorem 1, as the sample size $m \to \infty$ or $n \to \infty$, the bias of $\hat{\gamma}$ decrease while the variance of $\hat{\gamma}$ increase. As both $m \to \infty$ and $n \to \infty$, the bias goes to zero and the variance tends to the ordinary least squares variance which is the variance under no penalization.

2.3 Association Analysis

In this section, we introduce global as well as time-varying and space-varying association estimators for spatial-temporal processes $X(s,t)$ and $Y(s,t)$ with $s \in \mathcal{S}$ and $t \in \mathcal{T}$. The association estimators are derived using the model decomposition described in Section 4.2. Define

$$\begin{cases} X(s,t) = f_X(s,t) + \epsilon_X(s,t), f_X(s,t) = \sum_{k=0}^{n_S}\sum_{l=0}^{m_T} \gamma_{k,l}^X \psi_k(s)\phi_l(t) \\ Y(s,t) = f_Y(s,t) + \epsilon_Y(s,t), f_Y(s,t) = \sum_{k=0}^{n_S}\sum_{l=0}^{m_T} \gamma_{k,l}^Y \psi_k(s)\phi_l(t). \end{cases} \tag{22}$$

2.3.1 Estimation

We define the association measures using the standardized versions $f_X^*(s,t)$ and $f_Y^*(s,t)$ of f_X and f_Y. Ignoring the index X and Y, the standardized function $f^*(s,t)$ is

$$f^*(s,t) = \frac{\breve{f}(s,t) - \breve{R}_s}{\|\breve{f}(s,t) - \breve{R}_s\|_{\mathcal{T}}},$$

where $\breve{f}(s,t) = f(s,t) - < f(s,t), 1 >_{\mathcal{S}}$ and $\breve{R}_s =< \breve{f}(s,t), 1 >_{\mathcal{T}}$. The same standardization applies when we first subtract the spatial trend, $\bar{f}(s,t) = f(s,t) - < f(s,t), 1 >_{\mathcal{T}}$, and then subtract the temporal scaling factor $\bar{R}_t =< \bar{f}(s,t), 1 >_{\mathcal{S}}$.

For $\{1, \phi_1(t), \ldots, \phi_{m_T}(t)\}$ and $\{1, \psi_1(s), \ldots, \psi_{n_S}(s)\}$ orthonormal bases spanning $\mathcal{T}$ and $\mathcal{S}$ respectively, we define three scaling terms as follows

- The *space-varying* scaling factor $< f(s,t), 1 >_{\mathcal{T}} = \beta_{0,s} = \gamma_{0,0} + \sum_{k=1}^{n_S} \gamma_{k,0}\psi_k(s)$;
- The *time-varying* scaling factor $< f(s,t), 1 >_{\mathcal{S}} = \alpha_{0,t} = \gamma_{0,0} + \sum_{l=1}^{L} \gamma_{0,l}\phi_l(t)$;
- The *static* scaling factor $<< f(s,t), 1 >_{\mathcal{S}}, 1 >_{\mathcal{T}} =<< f(s,t), 1 >_{\mathcal{T}}, 1 >_{\mathcal{S}} = \gamma_{0,0}$.

The detrended and scaled function becomes

$$\breve{f}(s,t) - \breve{R}_s = \bar{f}(s,t) - \bar{R}_s = \sum_{l=1}^{m_T}\sum_{k=1}^{n_S} \gamma_{k,l}\psi_k(s)\phi_l(t)$$

Following these derivations, the association measures defined as the angle between two standardized surfaces become

- Global temporal association:

$$\rho_{gT} = \frac{< \mu_x(t), \mu_y(t) >_{\mathcal{T}}}{\|\mu_x(t)\|_{\mathcal{T}}\|\mu_y(t)\|_{\mathcal{T}}} = \frac{\sum_{l=1}^{L} \gamma_{0,l}^X \gamma_{0,l}^Y}{\sqrt{\left\{\sum_{l=1}^{L}(\gamma_{0,l}^X)^2\right\}\left\{\sum_{l=1}^{L}(\gamma_{0,l}^Y)^2\right\}}}.$$

where $\mu_.(t) =< f_.(s,t), 1 >_{\mathcal{S}} - \gamma_{0,0}^{\cdot}$ are temporal global trends.

- Global spatial association:

$$\rho_{gS} = \frac{< \tau_x(s), \tau_y(s) >_{\mathcal{S}}}{\|\tau_s(s)\|_{\mathcal{S}}\|\tau_y(s)\|_{\mathcal{S}}} = \frac{\sum_{k=1}^{K} \gamma_{k,0}^X \gamma_{k,0}^Y}{\sqrt{\left\{\sum_{k=1}^{K}(\gamma_{k,0}^X)^2\right\}\left\{\sum_{k=1}^{K}(\gamma_{k,0}^Y)^2\right\}}}.$$

where $\tau_.(s) =< f_.(s,t), 1 >_{\mathcal{T}} - \gamma_{0,0}^{\cdot}$ are spatial global trends.

• Local spatial association (*time-varying association*):

$$\rho(t, t+\eta) =< f_X^*(s,t), f_Y^*(s,t+\eta) >_{\mathcal{S}} \tag{23}$$

$$= \frac{\sum_{k=1}^{n_S} \left\{\sum_{l=1}^{m_T} \gamma_{k,l}^X \phi_l(t)\right\} \left\{\sum_{l=1}^{m_T} \gamma_{k,l}^Y \phi_l(t+\eta)\right\}}{\sqrt{\sum_{k=1}^{K} \left\{\sum_{l=1}^{L} \gamma_{k,l}^X \phi_l(t)\right\}^2} \sqrt{\sum_{k=1}^{K} \left\{\sum_{l=1}^{L} \gamma_{k,l}^Y \phi_l(t+\eta)\right\}^2}}$$

• Local temporal association (*space-varying association*):

$$\rho(s, s+\delta) =< f_X^*(s,t), f_Y^*(s+\delta,t) >_{\mathcal{T}} \tag{24}$$

$$= \frac{\sum_{l=1}^{m_T} \left\{\sum_{k=1}^{n_S} \gamma_{k,l}^X \psi_k(s)\right\} \left\{\sum_{k=1}^{n_S} \gamma_{k,l}^Y \psi_k(s+\delta)\right\}}{\sqrt{\sum_{l=1}^{L} \left\{\sum_{k=1}^{K} \gamma_{k,l}^X \psi_k(s)\right\}^2} \sqrt{\sum_{l=1}^{L} \left\{\sum_{k=1}^{K} \gamma_{k,l}^Y \psi_k(s+\delta)\right\}^2}}$$

Plugging in the estimates of γ discussed in Section 2.2.2, we get the estimated measures $\hat{\rho}(s)$ and $\hat{\rho}(t)$. We note here that the formulas described above are derived under the assumption of orthonormal and orthogonal (after some rescaling) bases of functions over both the time domain $\mathcal{T}$ and the space domain $\mathcal{S}$. While for one-dimensional classes of functions there are many orthogonal bases, for two- and higher dimensional spaces, non-orthogonal bases of functions have been used. Under non-orthogonality, the association measures do not have explicit formula, and thus we have to evaluate them by numerical integration. This approximation relies on the assumption of densely sampled domains.

2.3.2 Asymptotics

Using the asymptotic properties of the penalized estimator of γ, we derive the asymptotic properties of $\hat{\rho}(t, t+\eta)$ and $\hat{\rho}(s, s+\delta)$ described in Theorem 2. The proof is in Appendix B.

THEOREM 2: Under the assumptions A.1-A.3, $\hat{\rho}(t, t+\eta)$ *and* $\hat{\rho}(s, s+\delta)$ *are asymptotically unbiased as* $n \to \infty$ *and* $m \to \infty$

$$\mathbb{E}[\hat{\rho}(t, t+\eta)] \to \rho(t, t+\eta), \quad \textit{and} \ \mathbb{E}[\hat{\rho}(s, s+\delta)] \to \rho(s, s+\delta).$$

Using the notation $D_t = (\frac{\partial \rho(t,t+\eta)}{\partial \gamma_{kl}}, k = 0, \ldots, n_S; l = 0, \ldots, m_T)'$ *and* $D_s = (\frac{\partial \rho(s,s+\delta)}{\partial \gamma_{kl}}, k = 0, \ldots, n_S; l = 0, \ldots, m_T)$*, the variances of the association estimators are approximately*

$$\begin{aligned} \mathbb{V}[\hat{\rho}(t, t+\eta)] &\rightarrow D_t^{x'} \sigma_{\epsilon,x}^2 (B'B)^{-1} D_t^x + D_t^{y'} \sigma_{\epsilon,y}^2 (B'B)^{-1} D_t^y, \\ \mathbb{V}[\hat{\rho}(s, s+\delta)] &\rightarrow Dx'_s \sigma_{\epsilon,x}^2 (B'B)^{-1} D_s^x + Dy'_s \sigma_{\epsilon,y}^2 (B'B)^{-1} Dy_s. \end{aligned}$$

The asymptotic results in Theorem 2 state that the association estimates are asymptotically unbiased. Moreover, the variance estimates of the association measures depend on the error variance of the two processes and the first order derivatives of the association measures; therefore, these variance estimates cannot be used in assessing the accuracy of the association estimates unless we replace the true values with their corresponding estimates. Instead, in this chapter, we assess the accuracy of the association estimates using Monte Carlo simulations as described in the next section.

2.3.3 Interval Estimation

In this section, we provide interval estimates for the association measures by simulating from the posterior distribution in a Bayesian context (Silverman, 1985). Since the association measures, $\rho(t) = G_T(\gamma_x, \gamma_y; t)$ and $\rho(s) = G_S(\gamma_x, \gamma_y; s)$, are nonlinear functions of γ_x and γ_y, their posterior distributions are intractable. A commonly used approach to approximate the posterior cumulative distribution function $F_T(g; t)$ and $F_S(g; s)$ for G_T and G_S is using sampling techniques (Silverman, 1985; Wood, 2006). First, simulate the random vectors $\gamma_{x,b}^*$ and $\gamma_{y,b}^*$, $b = 1, \ldots, B$ from their posterior distribution

$$\gamma | Y \sim N(\hat{\gamma}, \sigma^2 (B'B + \lambda_s \tilde{B}_s + \lambda_t \tilde{B}_t)^{-1}),$$

and then compute the approximated empirical distributions F_T and F_S as

$$\hat{F}_T(g; t) = \frac{1}{B} \sum_{b=1}^{B} I(G_T(\gamma_{x,b}^*, \gamma_{y,b}^*; t) \leq g)$$

where $I(\cdot)$ is the indicator function. The Bayesian confidence intervals of $\rho(t)$ and $\rho(s)$ are obtained from the quantiles of this distribution.

Since a simulated sample may have large ranks at some time/spatial points and small ranks at others, we instead estimate simultaneous $1-\alpha$ confidence bands using the algorithm proposed by Mandela and Betensky (2008) by ranking each sample according to the time/spatial points which are most discrepant from the pointwise medians and then using these ranks to define the simultaneous confidence interval.

2.4 Simulation

Simulation Setting. We generate the penalized effects $\gamma^x_{k,l}$ and $\gamma^x_{k,l}$ from the standard normal distribution but correlated to induce cross-correlation between X and Y

$$cov(\gamma^x_{k,l}, \gamma^y_{k',l'}) = \begin{cases} 0.5, & k = k' \text{ and } l = l'; \\ 0, & k \neq k' \text{ or } l \neq l'. \end{cases}$$

The time points and spatial points are equally spaced. We simulate $X_{ij} = X(s_j, t_i)$ and $Y_{ij} = Y(s_j, t_i)$ using

$$X_{ij} = \sum_{k=1}^{n_S}\sum_{l=1}^{m_T} \gamma^x_{k,l}\phi_l(t_i)\psi_k(s_j), Y_{ij} = \sum_{k=1}^{n_S}\sum_{l=1}^{m_T} \gamma^y_{k,l}\phi_l(t_i)\psi_k(s_j)$$

where $\phi(t_i) = (\phi_0(t_i), \phi_1(t_i), \ldots, \phi_L(t_i))$ form an orthonormal basis,

$$\phi(t_i) = \{1, \sqrt{2}\sin(2\pi t_i), \sqrt{2}\cos(2\pi t_i), \sqrt{2}\sin(4\pi t_i), \sqrt{2}\cos(4\pi t_i), \ldots\}$$

and $\psi(\mathbf{s}_i) = (\psi_0(\mathbf{s}_i), \ldots, \psi m_S(\mathbf{s}_i))$ are constructed through tensor products of $\eta(s_{i,1}) = (\eta_0(s_{i,1}), \eta_1(s_{i,1}), \ldots, \eta_{K_1}(s_{i,1}))$ and $\varphi(s_{i,2}) = (\varphi_0(s_{i,2}), \varphi_1(s_{i,2}), \ldots, \varphi_{K_2}(s_{i,2}))$,

$$\psi(\mathbf{s}_i) = \eta(s_{i,1}) \otimes \varphi(s_{i,2})$$

where $\eta(s_{i,1})$ and $\varphi(s_{i,2})$ are both sine-cosine bases. Then the resulting spatial basis is a set of orthonormal basis spanning $[0,1] \times [0,1]$. The error term is simulated from normal distribution $N(0, \sigma^2_\epsilon)$. To control the noise level, we generate σ^2_ϵ proportional to the mean square of Y_{ij}.

The true temporal and spatial correlation are computed from (23) and (24) since we use orthonormal bases to generate $X(s,t)$ and $Y(s,t)$. However, we use the approximate versions to estimate the association measures $\hat{\rho}(t, t+\tau)$ and $\hat{\rho}(s, s+\delta)$ since in our estimation procedure, we use non-orthogonal basis of functions; specifically, we use the knots-based cubic regression spline for temporal domain and we use low-rank thin plate spline truncated using eigen-decomposition for spatial domain (Wood 2006).

Accuracy. We evaluate the accuracy of the correlation estimates using Average Squared Error for the time-varying correlation measure (ASE_t) and for the space-varying correlation measure (ASE_s) defined as follows

$$ASE_t = \frac{1}{m}\sum_{i=1}^{m}\left[\rho(t_i,t_i) - \hat{\rho}(t_i,t_i)\right]^2, ASE_s = \frac{1}{n}\sum_{j=1}^{n}\left[\rho(s_j,s_j) - \hat{\rho}(s_j,s_j)\right]^2.$$

The true temporal and spatial correlation are computed from (23) and (24) since we use orthonormal bases to generate the spatial-temporal local trends $X(s,t)$ and $Y(s,t)$. However, we use the numerical intergration to estimate the cross-correlation measures $\hat{\rho}(t, t+\tau)$ and $\hat{\rho}(s, s+\delta)$ since in our estimation procedure, we use cubic regression spline and thin plate spline which do not result in orthonormal bases.

In the tables below, we report the average ASE_t and ASE_s over 100 simulation runs. We summarize our findings as follows:

• Table 4 provides the accuracy of the association estimates for varying number of spatial and temporal design points. The results indicate that accuracy improves when the number of spatial/temporal points increases. This empirical study supports our asymptotic results.

• Table 5 shows the accuracy of the association estimates compared to the ad-hoc estimates ignoring the spatial dependence described in the Introduction section. The results indicate that accuracy improves when we account for the spatial dependence in the data.

Table 4: Accuracy of the local association estimates for the simulation model with varying number of temporal points m and spatial points n. The correlation measures are *estimated* using $n_S = 25$ and $m_T = 5$ ($ASE_t \times 10^{-4}$, $ASE_s \times 10^{-2}$)

	n=100	n=400	n=900
m=10	(33.5, 5.09)	(8.59, 2.72)	(5.26, 1.96)
m=20	(17.0, 3.26)	(5.74, 2.10)	(3.51, 1.75)
m=30	(12.9, 2.62)	(4.97, 1.88)	(2.66, 1.66)
m=50	(7.38, 1.92)	(3.21, 1.77)	(2.32, 1.60)

.

Table 5: Compare accuracy ($ASE_s \times 10^{-2}$) of our space-varying correlation estimates (Left) with estimates which ignore the spatial dependence (Right).

	n=100	n=400	n=900
m=10	(5.09,28.8)	(2.72,27.3)	(1.96,27.8)
m=20	(3.26,21.0)	(2.10,24.2)	(1.75,24.4)
m=30	(2.62,21.6)	(1.88, 20.8)	(1.66,22.7)
m=50	(1.92,19.7)	(1.77,19.0)	(1.60,20.0)

• Table 6 presents the coverage probability of 95% simultaneous confidence interval for varying number of spatial and temporal points. The results suggests that the simultaneous confidence interval has a reasonable coverage probability.

2.5 Service Accessibility and Income Level

Historically, income level has been one of the main drivers of inequities in service accessibility. Due to uncertainties in customer economic potential and the most often lack of infrastructure, low and medium income neighborhoods have not received sufficient attention from service providers to match their needs (PolicyLink, 2008). In response to this common practice, Community Reinvestment Act (1977) has been

Table 6: Coverage probability of 95% simultaneous confidence intervals ($B = 500$).

	n=100	n=400
m=10	96.48%, 95.49%	97.16%, 95.84%
m=20	98.04%, 96.31%	97.19%, 94.11%
m=30	98.05%, 96.25%	96.19%, 92.01%

established to ensure extent of financial services to under-served communities, specifically communities with low and medium income level. To this end, this research study assesses how the configuration of the financial service accessibility has changed in the past years for communities at all income levels.

Specifically, we estimate the association between income and service accessibility varying from one community to another and from one year to another. Although the change in this association is almost insignificant within a one year period, we would expect that under adherence to Community Reinvestment Act, this association has moved towards zero over the past years. Because of data scarcity, we only investigate the association between financial service accessibility and income level starting with 1996 to 2006. For brevity of the presentation, we focus on one state in the U.S., Georgia. However, the proposed methodology also applies to larger number of time points and larger geographic spaces.

2.5.1 Data Description

Service Location Data. The service site data in this study were acquired from the Federal Deposit Insurance Corporation (FDIC). The FDIC database provides address information about all regulated financial services but to use these addresses in geospatial data analysis and mapping, we first geocoded them into point locations, latitude and longitude, using the ArcGIS (ESRI) software.

Demographics Data. We use the population counts and per capita income data acquired from Sourcebook America - ESRI. These data are electronically released each year starting with 1996 to present. In this research, we use the census tract database since census tracts are used as proxy for communities. According to the Bureau of Census, census tracts are delineated with local input, and intended to represent neighborhoods. Since the boundaries of census tracts are updated by the Census Bureau every ten years (1980, 1990, 2000, 2010), our dataset includes a change of

boundaries. Bureau of Census provides the so called 'relationship files' to document the revisions of the 1990 to 2000 census tract boundaries. We therefore map the data collected before 1999 to 2000 boundaries using the information in these relationship files.

GIS Network Data. In service research, the distance between a service site and its customers is commonly evaluated using the Euclidean or the Manhattan distance between the centroid of the neighborhood and the location of the closest service site. GIS road network data allows including more realistic route distances. For example, Talen (1998, 2001) uses the street-network distance to compute the distance between the centroid of the neighborhood and the site location. Lovett et al.(2002) use road distance and travel time by car. In this research, we use the *street-network distance*. We acquired highway data as well as a TIGER street-detailed network for Georgia. We evaluated the street network distances based on both networks; we found that both provide similar distance values.

2.5.2 Service Accessibility

In this chapter, we measure service accessibility as the distance from an area or a community to a network of service sites within a geographic area also called *travel cost*. The travel cost is scaled to take into account the population per service rate at a specific location resulting into a 'utilization' adjusted measure for service accessibility.

Population Rate. We acquired population counts at the community (census tract) level which can be used to estimate the population rate varying over a continuum, the spatial domain under study (e.g. Georgia). Specifically, given the population counts and the boundaries of the contiguous areas forming the complete spatial domain we can further dis-aggregate the population counts into point-level data assuming that the population are uniformly distributed within each area or community. The assumption of uniformity is not realistic but it is reasonable as soon as the areal

units are small compared to the complete domain. Using methods for estimating the rate of point spatial processes, e.g. Kernel smoothing (Diggle, 1985), we can obtain a population rate estimate at any location. Denote this estimate $P(s),\ s \in \mathcal{S}$.

Distance to a Network of Services. One of the main challenges in measuring service accessibility is defining the distance of the residents of a community or small geographic area to the sites in a service network: given the space occupied by a community U and the service locations in the network, $\mathcal{S} = \{s_1, \ldots, s_n\}$, define this distance as $d(U, \mathcal{S})$. In the research works so far, the distance of a neighborhood to a network of sites ($d(U, \mathcal{S})$) is calculated as the distance between the centroid of the region U and the sites in the nearby locations in the service network (Lovett et al., 2002, Talen, 1998, Talen, 2001). We quantify $d(U, \mathcal{S})$ using a sampling procedure: 1. Sample the geographic space of the neighborhood U and obtain the neighborhood locations: $u_1, \ldots, u_B$ (B is the number of samples); and 2. Compute $d(U, \mathcal{S})$ as a summary of the street-network distances between the sample locations in U and all neighboring sites in the network $\mathcal{S}$: $d(u_b, s_i)$ for $b = 1, \ldots, m$ and $i = 1, \ldots, n$. In contrast to the existing methods for computing $d(U, \mathcal{S})$, this sampling technique assumes that neighborhoods occupy uneven geographic areas varying in size and shape.

In this research chapter, we measure the accessibility from a community U to the network $\mathcal{S}$ as a population-adjusted summary of the street-network distances $\{d(u_b, s_i)\}_{b=1,\ldots,B;i=1,\ldots,n}$ by modifying the travel cost measure discussed in Talen and Anselin (1998) to adjust for the need at a particular location and to incorporate the proposed distance to a network of services using the sampling technique above. Specifically, we use the travel cost to measure how much a person at a location in a given neighborhood U is required to travel to a service site and compute the accessibility of a neighborhood to a service network in year t,

$$Y(U, t) = \frac{1}{B} \sum_{b=1}^{B} \big(T(u_b, t)^{\beta} W(u_b, t) \big) \tag{25}$$

where $T(u_b, t)$ is the travel cost at the sample location u_b measured as the average

street-network distance to the closest K service sites available at time t (in our study, $K = 3$), $W(u_b, t)$ is the a population-based weight (in our study, it is equal to the population rate divided by the service rate) at location u_b and β is a *distance disutility parameter*. In most of accessibility studies β is arbitrarily selected to be equal to 2. In this chapter, we estimate β by linear regression: $\log(W(u_b, t)) \sim -\log(T(u_b, t))$.

In contrast to the existing research, in this chapter, the distance to a network of services varies not only with space but also with time. It is important to capture the temporal variations because both the demographic composition and the service network configuration change over time. The global time-dependent accessibility trends reflect the overall progress of the equity in service accessibility.

Dividing the geographic space into contiguous spatial units U_s, $s = 1, \ldots, S$ or communities, where each spatial unit corresponds to a neighborhood (e.g. census tract), the accessibility measure varies across the geographic space and time: $Y(U_s, t) = Y(s, t)$ defines the space- and time-dependent accessibility process.

2.5.3 Summary of the Results and Findings

Data Transformation. Because of the normality assumption, we transform both the accessibility measure (utilization-adjusted travel cost) and the income using the log-transformation. In this analysis, we only present our findings using this transformation.

Canonical Correlation Analysis. In the introduction, we briefly describe the existing common approach for estimating association patterns between two spatiotemporal processes. After scaling the processes to mean zero as suggested by Bretherton et al., (1992), this approach involves decomposing each process in a linear combination of variables with maximum covariance (Maximum Covariance Analysis - MCA) or correlation (Canonical Correlation Analysis - CCA). That is, given two processes observed discretely $X = \{X(s_j, t_i) = X_{ij}\}_{i=1,\ldots,m;j=1,\ldots,n}$ and

$Y = \{Y(s_j, t_i) = Y_{ij}\}_{i=1,\ldots,m;j=1,\ldots,n}$ find the first pair of spatial and temporal patterns

1. U_1 and V_1 with $U_1 = X\alpha_1$ and $V_1 = Y\beta_1$ to maximize $\text{cov}(U_1, V_1)$ or $\text{cor}(U_1, V_1)$ over all choices of α_1 and β_1 (S-mode); or

2. α_1 and β_1 with $\alpha_1 = XU_1$ and $\beta_1 = YV_1$ to maximize $\text{cov}(\alpha_1, \beta_1)$ or $\text{cor}(\alpha_1, \beta_1)$ over all choices of U_1 and V_1 (T-mode).

The second pair of spatial and temporal patterns is obtained similarly but under the constrain that they are orthogonal on the first pair.

In Figure 7, we show the first pair of spatial patterns for service accessibility and income. We use the CCA method under T-mode to obtain these patterns. The first pair of spatial patterns captures the global income trend with higher income levels in urban Georgia and lower income levels in rural Georgia and it captures a contrasting accessibility trend from the overall global trend with low access to financial services in urban areas especially the metropolitan Atlanta. Intuitively, we would expect that the first pair of spatial patterns would resemble more the global trends of observed processes as they explain the largest percentage of the total variance (25.5%); however, this is not the case for the accessibility process.

The time-varying association measures are as follows. The time-varying association between the accessibility spatial pattern provided by the first component and the observed income process is shown in red and the time-varying association between the income spatial pattern provided by the first component and the observed accessibility process is shown in green. The latter time-varying association is more meaningful than the former since it reveals the association between the accessibility process and the global income trend; the low negative association implies that high income is associated with high access to services and low income is associated with low access but at a low association level. Note that this finding is only based on the association between one component of the income level and the accessibility process which explains a small percentage of the total variability. Based on this analysis alone

we cannot conclude that there are inequities in financial service accessibility.

Although association techniques such as MCA and CCA are widely used in climatology studies, the interpretation of the results is not straightforward. There are multiple patterns describing the association modes and they are not all meaningful in the context of our application data. The first five components explain about 67% of the total variance. Moreover, because of the large discrepancy between the number of space points and time points, the S-mode association failed to provide the estimates for the space-varying association. In conclusion, using the MCA/CCA approach we cannot address the questions in our study.

Functional-based Association Analysis: Model Specifications. There are several specifications that may impact the accuracy of the association measure estimates - the selection of the spatial and temporal bases of functions and the selection of m_T and n_S.

In our implementation, we used cubic regression (knots-based) for the temporal basis and eigen-decomposition based low-rank thin plane spline for the spatial basis (Wood, 2006). We compared the association estimates for other bases of functions; the estimates do not change significantly for other bases. The association estimates are more accurate when using orthogonal basis because they have close form expressions; however, spatial domain bases of functions are commonly non-orthogonal.

We generally selected a small number of temporal basis functions ($m_T \leq 5$) since we have a small number of time points; the association measures change insignificantly for various values of $5 \leq m_T \leq 8$. However, the association patterns vary with the number of spatial basis functions, n_S. For small n_S, the space-varying association is very smooth. Ruppert (2002) empirically suggests that after a minimum n_S has been reached, the modelling bias is small. On the other hand, the shrinkage bias decreases with n and m as provided my our theoretical results. Therefore, we can only control the modeling bias by using a large enough n_S; note that in our application n_S can be

as large as $n = 1600$. In contrast, the larger n_S is, the more expensive the computation is. Consequently, we need to select n_S for an optimal trade-off between controlling the modeling bias and the computational feasibility. To select n_S, we use a residual-based analysis approach suggested by Wood (2006). The steps in the residual-based procedure are as follows: (i) fit the model and extract the deviance residuals; (ii) fit an equivalent model to the residuals using a substantially increased n_S to see if there is a significant spatial pattern in the residuals that could potentially be explained by a larger n_S. We performed this residual analysis for various numbers of spatial knots ($50 \leq n_S \leq 400$). We found that for $n_S = 300$ and higher the residuals have insignificant spatial pattern left.

In the following summary, our findings are based on association measures estimated using $m_T = 5$ and $n_S = 250$. Our inference is based on 95% simultaneous confidence bands.

Association Analysis: Figures. To analyze the association between service accessibility and income level using the proposed association analysis we investigated the global association along with the global trends, the time-varying association along with simultaneous confidence bands, the space-varying association, and the lagged time-varying association for the state of Georgia and the metropolitan Atlanta. In this manuscript, we included the visual displays for the global spatial trends (Figure 8) to be compared to the 1st patterns in CCA, and the time-varying and space-varying association estimates (Figure 9).

Remark: We interpret these plots as follows. *Large values of the utilization-adjusted travel cost correspond to low access to the network of financial services.* For example, negative association between the utilization-adjusted travel cost and the income level corresponds to positive association between access to financial services and income level.

Time-varying (Spatial) Association: Findings. In order to assess the equity

of service accessibility with respect to income we need to take into account two components, the association between the global spatial trends ($\tau_x(s)$ and $\tau_y(s)$) and the time-varying association (Figure 9) between the two processes.

The global spatial association is equal to -0.188 in Georgia and -0.804 in the metropolitan Atlanta area. This suggests that there is a positive low association between income level and access to financial service overall Georgia but a positive high association in Atlanta. In turn this implies significant inequities in service accessibility with respect to income in Atlanta but not overall Georgia. The primary reasoning behind this rather large difference in the global association estimates for Georgia and Atlanta is that Georgia is predominantly rural with low income population but medium to high access to financial services measured using the population-adjusted travel cost. On the other hand, the metropolitan Atlanta consists of mixed income population with mixed levels of urbanization and ethnicity. Atlanta is an example of the classical post World War II sprawl: the movement of jobs, people, investment, and infrastructure far from metropolitan regions leaving inner-city communities isolated and under-served while consuming unsustainable amounts of resources. High inequities with respect to income level therefore may be a consequence of this urbanization movement.

The time-varying association is close to zero while the confidence band includes the zero line (except for a change around 2001 which may be an artifact due to the change of boundaries). This suggests that the spatial association between service accessibility and income level has changed insignificantly over the period 1996-2006. On the other hand, the time-varying association for Atlanta alone is negative but reaching zero towards the end of the time period. We interpret this association pattern as a decrease in inequities of the financial service accessibility with respect to income level in the Atlanta area although the global association is rather high corresponding to high inequities. The simultaneous confidence band is much wider

than for Georgia since we base our estimates on a smaller number of spatial points; it is mainly below the zero line implying a negative association throughout the years 1996-2006.

Because both the income level and the change in service site configuration change smoothly over time, the lagged (lag =1-2 years) time-varying association for Georgia and Atlanta (not shown here) resemble the contemporaneous association (lag=0). This implies that the change in income level will not bring higher access to services in immediate years; reversely, an increase or decrease in access to financial services will have a delayed effect if any on the income level. The study of the time-varying association suggests that a longer period of time may be necessary to observe significant changes in the equity of financial services with respect to income.

Space-varying Association: Findings. The space-varying association between adjusted travel cost and income level is neither positive (blue) nor negative (red). Particularly, areas of positive association between travel cost and income level (negative association between access to services and income level) correspond to two trends: markets that experience increase in economic potential over the years but lag in access to financial services and markets that face economic decline but with constant or increasing access to services. For example, south Atlanta, Macon and Savannah, the association is predominantly negative where the access to financial services is also low. These are potential areas of growth that lag in service accessibility.

We note here that the simultaneous confidence band derived for the space-varying association measure is rather wide because of the large error variance and small m. Thus the estimates of the space-varying association are not precise; the inference based on confidence bands warrants us in making strong statements based on this association estimate.

2.6 Conclusions and Further Considerations

In this chapter, we introduce a means for summarizing global and local association patterns between random processes which co-evolve over time and space. The association analysis in this chapter allows borrowing information across time and space resulting in more accurate estimates than using ad-hoc approaches which ignore the functionality in the data. We show in the simulation study in Section 3.6 that the estimation error is larger when using the ad-hoc approaches as compared to the association measures introduced in this chapter.

We applied the association measures introduced in this chapter to summarize the spatial and temporal association between per capita income and service accessibility observed at the census tract level in the state of Georgia with a focus on the metropolitan Atlanta. The data are observed irregularly over the geographic space and the number of spatial design points is large. Therefore, the association analysis applies to irregular designs as well as to densely sampled space and time domains, common challenges in analysis of spatio-temporal data.

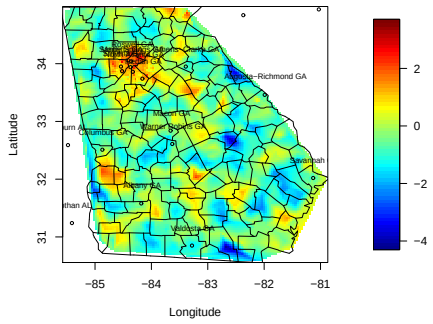
The primary objectives in our case study is to describe the temporal association between the two processes allowing identification of potential new markets for service delivery and to assess the equity of the financial service accessibility with respect to the income level. We find that urban areas including Atlanta particularly its southern communities, Macon and Savannah in Georgia have experienced increasing economic potential measured by the income level but low access to financial services; these areas are potential candidates for new delivery markets.

From the time-varying association analysis, we conclude that there are low inequities overall Georgia which is predominantly rural with low population density but significant inequities in the metropolitan Atlanta which features mixed income levels and high population density. Importantly, our findings are based on an accessibility measure which accounts for the utilization level of the financial services through

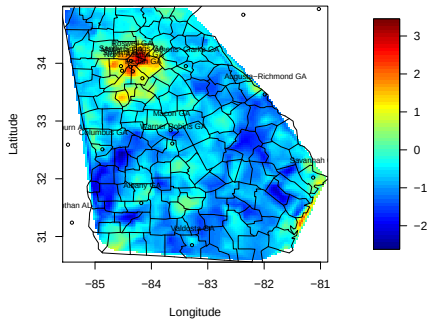
a population-based adjustment. Without this utilization-based adjustment, our findings would have changed. Thus, our equity evaluation depends on the definition of the accessibility measure. In this chapter, we advance the study of the *vertical equity* which accounts for the expected utilization of a service.

Within our case study, we also compare our approach with existing association methods based on maximum correlation analysis. This approach decomposes the association between two space-time varying processes into multiple outputs (spatial and temporal patterns, space and time-varying association measures) which are difficult to interpret. Moreover, this approach fails to provide estimates for the space-varying association measure because there is a large discrepancy between the number of spatial locations and the number of time points.

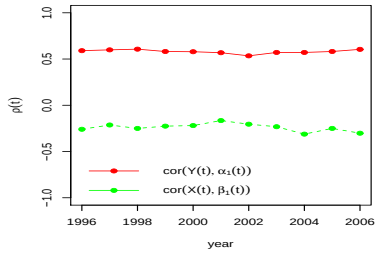
Finally, we complement our association measures with an understanding of their asymptotic properties and with inference based on confidence bands. From our theoretical results, we learn that the association estimates are asymptotically unbiased and consistent as soon as n and m are large. In our application, the number of time points m is small which in turn leads to low accuracy estimates for the space-varying association measure. To improve the accuracy of this association estimate we recommend using a larger number of time points when available. Our simulation study shows a significant increase in the accuracy of the association estimates for $m \geq 20$.



(a) Accessibility - 1st Spatial Pattern

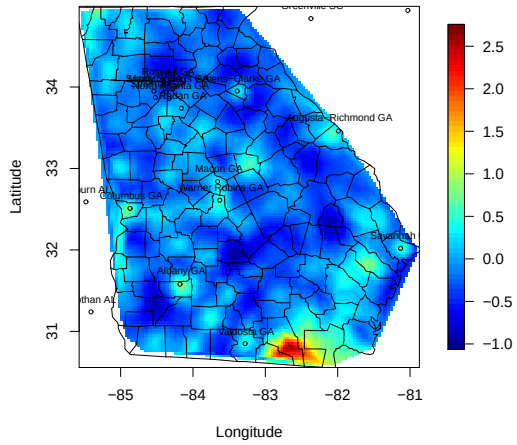


(b) Income - 1st Spatial Pattern

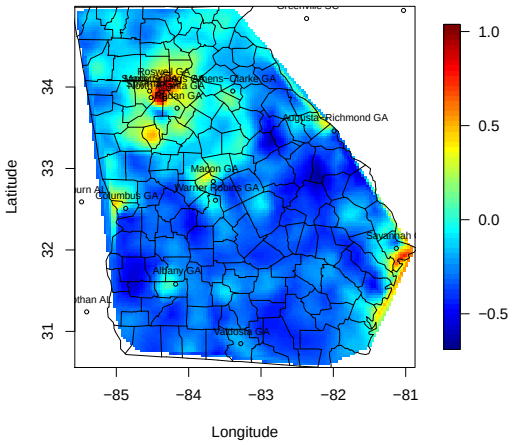


(c) Time-Varying Association

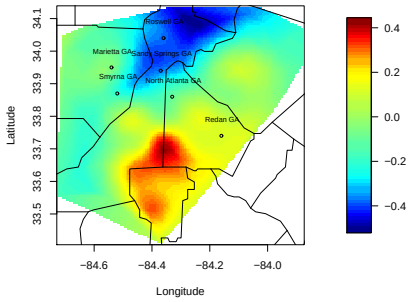
Figure 7: 1st pair of the canonical correlation decomposition; the time-varying association is described by two patterns.



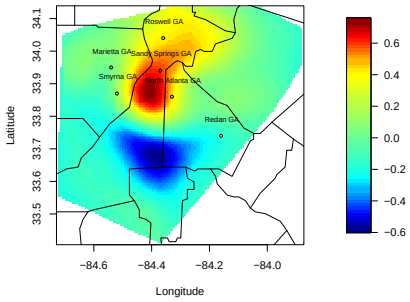
(a) Georgia: log(Accessibility)



(b) Georgia: log(Income)

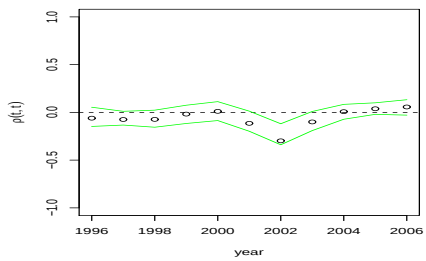


(c) Atlanta: log(Accessibility)

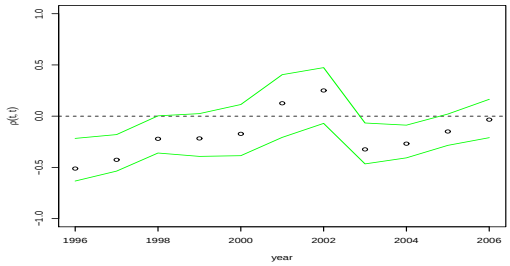


(d) Atlanta: log(Income)

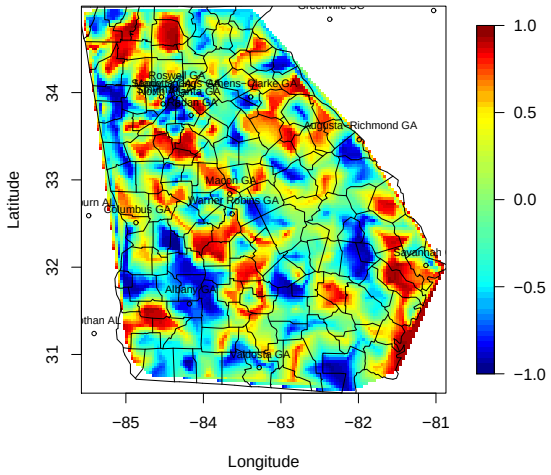
Figure 8: Global spatial trends for Georgia and Atlanta.



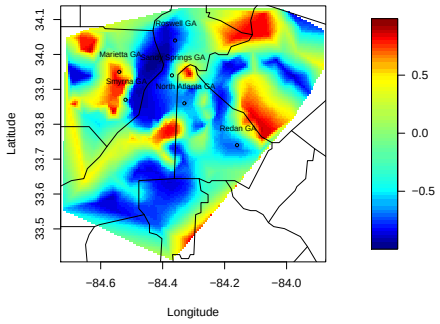
(a) Time-varying Association: Georgia



(b) Time-varying Association:Atlanta



(c) Space-varying Association: Georgia



(d) Space-varying Association: Atlanta

Figure 9: Time-varying measure of the spatial association and space-varying measure of the temporal association for the state of Georgia and the metropolitan Atlanta.

CHAPTER III

MULTI-LEVEL FUNCTIONAL CLUSTERING ANALYSIS

3.1 Introduction

Due to an increasing number of applications requiring analysis of a large number of observed random functions, exploratory tools such as unsupervised or supervised clustering play an important role in uncovering prevalent patterns among the observed random functions. Specific applications include gene expression profiling from microarray studies (Hastie et al., 2000; Bar-Joseph et al., 2002; Wakefield et al., 2002; Serban and Wasserman, 2005; Booth et al., 2008), clustering subjects by their spinal bone mineral density (James and Sugar, 2003), and summarizing the market value trends for manufacturing companies (Serban, 2009) among others.

Functional clustering methods group into hard and soft (model-based) methods. Hard clustering divides the set of functions to be clustered into a partition of non-overlapping subsets according to a similarity measure (e.g. Euclidean distance or correlation). In hard clustering, an observed random function will be assigned to one and only one cluster. On the other hand, in soft clustering, the underlying assumption is that the observed random functions are realizations from a mixture process where the mixture weights are the cluster probabilities. The cluster membership is not fixed as in hard clustering but random following a multinomial distribution. Examples of hard clustering methods are by Hastie et al. (2000), Bar-Joseph et al. (2002); Serban and Wasserman (2005); Chiou and Li (2008). Examples of model-based clustering are by James and Sugar (2003); Fraley and Raftery (2002); Wakefield et al.(2002) and Booth, Casella & Hobert (2008).

In the clustering literature so far, the data are assumed to be observed only at one

level. That is, observe random functions $X_i(t)$ for $i = 1, \ldots, I$ where a clustering is a hard or soft division of the index set $\mathcal{I} = \{1, 2, \ldots, I\}$ into a partition of K subsets. In this chapter, we pursue a more complex problem: clustering data observed at multi-levels. For simplicity, we will focus on two-level data, but the proposed methods extend to more than two levels. Particularly, the statistical problem is to cluster $X_{ij}(t)$ for $j = 1, \ldots, J$ and $i = 1, \ldots, I$ where j indexes the measurement type and i indexes the case type; that is, for each case (e.g. subject, product or gene), we observe J random functions each corresponding to a different measurement. The underlying model is functional ANOVA

$$X_{ij}(t) = \alpha(t) + \beta_j(t) + Y_i(t) + W_{ij}(t) + \varepsilon_{ij}(t) \tag{26}$$

where $\alpha(t)$ and $\beta_j(t)$ for $j = 1, \ldots, J$ are fixed functional means specifying the overall trend, and respectively, between-measurement functional trends. For simplicity, we assume $\alpha(t) = 0$ and $\beta_j(t) = 0$; when non-zero, we can use standard nonparametric methods to estimate them. Under this framework, there are two problems that we pose:

- *Clustering by similarity of within-case means (at level 1)*: two cases i_1 and i_2 are in the same cluster if their within-case means $Y_{i_1}(t)$ and $Y_{i_2}(t)$ are similar in shape.
- *Clustering by similarity of between-case deviations (at level 2)*: two cases i_1 and i_2 are in the same cluster if their corresponding deviations from the within-case means, $\{W_{i_1 j}\}_{j=1,\ldots,J}$ and $\{W_{i_2 j}\}_{j=1,\ldots,J}$, are dynamically similar or they move together over time.

The first clustering problem could be simply carried out by estimating the within-case means $Y_i(t)$ using nonparametric methods and cluster the smooth means using functional clustering algorithms. In this chapter, we call this method the *level-1 naive approach*. A second modeling alternative is to decompose the functional ANOVA model following the multilevel functional principal component analysis (MFPCA)

introduced by Di et al. (2008) and Di and Crainiceanu (2010) and cluster the level-1 estimated scores using common clustering methods such as k-means, hierarchical clustering and others. We call this method the *level-1 hard clustering approach*. The third approach is model-based clustering using the MFPCA decomposition. We call this method the *level-1 model-based clustering approach*.

The second clustering problem can be reduced to estimation of the correlation between two samples of random functions and cluster based on the correlation estimates. For example, apply the dynamical correlation analysis introduced by Dubin and Müller (2005) to each case pair $\{W_{i_1 j}\}_{j=1,\ldots,J}$ and $\{W_{i_2 j}\}_{j=1,\ldots,J}$ to obtain a correlation value ρ_{i_1,i_2} and further apply hierarchical or other distance-based clustering to the correlation values $\{\rho_{i_1,i_2}\}_{i_1=1,\ldots,I;i_2=1,\ldots,I}$. However, this approach assumes large J and large number of time points, assumption that does not hold in many applications. Instead, we can apply the MFPCA approach to the multilevel data and cluster the level-2 estimated scores. We call this *level-2 hard clustering approach*. Similarly to level-1 clustering, an alternative approach is model-based using the MFPCA decomposition. We call this method the *level-2 model-based clustering approach*.

In this chapter, we discuss advantages and disadvantages of these functional clustering approaches and validate their performance within a simulation study. We point out here that one underlying advantage of the model-based approach is that it provides a natural framework for inference on the cluster means and imputed cluster memberships, and it allows incorporating information about the dependence between functions at various levels. However, a drawback is that it is computational intensive as estimation of the model-based clustering is often based on a Expectation-Maximization algorithm.

The rest of the chapter is organized as follows. In Section 3.2, we review the ANOVA functional model and its decomposition using the MFPCA approach. We will continue in Section 3.3 with the description of a series of hard clustering approaches

and in Section 3.4 with the presentation of the model-based clustering method. An important aspect of unsupervised clustering is that the number of clusters is unknown. Under the clustering model described in Section 3.5, we discuss a selection method for the number of clusters in Section 3.5. We assess the performance of the clustering approaches discussed in this chapter within a simulation study in Section 3.6 and within a case study in Section 3.7. Some technical details are deferred to the Appendix.

3.2 Multi-level Functional Model

Let $\{X_{ij}(t),\ j = 1, \ldots, J\}$ be a group of random functions observed over a continuous variable t for the ith case with $i = 1, \ldots, I$ (I is the number of cases). For each experimental case, we observe a set of J random functions, which are functional observations resulting from different types of measurement. Generally, the number of cases I ($I >> 100$'s) is large whereas the number of measurements per subject J is small ($J \sim 2 - 5$).

In this chapter, the underlying model is a functional ANOVA model

$$X_{ij}(t) = \alpha(t) + \beta_j(t) + Y_i(t) + W_{ij}(t) + \varepsilon_{ij}(t) \tag{27}$$

where $t \in \mathcal{T}$ ($\mathcal{T}$ is the functional domain). For brevity of the model description, we assume $\alpha(t) = 0$ and $\beta_j(t) = 0$. Assuming unknown functional effects, we employ a nonparametric decomposition of the model

$$X_{ij}(t) = \sum_{s=1}^{N_1} \xi_{i,s}\phi_s^{(1)}(t) + \sum_{r=1}^{N_2} \zeta_{ij,r}\phi_r^{(2)}(t) + \varepsilon_{ij}(t) \tag{28}$$

where $\{\xi_{i,s}\}_{s=1,\ldots,N_1}$ and $\{\zeta_{ij,r}\}_{r=1,\ldots,N_2,j=1,\ldots,J}$ are the level-1 and level-2 *unconditional scores* for the ith case. In this chapter, we use the term 'unconditional' in contrast to the term 'conditional' which refers to conditionality on the cluster membership variable in the clustering model. In this chapter, we assume

A.1 $\mathbb{E}(\xi_{i,s}) = 0$, $Var(\xi_{i,s}) = \tau_s^{(1)}$ for any case i and $\mathbb{E}(\xi_{i,s_1}\xi_{i,s_2}) = 0$ for $s_1 \neq s_2$.

A.2 $\{\phi_s^{(1)}(t), s = 1, 2, \ldots\}$ is an orthogonal basis in $L^2(\mathcal{T})$.

A.3 $\mathbb{E}(\zeta_{ij,r}) = 0$, $Var(\zeta_{ij,r}) = \tau_{j,r}^{(2)}$ and $\mathbb{E}(\zeta_{ij,r_1}, \zeta_{ij,r_2}) = 0$ for any case i and any measurement type j and for $r_1 \neq r_2$.

A.4 $\{\phi_r^{(2)}(t), r = 1, 2, \ldots\}$ is an orthogonal basis in $L^2(\mathcal{T})$.

A.5 $\{\xi_{i,s},\ s = 1, 2, \ldots\}$ are uncorrelated with $\{\zeta_{ij,r},\ r = 1, 2, \ldots\}$.

There are various approaches for estimating functional ANOVA. Recent methods include Bugli and Lambert (2006), who assume that the bases of functions in A.2 and A.4 are fixed and estimate the scores using penalized splines; Di et al. (2008) and Di and Crainiceanu (2010), who base their estimation procedure on functional principal component analysis, and Kaufman and Sain (2010), who pursue a fully Bayesian approach. An advantage of employing the MFPCA approach is its computational efficiency; the bases of functions are functional principal components which allow reducing the functional space to a lower dimensional space than when fixing the bases of functions as proposed by Bugli and Lambert (2006). Moreover, it applies to both densely observed as well as sparse data, which is an important aspect in functional data analysis. To this end, our clustering model is based on the MFPCA decomposition.

Remark: We note here that assumption A.3 is more restrictive in the MFPCA implementation. Specifically, MFPCA assumes that $Var(\zeta_{ij,r}) = \tau_r^{(2)}$; that is, the variances do not vary with the measurement type j. However, as we will discuss in Section 3.4, the model based clustering is subject to the more general assumption A.3 when the cluster means vary with the measurement type. We will expand on this comment later in the text.

3.3 Alternative Clustering Approaches

3.3.1 Level-1 Clustering

In this section, we describe two alternative approaches to clustering by similarity of within-case means; they are both hard clustering approaches. Generally, in hard clustering, the underlying assumption is that the set of cases to be clustered $\mathcal{I} = \{1, 2, \ldots, I\}$ is divided into a partition of K subsets, $\{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ with $\mathcal{C}_{k_1} \cap \mathcal{C}_{k_2} = \emptyset$ for any k_1 and k_2. Two cases are in the same cluster if they are similar according to a similarity measure (e.g. Euclidean distance, correlation). When the objective is to cluster random functions by shape regardless of scale, the similarity measure is often the correlation between two functions. One common approach to clustering functional data is to first project the functional data from the functional space to a finite dimensional space using nonparametric decompositions, and cluster based on similarity of the transform coefficients. James and Sugar (2003) dubbed this approach as *filtering*. Clustering functions by shape using the correlation measure in the functional domain is equivalent to clustering the transform coefficients using the Euclidean distance in the transform domain (Serban and Wasserman, 2005).

For multi-level functional data, a naive clustering approach is to first decompose the random functions using an orthogonal basis of functions $\{\psi_1(t), \psi_2(t), \ldots\}$:

$$X_{ij}(t) = \sum_{p=1}^{\infty} \theta_{p,ij}\psi_p(t) = \Psi(t)\boldsymbol{\theta}'_{ij}$$

where $\boldsymbol{\theta}_{ij} = (\theta_{1,ij}, \theta_{2,ij}, \ldots)$ is the vector of coefficients of the random functions observed for case i in the transform domain. Since we observe the random functions at a finite number of time points, we need to truncate the summation in the decomposition above. That is, estimate up to P_i coefficients where $P_i < \infty$. In the model formulation above, P_i controls the smoothness of the estimated within-case mean $Y_i(t)$, and therefore, its selection will impact the accuracy of the estimated cluster memberships. Bugli and Lambert (2006) proposed using a large $P_i = P$ to reduce the modeling bias

but penalize the influence of the coefficients - penalized spline smoothing. Further, we cluster the estimated mean coefficients $\hat{\boldsymbol{\theta}}_i = \frac{1}{J}\sum_{i=1}^{J}\hat{\boldsymbol{\theta}}_{ij}$ using common clustering approaches for multivariate data (e.g. hierarchical clustering, k-means).

For densely observed random functions, we expect that this approach will perform reasonably well since the coefficients $\boldsymbol{\theta}_{ij}$ are accurately estimated - $\hat{\boldsymbol{\theta}}_{ij}$ are asymptotically unbiased and consistent. On the other hand, under sparse design (i.e. each random function is observed at a small number of design points), the coefficients $\boldsymbol{\theta}_{ij}$ are inaccurately estimated which in turn, will result in inaccurate clustering membership estimation.

We overcome this difficulty by employing an estimation method which allows borrowing strength across within-case measurements to improve the accuracy of the estimated coefficients for individual cases. Our proposed algorithm for clustering at level 1 is:

1. Apply MFPCA to impute the scores at level 1: $\hat{\xi}_{i,s}$.

2. Apply a multivariate clustering algorithm to the estimated scores $\hat{\xi}_{i,s}$ where the similarity measure is the Euclidean distance ($d(i_1, i_2) = ||\hat{\boldsymbol{\xi}}_{i_1} - \hat{\boldsymbol{\xi}}_{i_2}||^2$ for $i_1, i_2 \in \mathcal{I}$).

This algorithm is equivalent to clustering the within-case means $Y_i(t)$ by shape regardless of scale, or, more precisely, clustering by correlation in the functional space. By borrowing strength across measurements, the clustering membership is more accurately estimated than for the naive approach as supported by our simulation study (see Section 3.6).

3.3.2 Level-2 Clustering

Clustering by similarity of between-case deviations reduces to estimation of a similarity measure for within-case deviations $\{W_{i_1 j}\}_{j=1,\ldots,J}$ and $\{W_{i_2 j}\}_{j=1,\ldots,J}$. For large J and densely sampled time domain, one such measure is the dynamical correlation for multivariate longitudinal data by Dubin and Müller (2005). However, it is rarely

the case that we will have available a large number of measurements J per each case observed at a large number of time points. Because of this limitation, we propose a MFPCA-based approach as follows

1. Apply MFPCA to impute the scores at level 2: $\hat{\zeta}_{ij,r}$.

2. Apply a multivariate clustering algorithm to the estimated coefficients $\hat{\zeta}_{ij,r}$.

The similarity measure is the average

$$d(i_1, i_2) = \sum_{j=1}^{J} ||\hat{\boldsymbol{\zeta}}_{i_1 j} - \hat{\boldsymbol{\zeta}}_{i_2 j}||^2.$$

3.4 Model-based Clustering

An alternative approach to the hard clustering methods is to borrow strength across random functions within the same cluster (James and Sugar, 2003) to improve the estimation accuracy of the transform coefficients and the cluster memberships. In this section, we introduce a model-based clustering approach which combines both ideas - borrowing strength across random functions within the same cluster and within the same case.

In model-based clustering, the underlying assumption is that the complete data are bivariate variables (X_i, Z_i) for $i = 1, \ldots, I$ where X_i are case-specific realizations from a multivariate distribution and the cluster membership Z_i is a latent variable (Fraley and Raftery, 2002). A common estimation method for model-based clustering is the Estimation-Maximization algorithm where at the Estimation step, we impute or predict the cluster membership $Z = (Z_1, \ldots, Z_I)$ along with estimation of the cluster weights $\pi_1, \ldots, \pi_K$, and at the Maximization step, we estimate the parameters specifying the conditional distribution of $X_i|Z_i,\ i = 1, \ldots, I$. Therefore, we need to specify the conditional distribution $X_i|Z_i,\ i = 1, \ldots, I$ and the distribution of the latent variable Z_i, which in turn, specify the distribution of the complete data. The cluster membership of the ith case Z_i follows a multinomial distribution with proportion parameters $\pi_1, \ldots, \pi_K$ where K is the number of clusters. $X_i|Z_i = k,\ i =$

$1, \ldots, I$ are commonly assumed conditionally independent following a distribution with cluster mean $\mu_k(t)$ and covariance function $\Sigma_k(t, t')$.

Using a similar framework for clustering multilevel data, we assume that the complete data are $(X_{ij}, Z_i^{(1)}, Z_i^{(2)})$ for $i = 1, \ldots, I$ and $j = 1, \ldots, J$ where $Z_i^{(1)}$ and $Z_i^{(2)}$ are latent variables specifying the clustering membership at level 1, and respectively, at level 2. We assume:

• The cluster membership $Z_i^{(1)}$ of the ith case has a multinomial distribution with proportion parameters $\pi_1^{(1)}, \ldots, \pi_{C_1}^{(1)}$ where C_1 is the number of clusters at level 1.

• The cluster membership $Z_i^{(2)}$ of the ith case has a multinomial distribution with proportion parameters $\pi_1^{(2)}, \ldots, \pi_{C_2}^{(2)}$ where C_2 is the number of clusters at level 2.

Level-1 Clustering. For clustering at level 1, we assume $C_2 = 1$ but $C_1 \geq 1$. Therefore, the joint data are $(X_{ij}, Z_i^{(1)})$. However, to model the distribution of the joint data we need to specify the conditional distribution of $X_i | Z_i^{(1)}$. Following the unconditional distribution of X_i in (27), the conditional distribution is:

$$X_{ij}(t) | (Z_i^{(1)} = k) = \sum_{s=1}^{N_1} \nu_{i,s,k} \phi_s^{(1)}(t) + \sum_{r=1}^{N_2} \zeta_{ij,r} \phi_r^{(2)}(t) + \varepsilon_{ij}(t) \text{ with} \tag{29}$$

$$\boldsymbol{\nu}_{i,k} = (\nu_{i,1,k}, \ldots, \nu_{i,N_1,k})' \sim N(\mu_k, \Lambda_k^{(1)})$$
$$\boldsymbol{\zeta}_{ij} = (\zeta_{ij,1}, \ldots, \zeta_{ij,N_2})' \sim N(0, \Lambda_j^{(2)})$$

where $\boldsymbol{\mu}_k = (\mu_{1,k}, \ldots, \mu_{N_1,k})'$ and $\Lambda_k^{(1)}$ is a $N_1 \times N_1$ diagonal matrix with diagonal elements $\boldsymbol{\lambda}_k^{(1)} = (\lambda_{1,k}^{(1)}, \ldots, \lambda_{N_1,k}^{(1)})'$. Under this conditional model, the *conditional scores* $\nu_{i,s,k} = (\xi_{i,s} | Z_i^{(1)} = k)$ for $k = 1, \ldots, C_1$ are assumed independent with conditional mean $\mu_{s,k}$ and conditional variance $\lambda_{s,k}^{(1)}$. Here $\xi_{i,s}$ for $i = 1, \ldots, I$ and $s = 1, \ldots, N_1$ are the unconditional scores at level 1 with a distribution following assumption A.1. For this model, the scores at level 2 are unconditional of the clustering latent variable $Z^{(1)}$, and therefore, their distribution follows the assumption A.3.

From the conditional and unconditional models, we derive

$$0 = \mathbb{E}(\xi_{i,s}) = \mathbb{E}(\mathbb{E}(\xi_{i,s}|Z_i^{(1)})) = \sum_{k=1}^{C_1} \pi_k^{(1)} \mathbb{E}(\nu_{i,s,k}) = \sum_{k=1}^{C_1} \pi_k^{(1)} \mu_{s,k} \tag{30}$$

$$\tau_s^{(1)} = \mathbb{V}(\xi_{i,s}) = \sum_{k=1}^{C_1} \pi_k^{(1)} (\lambda_{s,k}^{(1)} + \mu_{s,k}^2) - (\sum_{k=1}^{C_1} \pi_k^{(1)} \mu_{s,k})^2 = \sum_{k=1}^{C_1} \pi_k^{(1)} (\lambda_{s,k}^{(1)} + \mu_{s,k}^2) \tag{31}$$

It follows that the clustering model at level 1 (*Model 1*) is

$$\begin{cases} X_{ij}(t) = \sum_{s=1}^{N_1} \xi_{i,s} \phi_s^{(1)}(t) + \sum_{r=1}^{N_2} \zeta_{ij,r} \phi_r^{(2)}(t) + \varepsilon_{ij}(t) \\ \xi_{i,s}|(Z_i^{(1)} = k) \sim N(\mu_{s,k}, \lambda_{s,k}^{(1)}) \\ Z_i^{(1)} \sim \text{ Multinomial}(1; \pi_1^{(1)}, \ldots, \pi_{C_1}^{(1)}) \\ \zeta_{ij,r} \sim N(0, \lambda_{j,r}^{(2)}) \text{ indep. of } \xi_{i,s,k}, \ Z_i^{(1)} \end{cases} \tag{32}$$

subject to the constrain $\sum_{k=1}^{C_1} \pi_k^{(1)} \mu_{s,k} = 0$ by (30). We note that the relationship between conditional and unconditional variances in equation (31) does not impose a constraint.

Under this clustering set up, the kth cluster mean is

$$\mathbb{E}(X_{ij}(t)|Z_i^{(1)} = k) = \mathbb{E}(Y_i(t)|Z_i^{(1)} = k) = \sum_{s=1}^{N_1} \mu_{s,k} \phi_s^{(1)}(t).$$

Level-2 Clustering. For clustering at level 2, we assume $C_1 = 1$ but $C_2 \geq 1$. Therefore, the joint data are $(X_i, Z_i^{(2)})$ and the conditional distribution of $X_i|Z_i^{(2)}$ is:

$$X_{ij}(t)|(Z_i^{(2)} = k) = \sum_{s=1}^{N_1} \xi_{i,s} \phi_s^{(1)}(t) + \sum_{r=1}^{N_2} \delta_{ij,r,k} \phi_r^{(2)}(t) + \varepsilon_{ij}(t) \text{ with} \tag{33}$$

$$\xi_i = (\xi_{i,1}, \ldots, \xi_{i,N_1})' \sim N(0, \Lambda^{(1)})$$

$$\boldsymbol{\delta}_{ij,k} = (\delta_{ij,1,k}, \ldots, \delta_{ij,N_2,k})' \sim N(\eta_{jk}, \Lambda_{j,k}^{(2)})$$

where $\boldsymbol{\eta}_{jk} = (\eta_{j,1,k}, \ldots, \eta_{j,N_2,k})'$ and $\Lambda_{jk}^{(2)}$ is an $N_2 \times N_2$ diagonal matrix with diagonal elements $\boldsymbol{\lambda}_{jk}^{(2)} = (\lambda_{j,1,k}^{(2)}, \ldots, \lambda_{j,N_2,k}^{(2)})'$. Under this conditional model, the *conditional scores at level 2*, $\delta_{ij,r,k} = (\zeta_{ij,r}|Z_i^{(2)} = k)$, are assumed independent with conditional mean $\eta_{j,r,k}$ and conditional variance $\tau_{j,r,k}^{(2)}$ for $k = 1, \ldots, C_2$. Here $\zeta_{ij,r}$'s are the unconditional scores in the unconditional model (28) assumed independent with mean

zero ($\mathbb{E}(\zeta_{ij,r}) = 0$) and constant variance across cases ($\mathbb{V}(\zeta_{ij,r}) = \tau^{(2)}_{j,r}$) as provided in assumption A.3.

From the conditional and unconditional models, we derive

$$0 = \mathbb{E}(\zeta_{i,s}) = \mathbb{E}(\mathbb{E}(\zeta_{i,s}|Z^{(2)}_i)) = \sum_{k=1}^{C_2} \pi^{(2)}_k \mathbb{E}(\delta_{i,s,k}) = \sum_{k=1}^{C_2} \pi^{(2)}_k \eta_{s,k} \tag{34}$$

$$\tau^{(2)}_{j,r} = \mathbb{V}(\zeta_{ij,r}) = \sum_{k=1}^{C_2} \pi^{(2)}_k (\lambda^{(1)}_{j,r,k} + \eta^2_{j,r,k}) - (\sum_{k=1}^{C_2} \pi^{(2)}_k \eta_{j,r,k})^2 = \sum_{k=1}^{C_2} \pi^{(2)}_k (\lambda^{(1)}_{j,r,k} + \eta^2_{j,r,k}) \tag{35}$$

Similar to the clustering model at level 1, the clustering model at level 2 (*Model 2*) is

$$\begin{cases} X_{ij}(t) = \sum_{s=1}^{N_1} \xi_{i,s}\phi^{(1)}_s(t) + \sum_{r=1}^{N_2} \zeta_{ij,r}\phi^{(2)}_r(t) + \varepsilon_{ij}(t) \\ \zeta_{ij,r}|(Z^{(2)}_i = k) \sim N(\eta_{j,k}, \Lambda^{(1)}_{j,k}) \\ Z^{(2)}_i \sim \text{ Multinomial}(1; \pi^{(2)}_1, \ldots, \pi^{(2)}_k) \\ \xi_{i,s} \sim N(0, \tau^{(1)}_s) \text{ indep. of } \zeta_{ij,r,k},\ Z^{(2)}_i \end{cases} \tag{36}$$

subject to the constraint $\sum_{k=1}^{C_2} \pi^{(2)}_k \eta_{j,r,k} = 0$ by (34). On the other hand, the relationship between unconditional and conditional variances in equation (35) does not imposes a constraint but it requires that the unconditional variances vary with the measurement type when η_{jk} also vary with the measurement type leading to assumption A.3 in Section 3.2. However, MFPCA as introduced by Di et al. (2009) does not allow for the eigenvalues at level-2 to vary with the measurement type. For this reason, the estimated level-2 scores will provide lower accuracy clustering for larger number of measurement types J; this observation is supported in our simulation study.

Under this clustering set up, the kth cluster trend for the jth condition is

$$\mathbb{E}(X_{ij}(t)|Z^{(2)}_i = k) = \mathbb{E}(W_{ij}(t)|Z^{(2)}_i = k) = \sum_{r=1}^{N_2} \eta_{j,r,k}\phi^{(2)}_r(t).$$

The formulation of the level-1 and level-2 joint clustering model is provided in the Appendix along with the estimation method that applies not only to the joint model but also to the reduced models discussed in this section.

3.5 Model Selection

The clustering models described in the previous section depend on a series of parameters which are assumed fixed: C_1, C_2, N_1 and N_2. We identify two model selection problem (1) Selecting the number of eigenfunctions which explain a large percentage of the variability between cases (selecting N_1) and within cases (selecting N_2); and (2) Selecting the number of clusters at level 1 (selecting C_1) and/or the number of clusters at level 2 (selecting C_2). We can select N_1 and N_2 using the unconditional model (MFPCA). Di et al. (2008) and Di and Crainiceanu (2010) discuss various alternative methods for selection of the number of basis functions and we follow their direction.

There are several existing methods for estimating the number of unknown clusters for model-based clustering (Fraley and Raftery, 2002; Sugar and James, 2003). Since our clustering algorithm is model-based, the problem of identifying the number of clusters is equivalent to a model selection problem since each number of clusters corresponds to a different model. Consequently, we will focus our attention on likelihood-based approaches.

Common variable selection methods, such as the Akaike information criterion (AIC), and Bayesian information criterion (BIC) have been employed for estimating the number of clusters (Fraley and Raftery, 2002). Both criteria select the number of clusters which minimizes the objective function of the form

$$-2\log L(\Psi) + 2J(C_1, C_2)$$

where $\log L(\hat{\Psi})$ is the log likelihood of observed data which measures the lack of fit. In our multi-level clustering model,

$$\log L(\Psi) = \sum_{i=1}^{I}\sum_{k=1}^{C_1}\sum_{k'=1}^{C_2} \pi_k^{(1)}\pi_{k'}^{(2)} \log f(X_i; \mu_k, \eta_{k'}, \Lambda_k^{(1)}, \Lambda_{k'}^{(2)}, \sigma^2).$$

The second term $2J(C_1, C_2)$ is the penalty term that measures the complexity of the

model. For AIC, $2J = 2d$ and $2J = (\log I)d$ for BIC where $d = 2C_1K_1 + 2C_2K_2 - K_1 - K_2 + C_1 + C_2 - 1$ is the number of parameters.

Many authors (for example, Koehler and Murphee, 1988) observed that models selected using AIC tend to overfit as AIC prefers larger models. In the model-based clustering context, this translates into overestimation of the number of clusters (Soromenho, 1933; Celeux and Soromenho, 1996). Alternatively, the likelihood correction using BIC selects more parsimonious models. Consequently, BIC selection criteria has been often used in model-based clustering (Fraley and Raftery, 1998). Lereoux (1992) has shown that BIC does not underestimate the true number of components, asymptotically.

Indeed, in our simulation studies (not reported here), we have assessed the number of clusters for various settings and with the number of clusters ranging from 2 to 10. Similarly to past research, BIC most often correctly identifies the number of clusters whereas AIC overestimates the number of clusters in average adding 2 additional clusters from the true clustering.

3.6 Simulation Studies

3.6.1 Level-1 Clustering

The first objective of this research is to assess the accuracy of the clustering membership under two comparative settings: 1. Sparse vs. dense sampling design; and 2. Naive vs. hard vs. soft clustering.

We generate samples of functions from the joint model $(X_i, Z_i^{(1)})$ described in Section 3.4. Specifically, we generate $Z_i^{(1)}$, the clustering membership, from multinomial distribution with fixed cluster weights $\pi_1^{(1)}, \ldots, \pi_{C_1}^{(1)}$ across all simulations. For simplicity, we choose $C_1 = 2$ with $\pi_1^{(1)} = 1/3$ and $\pi_2^{(1)} = 2/3$.

We simulate for $I = 100$ cases with $N_1 = 4$ eigenfunctions at level 1 and $N_2 = 4$ eigenfunctions at level 2. The conditional variances at level 1 are generated according

to the two different settings:

- Equal conditional variances across clusters: $\lambda_{s,k} = 0.9^{s-1}$ for $k = 1, \ldots, C_1$; and
- Varying conditional variances across clusters: $\lambda_{s,k} = 2^{2(k-s)-1}$.

The unconditional variances (true eigenvalues) at level 2 are $\tau_{jr} = \frac{j+1}{2^{2r}}$. The conditional means at level 1 are $\mu_1 = (3, 2, 1, 0)$ and $\mu_2 = (-1.5, -1, -0.5, 0)$ selected such that $\sum_{k=1}^{C_1} \pi_k^{(1)} \mu_{s,k} = 0$. The eigenfunctions are

$$\Phi^{(1)}(t) = (\sqrt{2}\sin(2\pi t), \sqrt{2}\cos(2\pi t), \sqrt{2}\sin(4\pi t), \sqrt{2}\cos(4\pi t))$$
$$\Phi^{(2)}(t) = (1, \sqrt{3}(2t-1), \sqrt{5}(6t^2 - 6t + 1), \sqrt{7}(20t^3 - 30t^2 + 12t - 1).$$

Note that the basis of function at level 2 is mutually non-orthogonal. The noise level is $\sigma = 2$. We vary the number of maximum observations per random function, $m = 4, 6, 10, 15$ and the number of measurement types per case, $J = 2, 3, 4, 5$. We expect higher accuracy of the clustering membership with a larger number of observations and a larger number of repeated measurements when m small.

In our simulation example, because we have the true clustering membership, we can assess the accuracy of the clustering prediction for the method introduced in this chapter and other existing methods using a clustering/classification error.

We measure the clustering error using Rand index (Rand, 1971), which is the fraction of all misclustered pairs of functions. Let $\mathcal{C} = \{f_1, \ldots, f_S\}$ denote the set of true functions, $\hat{\mathcal{C}} = \{\hat{f}_1, \ldots, \hat{f}_S\}$ denote the set of estimated functions, and T and $\hat{T}$ denote the true and estimated clustering maps, respectively. Rand index is defined by

$$\mathcal{R}(\mathcal{C}, \hat{\mathcal{C}}) = \frac{\sum_{r<s} I(T_k(f_r, f_s) \neq \hat{T}_k(f_r, f_s))}{\binom{N}{2}}.$$

Therefore, the Rand index is low when there are only few misclustered functions.

We report the estimation accuracy of the clustering membership under the assumption of hard clustering for varying number of maximum number of design points per function (N) and maximum number of repeated measurements per case (J). We

compare the naive clustering and MFPCA clustering approaches discussed in Section **??** to the model-based clustering approach discussed in Section 3.4. In the tables below, we denote with $*$ simulations in which the naive clustering algorithm was computationally unstable; these are settings in which m is small.

Tables 7 and 8 provide the 1-Rand index for clustering data generated as described above; the Rand index is used to assess the clustering accuracy. Tables 9 and 10 provide the relative mean square error (RMSE) for the clustering patterns calculated as

$$RMSE = \frac{1}{C_1} \sum_{k=1}^{C_1} \frac{\int_{\mathcal{T}} (\mu_k(t) - \hat{\mu}_k(t))^2 \, dt}{\int_{\mathcal{T}} \mu_k^2(t) \, dt}.$$

The values reported for the Rand index and mean squared errors are averages over the 100 simulations. The RMSE is used to assess the accuracy of the clustering patterns. We brief the estimation accuracy results as follows:

• There is a significant improvement in estimation accuracy for both the clustering membership and clustering patterns from the naive approach to MFPCA-based clustering;

• For equal conditional variances, the MFPCA-based and model-based clustering methods perform similarly whereas for varying conditional variances, a more realistic setting, the model based clustering performs better uniformly over all settings;

• As J and m increase, the clustering estimation accuracy increases; when comparing $m = 10$ to $m = 15$ there is little gain in the estimation accuracy which indicates that the extra number of observations per random functions will not add much to the accuracy for this simulation setting. On the other hand, an increase in J leads to more significant increase in estimation accuracy.

3.6.2 Level-2 Clustering

To assess the clustering performance of our model-based method at level 2, we simulate $C_2 = 2$ clusters with $\pi_1^{(2)} = 1/3$ and $\pi_2^{(2)} = 2/3$. The true eigenfunctions are

Table 7: Equal variance: Rand index for the clustering membership at level 1: Naive Approach|Hard Clustering|Soft Clustering

	$m = 4$	$m = 6$
$J = 2$	* | **0.083** | 0.097	* | 0.054 | **0.054**
$J = 3$	* | **0.068** | 0.070	0.127 | **0.038** | 0.040
$J = 4$	* | 0.056 | **0.047**	0.093 | **0.037** | **0.037**
$J = 5$	0.182 | **0.045** | 0.045	0.077 | **0.038** | 0.047
	$m = 10$	$m = 15$
$J = 2$	0.086 | 0.041 | **0.029**	0.057 | 0.028 | **0.018**
$J = 3$	0.061 | **0.028** | 0.034	0.035 | 0.026 | **0.023**
$J = 4$	0.040 | **0.028** | 0.036	0.029 | 0.023 | **0.017**
$J = 5$	0.037 | **0.023** | 0.035	0.025 | **0.017** | 0.027

Table 8: Varying variance: Rand index for the clustering membership at level 1: Naive Approach|Hard Clustering|Soft Clustering

	$m = 4$	$m = 6$
$J = 2$	* | 0.106 | **0.061**	0.190 | 0.090 | **0.039**
$J = 3$	* | 0.092 | **0.052**	0.110 | 0.091 | **0.036**
$J = 4$	0.170 | 0.072 | **0.026**	0.105 | 0.054 | **0.024**
$J = 5$	0.114 | 0.070 | **0.039**	0.071 | 0.067 | **0.022**
	$m = 10$	$m = 15$
$J = 2$	0.100 | 0.089 | **0.025**	0.082 | 0.075 | **0.018**
$J = 3$	0.070 | 0.057 | **0.016**	0.092 | 0.089 | **0.029**
$J = 4$	0.056 | 0.052 | **0.023**	0.061 | 0.071 | **0.028**
$J = 5$	0.061 | 0.091 | **0.035**	0.068 | 0.075 | **0.025**

Table 9: Equal variance: RMSE for the clustering patterns at level 1: Naive Approach|Hard Clustering|Soft Clustering

	$m = 4$	$m = 6$
$J = 2$	* | **0.0822** | 0.0934	* | **0.0404** | 0.0535
$J = 3$	* | **0.0546** | 0.0772	0.170 | **0.037** | 0.05
$J = 4$	* | **0.0486** | 0.0559	0.11 | **0.032** | 0.047
$J = 5$	0.19 | **0.043** | 0.055	0.055 | **0.029** | 0.066
	$m = 10$	$m = 15$
$J = 2$	0.085 | **0.028** | 0.03	0.046 | **0.021** | 0.023
$J = 3$	0.048 | **0.022** | 0.045	0.025 | **0.020** | 0.027
$J = 4$	0.031 | **0.026** | 0.046	**0.020** | 0.021 | 0.023
$J = 5$	0.025 | **0.020** | 0.047	0.018 | **0.017** | 0.042

Table 10: Varying variance: RMSE for the clustering patterns at level 1: Naive Approach|Hard Clustering|Soft Clustering

	$m = 4$	$m = 6$
$J = 2$	* | 0.087 | **0.076**	0.329 | 0.071 | **0.043**
$J = 3$	* | **0.05** | 0.058	0.120 | 0.073 | **0.042**
$J = 4$	0.177 | 0.063 | **0.039**	0.11 | 0.047 | **0.037**
$J = 5$	0.12 | 0.059 | **0.057**	0.069 | 0.062 | **0.036**
	$m = 10$	$m = 15$
$J = 2$	0.13 | 0.066 | **0.032**	0.097 | 0.057 | **0.022**
$J = 3$	0.075 | 0.047 | **0.025**	0.085 | 0.075 | **0.042**
$J = 4$	0.052 | 0.038 | **0.035**	0.061 | 0.054 | **0.038**
$J = 5$	0.048 | 0.064 | **0.051**	0.066 | 0.083 | **0.029**

the same as in the previous section and the unconditional variances (true eigenvalues) at level 1 are $\lambda_{s,k} = 0.9^{s-1}$. The conditional means at level 2, $\eta_{j,k}$, are selected such that $\sum_{k=1}^{C_2} \pi_k^{(2)} \eta_{j,k} = 0$. Since in our simulations we compare the accuracy with $J = 2, 3, 4, 5$, the conditional means for the cluster 1 are as follows

$$\begin{aligned} \eta_{1,1} &= (4, 3, 2, 1), \eta_{2,1} = (4, -3, 2, -1), \eta_{3,1} = (4, -3, -2, 1), \\ \eta_{4,1} &= (-4, 3, -2, 1), \eta_{5,1} = (-4, -3, -2, -1) \end{aligned}$$

and the means for the cluster 2 are

$$\begin{aligned} \eta_{1,2} &= (-2, -1.5, -1, -0.5), \eta_{2,2} = (-2, 1.5, -1, 0.5), \eta_{3,2} = (-2, 1.5, 1, -0.5), \\ \eta_{4,2} &= (2, -1.5, 1, -0.5), \eta_{5,2} = (2, 1.5, 1, 0.5). \end{aligned}$$

The conditional variances at level 2 are $\lambda_{jr,k} = \frac{a_{kj}}{2^{(2(r-1))}}$ where a_{kj} is a scaling constant randomly generated from $Unif(0.5, 1.5)$ (varying across clusters and across replicates within each case).

Tables 11 and 12 provide the accuracy of the clustering membership measured by the Rand index and the accuracy of the clustering patterns measured by the mean square error for the simulation setting above. Note that we don't show the results for equal level-2 conditional variances as this is not a realistic assumption because of the constraint given by (35). We brief the estimation accuracy results as follows:

Table 11: Rand Index for the clustering membership at level 2: Hard Clustering|Soft Clustering

	$m = 4$	$m = 6$	$m = 10$	$m = 15$
$J = 2$	0.149 | **0.090**	0.073 | **0.035**	0.005 | **0.003**	0.001 | **0.000**
$J = 3$	0.468 | **0.202**	0.478 | **0.177**	0.475 | **0.143**	0.470 | **0.093**
$J = 4$	0.468 | **0.071**	0.474 | **0.079**	0.481 | **0.098**	0.473 | **0.091**
$J = 5$	0.461 | **0.182**	0.464 | **0.165**	0.477 | **0.116**	0.469 | **0.114**

Table 12: RMSE for the clustering patterns at level 2: Hard Clustering|Soft Clustering

	$m = 4$	$m = 6$	$m = 10$	$m = 15$
$J = 2$	**1.232** | 1.284	**1.268** | 1.276	**1.242** | 1.243	1.248 | **1.235**
$J = 3$	1.234 | **1.221**	1.202 | **1.193**	1.213 | **1.208**	**1.208** | 1.252
$J = 4$	1.170 | **1.094**	1.198 | **1.184**	1.223 | **1.187**	1.253 | **1.124**
$J = 5$	1.307 | **1.213**	1.342 | **1.260**	**1.271** | 1.325	**1.186** | 1.207

• As J increases, the clustering membership accuracy improves significantly for the model-based clustering approach as compared to the hard-clustering on the estimated level-2 MFPCA scores. One reason for this significant improvement is that the MFPCA approach assumes equal conditional variances of the scores whereas the model-based clustering does not (assumption A.3).

• An increase in m does not improve the accuracy of the clustering membership estimated using the MFPCA-based hard clustering approach but it does improve the accuracy for the model-based clustering.

• The accuracy of the model-based clustering also decreases as J increases; this is because the initial clustering membership for the model-based clustering method is based on the MFPCA estimated level-2 scores.

• Although the clustering membership is not accurately estimated by the hard clustering approach as compared to the model-based approach, the RMSE of the clustering patterns is comparable for both methods. Therefore, both methods capture the clustering patterns equally well but they provide different clustering memberships.

3.7 Case Study

Sales forecasting plays a fundamental role in retail business strategy. Accurate sales forecasting can help a retail provider make appropriate decisions on inventory management as well as provide valuable input to the company's operating and financial planning systems. For perishable goods, accurate forecasting results in decreased loss due to over- or under-supplying. Sales demand data may show high variation or may be insufficient to construct reliable forecast models at the individual product level.

Recent retail sales forecasting methods have taken advantage of the intrinsic hierarchy of product categorizations. The hierarchical structure extends from category of products to multiple stores. When forecast demand at the detailed level (single product, single store, single week level) is of primary interest, retailers commonly forecast sales in a higher aggregation level because detailed forecasting can be more difficult than aggregated forecasting. An important difficulty in aggregate forecasting is identifying a meaningful categorization.

To this end, we apply the clustering approach introduced in this chapter to a database of product sales from a large retailer in the U.S. The sales for each product are aggregated at the geographic region to overcome the sparsity of the sales. Specifically, we observe monthly count sales of I products within J geographic regions: $X_{ij}(t)$ where the t is the month index varying over a period of 2 years (i.e. $m = 24$). Although the data are discretely observed since the number of counts is large we use the variance stabilization transformation $\sqrt{X_{ij} + 3/8}$ and apply the clustering algorithm under the assumption of normality. The extension of the clustering algorithm to count data will be discussed elsewhere.

The objective of this study is to obtain a clustering of products by similarity in sales across geographic regions. The clustering of products may be used in conjunction with forecasting methods that allow borrowing information across products to enhance the prediction accuracy.

Remarks: right now we are still waiting for the data and will complete the analysis once the data are in hands.

3.8 Discussions

In this chapter, we introduce a means for clustering functional data with an intrinsic hierarchical structure; the clustering algorithm identifies groups of functions which are similar in their within- or/and between random trends. The underlying clustering (hard or soft) begins with the specification of a model using functional principal component analysis and either clusters the resulting estimated scores using common hard clustering or updates the estimated scores assuming a clustering model. The estimation procedure for the latter approach is iterative and therefore more computational expensive.

From our simulation studies, we find that clustering by similarity of within-case means at level-1 using either of the two approaches will provide similar results as soon as there is not a significant difference in within-cluster variability across clusters. Therefore, the extra computational cost incurred by updating the scores using the clustering model will be worth when the variability across functions assigned to the same cluster will largely vary from one cluster to another. If the number of cases to be clustered is not large ($I \sim 100 - 1000$), we advice in proceeding with the model-based clustering as the additional computational cost is not great. On the other hand, for large I either a more computational efficient implementation of the model-based clustering method should be considered or simply the application of the hard clustering approach with the understanding of its shortcomings.

Clustering by similarity of between-case deviations at level 2 is more difficult as it pools information across multiple functions simultaneously. The hard clustering approach using the estimated scores from MFPCA provides inaccurate clustering as J,

the number of measurements, increases. This may primarily be because of the restrictive assumption that the eigenvalues do not vary with the measurement type. After updating the scores using the clustering model, the accuracy of the estimated cluster membership improves significantly however not so the accuracy of the estimated cluster patterns. The reason for poor estimation of the cluster patterns regardless of the clustering approach is the estimation inaccuracy of the eigen-functions provided by the MFPCA approach. One way to overcome this problem is to update the eigen-function within a more complex clustering model or to assume fixed eigen-functions from a known basis of functions (James et al., 2000). These alternative methods will be discussed elsewhere.

CHAPTER IV

FUTURE WORK: A MULTILEVEL SPACE-TIME AUTOREGRESSIVE MODEL

4.1 Introduction

In many real applications, the data are areal data collected over geographic regions such as states, counties or census tracts and thus with an intrinsic hierarchical structure. Often, the variables of interest come from different data sources observed at different spatial resolution. In this chapter, our focus is on developing a statistical model for estimating temporal and spatial associations of a series of time-varying variables observed at different spatial resolution levels to a primary response variable observed at the highest spatial resolution level. Specifically, we observe a response variable $Y_{ij,k,t}$ at the kth census tract in the jth county and ith state and at the time point t. We also observe multiple predictor variables at varying geographic levels: $X_{ij,k,t}$ at the census tract level, $R_{i,j,t}$ at the county level and $Z_{i,t}$ at the state level. Importantly, we are not only interested in contemporaneous associations but also in spatial and temporal lagged associations often referred to as Granger causal effects.

Our underlying objective is to simultaneously estimate and make inference on

- Autoregressive spatial-temporal associations:

$$Y_{ij,k,t} \sim Y_{ij,\partial_{\delta_1} k,t-l_1}.$$

- Autoregressive spatial-temporal associations at lower spatial resolutions:

$$Y_{ij,k,t} \sim Y_{i,\partial_{\delta_2} j,k,t-l_2},\ Y_{\partial_{\delta_3} i,j,k,t-l_3}.$$

- Exogenous lagged and concurrent spatial-temporal associations:

$$Y_{ij,k}(t) \sim X_{ij,\partial_{\delta_1} k,t-l_1},\ X_{i,\partial_{\delta_2} j,k,t-l_1},\ X_{\partial_{\delta_3} i,j,k,t-l_1},\ R_{i,\partial_{\delta_2} j,t-l_2},\ R_{\partial_{\delta_1} i,j,t-l_2},\ Z_{\partial_{\delta_3} i,t-l_3}$$

for $\delta_1, \delta_2, \delta_3 \geq 0$, and $l_1, l_2, l_3 \geq 0$. In the lagged spatial associations above, $\partial_{\delta_1} k$ is the index of the δ_1-closest census tract from the kth census tract, $\partial_{\delta_2} j$ is the index of the δ_2 -closest counties from the jth county and $\partial_{\delta_i} i$ is the index of the δ_3-closest state from the ith state.

To study the association between variables observed at different resolution levels, one alternative is to use techniques in the literature of 'Change of Support Problem' (COSP) to bring all variables to a homogeneous resolution level or support. Alternative approaches in COSP have been reviewed by Gottway and Young (2002). Challenges arising in the application of these techniques to our statistical model are multiple. First, the variables in our study are not only varying with space but also with time. Second, we have multiple predictors at varying resolutions levels that require undergoing a change of support. Third, we estimate various associations (e.g. lagged vs. concurrent, spatial vs. temporal). Therefore, the application of COSP approaches prior to modeling the underlying associations will lead to a computationally complex approach with intricate and difficult to assess modeling biases due to the disaggregation effect. To sidestep the difficulties arising from changing the support of one or more predictor variables, we introduce a multilevel autoregressive model in which spatial and temporal associations are modelled at each spatial resolution level.

Our methodological contribution is two fold. First, we introduce a multilevel autoregressive model which allows estimation of spatial-temporal concurrent and lagged effects of the response variable itself as well as of exogenous variables which are observed at various spatial resolution levels. importantly, autoregressive spatial effects enter the model not only at the highest resolution level at which the response variable is observed but also at the lower resolution levels.

Second, in order to assess the significance of the autoregressive and exogenous spatial-temporal effects we investigate the classical boosting method for variable selection in the context of the more complex modeling framework in this paper. In this

chapter, we use boosting as the basis for our model selection method because it can be extended to the multilevel autoregressive model discussed in this paper and it does not require orthogonality between the variables to be selected.

The endpoint of the statistical model developed in this paper is to infer a multilevel graphical model describing the Granger causality of a series of predictors which are observed at various spatial resolutions to a space-time varying response variable. Granger causality (Granger 1969) developed by the Nobel prize winning economist, Clive Granger, is a statistical concept of causality that is based on prediction. According to Granger causality, if the past values of a variable X contain information that helps predict Y beyond the information contained in past values of Y alone, then X is said to Granger-cause Y. In the time series literature, Granger causality has been mainly focused in temporal causality effects. An important contribution of this work is that the resulting graphical model will not only describe temporal causal effects but also spatial causal effects. In many applications such as the accessibility study in this paper, the accessibility at a certain location will not only be influenced by the past activities within the local area but also those of the neighboring/surrounding regions.

4.2 The Model

The observed data are:

• Sampling spatial units: census tracts $(s_1, \ldots, s_{K_1})$ which are nested in counties $(u_1, \ldots, u_{K_2})$ which in turn are nested in states $(v_1, \ldots, v_{K_3})$.

• An underlying response variable that varies in space and time: $Y_{ijk,t} = Y_{ij}(s_k, t)$ where s_k is a census tract in county u_j and state v_i.

• Predictor variables observed at the census tract level: $X_{ijk,t} = X_{ij}(s_k, t)$ where s_k is a census tract in county u_j and state v_i.

• Predictor variables observed at the county level: $R_{ij,t} = R_i(u_j, t)$ where u_j is a

Table 13: Multi-level space-time autoregressive model

	Spatio-temporal autoregressive	Spatio-temporal casual
Census Tract	$Y_{ij,\partial_d k,t-l}$	$X_{ij,\partial_d k,t-l}$
County	$\mu^Y_{i,\partial_d j,t-l}$	$R_{i,\partial_d j,t-l}, \mu^X_{i,\partial_d j,t-l}$
State	$\mu^Y_{\partial_d i,t-l}$	$X_{\partial_d i,t-l}, \mu^X_{\partial_d i,t-l}, \mu^R_{\partial_d i,t-l}$

county within the state v_i.

• Predictor variables observed at the state level: $Z_{i,t} = Z(v_i, t)$.

The model we investigate is:

$$Y_{ijk,t} = \sum_{d=0}^{D_{1y}} \sum_{l=1}^{L_{1y}} \alpha^Y_{dl} Y_{ij,\partial_d k,t-l} + \sum_{d=0}^{D_{1x}} \sum_{l=0}^{L_{1x}} \alpha^X_{dl} X_{ij,\partial_d k,t-l} + \sum_{l=1}^{L_{2y}} \mu^Y_{ij,t-l} + \sum_{l=1}^{L_{2x}} \mu^X_{ij,t-l} + \epsilon_{ijk,t}$$

where $\mu^Y_{ij,t-l}$ and $\mu^X_{ij,t-l}$ are county-level random effects decomposed as follows

$$\begin{aligned}
\mu^Y_{ij,t-l} &= \sum_{d=1}^{D_{2y}} \beta^Y_{dl} \mu^Y_{i,\partial_d j,t-l} + \sum_{d=0}^{D_{2r}} \beta^R_{dl} R_{i,\partial_d j,t-l} + \mu^Y_{i,t-l} + \mu^R_{i,t-l} + \eta^Y_{ij,t-l} \\
\mu^X_{ij,t-l} &= \sum_{d=1}^{D_{2x}} \beta^X_{dl} \mu^X_{i,\partial_d j,t-l} + \mu^X_{i,t-l} + \eta^X_{ij,t-l}
\end{aligned}$$

where, in turn, $\mu^Y_{i,t-l}$, $\mu^R_{i,t-l}$ and $\mu^X_{i,t-l}$ are state-level random effects further decomposed

$$\begin{aligned}
\mu^Y_{i,t-l} &= \sum_{d=1}^{D_{3y}} \gamma^Y_{dl} \mu^Y_{\partial_d i,t-l} + \sum_{d=0}^{D_{2z}} \gamma^Z_{dl} Z_{\partial_d i,t-l} + \mu^Y_{t-l} + \mu^Z_{t-l} + \xi^Y_{i,t-l} \\
\mu^X_{i,t-l} &= \sum_{d=1}^{D_{3x}} \gamma^X_{dl} \mu^X_{\partial_d i,t-l} + \mu^X_{t-l} + \xi^X_{i,t-l} \\
\mu^R_{i,t-l} &= \sum_{d=1}^{D_{3r}} \gamma^R_{dl} \mu^R_{\partial_d i,t-l} + \mu^R_{t-l} + \xi^R_{i,t-l}
\end{aligned}$$

Finally, $\mu^Y_{t-l} \sim N(0, \sigma^2_{y,l})$, $\mu^Z_{t-l} \sim N(0, \sigma^2_{z,l})$, $\mu^X_{t-l} \sim N(0, \sigma^2_{x,l})$ and $\mu^R_{t-l} \sim N(0, \sigma^2_{r,l})$ are random coefficients accounting for the unexplained variability at the lth lag. The error terms in the model are:

• $\epsilon_{ijk,t}$ assumed independently and identically distributed with mean 0 and variance σ^2_1;

• $\eta^Y_{ij,t-l}$ and $\eta^X_{ij,t-l}$ assumed independent with standard deviances σ^Y_2, and respectively, σ^X_2; and

• $\xi^Y_{i,t-l}$, $\xi^X_{ij,t-l}$ and $\xi^R_{i,t-l}$ assumed independent with standard deviances σ^Y_3, σ^X_3, and respectively, σ^R_3.

The parameters to be estimated are:

• Level 1: α^Y_{dl} and α^X_{dl} as well as the error standard deviation σ_1;

• Level 2: β^Y_{dl}, β^X_{dl} and β^R_{dl} as well as the error standard deviations, σ^Y_2 and σ^X_2;

• Level 3: γ^Y_{dl}, γ^X_{dl}, γ^R_{dl} and γ^Z_{dl} as well as the error standard deviations, σ^Y_3, σ^X_3, and σ^R_3.

This is still an on-going research work and I will continue working on it.

APPENDIX A

FITTING ALGORITHM OF SPATIAL CLUSTERING MODEL

The details of the E and M iterative steps in the modified EM estimation algorithm are provided below:

E-step In the E-step, we need to calculate the conditional expectation of $L(\Theta)$, given the observed data Y.

$$Q(\Theta) = E[-\log L(\Theta)|Y] = -E[\log f(Y|Z)] - E[\log f(\gamma)] - E[\log f(z_1, \ldots, z_S)]$$

$$\approx -E[\log f(Y|Z)] - E[\log f(\gamma)] - \sum_{j=1}^{S} E[\log f(z_{s_j}|z_{\partial s_j}; \psi)]$$

$$= \frac{1}{2}\sum_{j=1}^{S}\sum_{k=1}^{C} z_{jk}[T\log\sigma_\varepsilon^2 + \frac{1}{\sigma_\varepsilon^2}\sum_{j=1}^{S}\|Y_j - \mathbf{1}_T\mu_{0,j} - \Phi_T\beta_k - X_{S,j}\alpha - H_T E(u_k|Y) - H_{S,j}E(w_J|Y)\|^2$$

$$+ \frac{1}{2}\sum_{k=1}^{C}[T\log\sigma_k^2 + E(u_k'u_k|Y)/\sigma_k^2] + \frac{1}{2}[J\log\sigma_s^2 + E(w_J'w_J|Y)/\sigma_s^2] - \sum_{j=1}^{S}\sum_{k=1}^{C} z_{jk}\log(\pi_{jk})$$

where $X_{S,j} = \Phi_{S,j} \otimes 1_T$ and $H_{S,j} = \tilde{H}_{S,j} \otimes 1_T$; $X_{S,j}$ and $\tilde{H}_{S,j}$ are the jth row of Φ_S and $\tilde{H}$. We use the pseudo-likelihood 7 to approximate the joint likelihood $f(z_1, \ldots, z_S)$. The predicted cluster membership $z_{jk} = E[z_{s_j} = k|Y_j] = \frac{\pi_{jk}f(Y_j|z_{s_j})}{\sum_{k=1}^{C}\pi_{jk}f(Y_j|z_{s_j})}$. In this chapter, we define π_{jk} defined through the Gibbs distribution as in the Equation 6, and thus we perform a hard clustering in the each iteration. We predict $Z_j = k$ if $z_{jk} = E[z_{s_j} = k|Y_j]$ is maximized among all the k.

In addition to reconstructing the cluster membership, at the E-step, the random effects $\gamma = (u_1', \ldots, u_C', w_J')'$ are also predicted from the conditional distribution $\gamma|Y$ given by

$$N((\sigma_\varepsilon^2\Gamma^{-1} + H'H)^{-1}H'\left(Y - \mathbf{I}\mu_0 - X_T\beta - X_S\alpha\right), (\Gamma^{-1} + H'H/\sigma_\varepsilon^2)^{-1}).$$

M-step The parameters $\mu_0, \beta_1, \ldots, \beta_C, \alpha, \sigma_\varepsilon^2$ are estimated by maximizing the expectation of the likelihood function $f(Y|\gamma, Z; \mu_0, \beta_1, \ldots, \beta_C, \alpha, \sigma_\varepsilon^2)$. The constraint on the cluster fixed effects ($\beta_1 + \ldots + \beta_C = 0$) ensures identifiability of the model parameters.

The estimates of curve-specific offset parameters are

$$\hat{\mu}_{0,j} = \frac{1}{T}\sum_{i=1}^{T}\left[Y_{ij} - \sum_{\nu=1}^{p}\phi_{T,\nu}(t_i)\hat{\beta}_{z_{s_j},\nu} - \sum_{\nu=1}^{q}\phi_{S,\nu}(s_j)\hat{\alpha}_\nu - \sum_{m=1}^{T}\{H_T\}_{i,m}\hat{u}_{z_{s_j},m} - \sum_{n=1}^{J}\{\tilde{H}_S\}_{j,n}\hat{w}_{J,n}\right]$$

$\hat{\beta}_1, \ldots, \hat{\beta}_C$ are estimated by solving a Lagrange multiplier problem,

$$\min_{\beta_1,\ldots,\beta_C,\lambda}\sum_{j=1}^{S}\|Y_j - \mathbf{1}_T\hat{\mu}_{0,j} - \Phi_T\beta_{z_j} - X_{S,j}\hat{\alpha} - H_T\hat{u}_{z_j} - \tilde{H}_{S,j}\hat{w}_J\|^2 + \lambda(\beta_1 + \ldots + \beta_C)$$

where λ is the Lagrange multipier. The estimates of the spatial fixed effects are

$$\hat{\alpha} = \frac{1}{T}(\Phi_S'\Phi_S)^{-1}\sum_{j=1}^{S}\Phi_S'\left(Y_j - \mathbf{1}_T\hat{\mu}_{0,j} - \Phi_T\hat{\beta}_{z_{s_j}} - H_T\hat{u}_{z_{s_j}} - \tilde{H}_{S,j}\hat{w}_J\right).$$

Denote $\varepsilon = Y - \mathbf{I}\mu_0 - X_T\beta - X_S\alpha - H\gamma$, the variance component of the measurement error is estimated by

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{ST}E[\varepsilon'\varepsilon|Y, Z] = \frac{1}{ST}\hat{\varepsilon}'\hat{\varepsilon} + \frac{1}{ST}\text{trace}[Hcov(\gamma|Y, Z)H'].$$

The variance components of the random effect γ are estimated by maximizing the expectation of $f(\gamma|Y, Z; \sigma_1^2, \ldots, \sigma_K^2, \sigma_s^2)$. We derive that

$$\hat{\sigma}_k^2 = \frac{1}{T}E[u_k'u_k|Y, Z] = \frac{1}{T}\text{trace}[cov(u_k|Y, Z)] + \frac{1}{T}\hat{u}_k'\hat{u}_k,$$
$$\hat{\sigma}_s^2 = \frac{1}{J}E[w_J'w_J|Y, Z] = \frac{1}{J}\text{trace}[cov(w_J|Y, Z)] + \frac{1}{J}\hat{w}_J'\hat{w}_J.$$

The interaction parameter ψ in the Gibbs distribution 6 is computed through maximizing the pseuodo-likehhood.

$$E[f(z_1, \ldots, z_S)|Y] \approx \prod_{j=1}^{S}E[f(z_{s_j}|z_{\partial s_j}; \psi)]$$

It does not have an explicit formula; we use a numeric approach to estimate this parameter.

APPENDIX B

SERVICE ACCESSIBILITY CLUSTERING: COMPARATIVE PLOTS

In this section, we complement our discussion about the service accessibility clustering in California and Georgia with two sets of comparative plots. The first set of plots shows the accessibility curves assigned to clusters as identified with the method introduced in this paper. Each figure corresponds to one cluster and the highlighted red line is the cluster mean. Similarly, the second set of plots shows the accessibility curves assigned to clusters generated by the comparative clustering method. We applied the Fclust method after re-scaling the curves to cluster by shape regardless of scale.

From the set plots for California, we conclude:

• The clustering method introduced in this paper assigns most of the flat curves in cluster 3 which consists of about 83% of the accessibility curves. The rest of the clusters have defined patterns with no overlapping patterns across clusters. (Figure 10)

• The comparative clustering method, *Mclust*, assigns a large percentage of the flat curves throughout all clusters. Subsequently, most of the cluster patterns are flat without significant differences between clusters. (Figure 11)

• Because of numerical instability, we could not obtain the cluster patterns for California using Fclust.

• The Rand index $\mathcal{R}(\tilde{\mathcal{C}}, \hat{\mathcal{C}})$ is about 0.575 where $\tilde{\mathcal{C}}$ is the Mclust clustering membership and $\hat{\mathcal{C}}$ is the FSCM clustering membership.

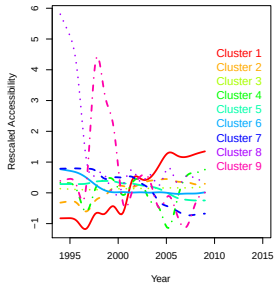
From the set plots for Georgia, we conclude:

• The clustering method introduced in this paper assigns most of the flat curves in cluster 2 which consists of about 83% of the accessibility curves. The cluster patterns are similar for Clusters 5 and 6, and for Clusters 2 and 4. (Figure 12)

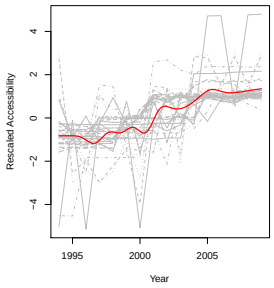
• The comparative clustering method, *Mclust*, assigns a large percentage of the flat curves throughout all clusters. Subsequently, most of the cluster patterns are flat (except Cluster 1) without significant differences between clusters. The accessibility functions are approximately uniformly spread over the 7 clusters. (Figure 11)

• The comparative clustering method, *Fclust*, discovers similar patterns to FSCM but smoother. Similarly, the largest cluster consists of the constant curves with a small number of curves assigned to the other 6 clusters. The cluster patterns are similar for Clusters 1 and 2, for Clusters 3 and 4, and for Clusters 6 and 7. The outlying cluster 7 provided by FCSM is not discovered by Fclust (Figure 14)

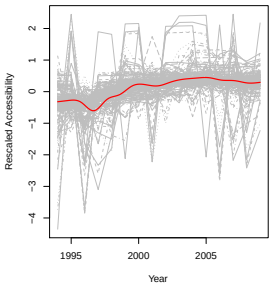
• The Rand index $\mathcal{R}(\tilde{\mathcal{C}}, \hat{\mathcal{C}})$ is about 0.413 when $\tilde{\mathcal{C}}$ is the Mclust clustering membership and it is about 0.161 when $\tilde{\mathcal{C}}$ is the Fclust clustering membership.



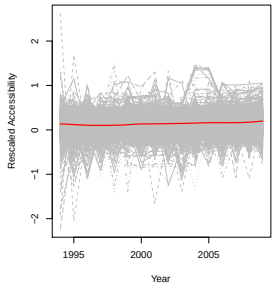
(a) Cluster Patterns



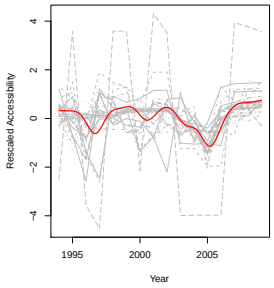
(b) Cluster 1 (39)



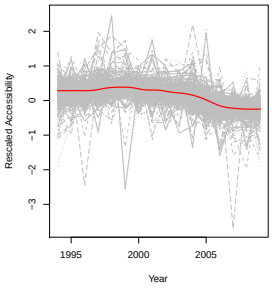
(c) Cluster 2 (196)



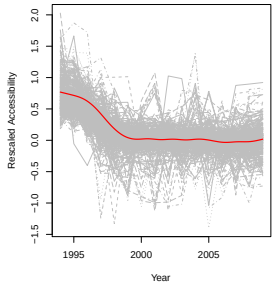
(d) Cluster 3 (5914)



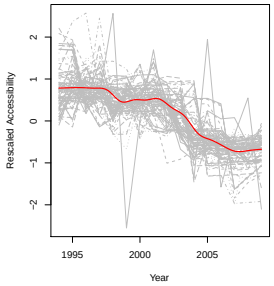
(e) Cluster 4 (30)



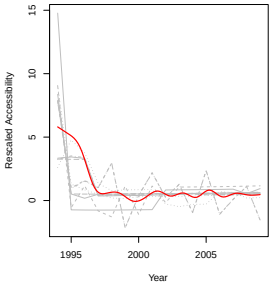
(f) Cluster 5 (408)



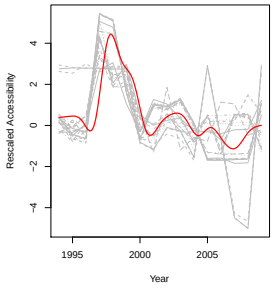
(g) Cluster 6 (403)



(h) Cluster 7 (91)

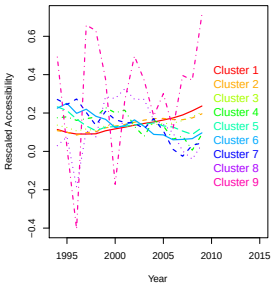


(i) Cluster 8 (11)

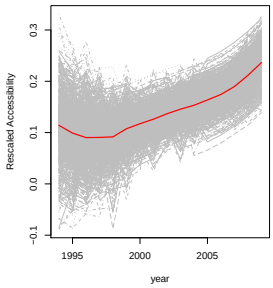


(j) Cluster 9 (23)

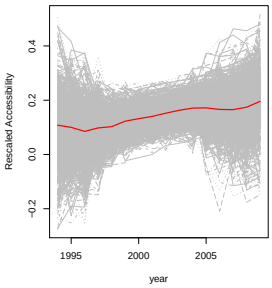
Figure 10: California: $\mu_k(t)$ for 9 Clusters provided by FSCM



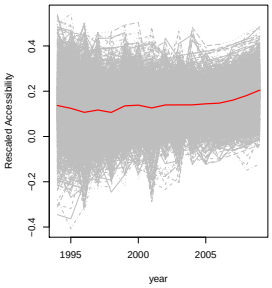
(a) Cluster Patterns



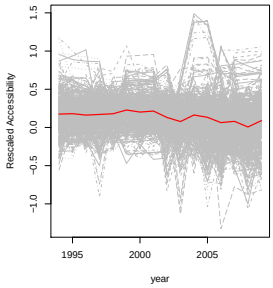
(b) Cluster 1 (901)



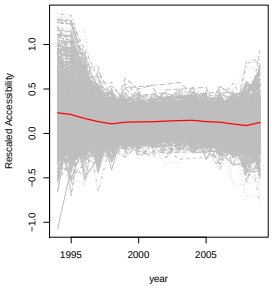
(c) Cluster 2 (1242)



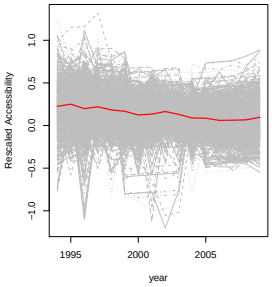
(d) Cluster 3 (1740)



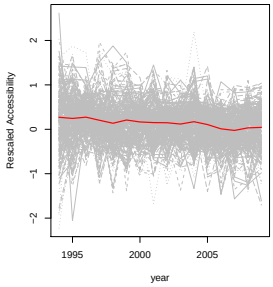
(e) Cluster 4 (509)



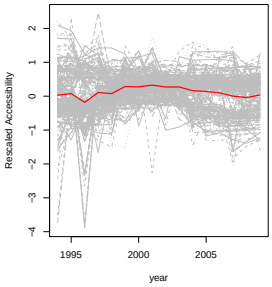
(f) Cluster 5 (1397)



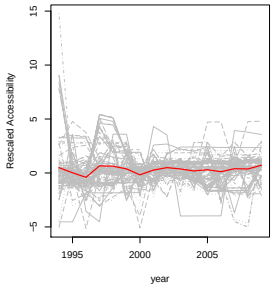
(g) Cluster 6 (638)



(h) Cluster 7 (294)

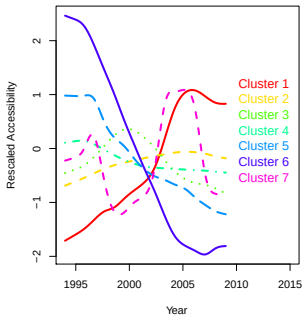


(i) Cluster 8 (279)

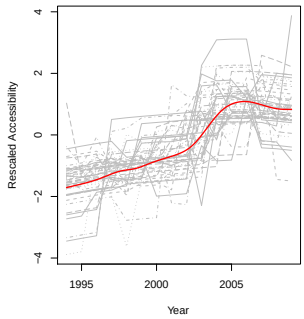


(j) Cluster 9 (115)

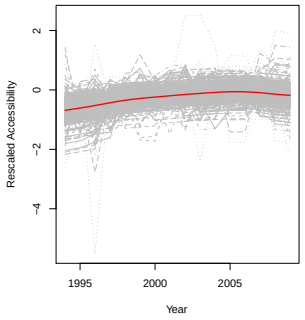
Figure 11: California: $\mu_k(t)$ for 9 Clusters provided by Mclust.



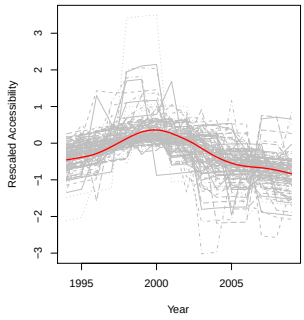
(a) Cluster Patterns $\mu_k(t)$



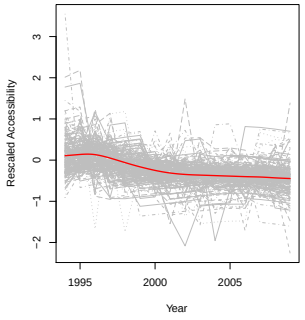
(b) Cluster 1 (50)



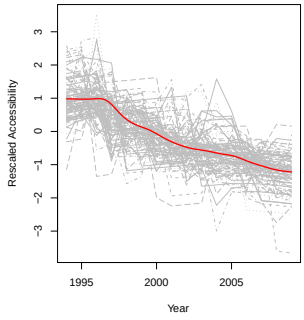
(c) Cluster 2 (1062)



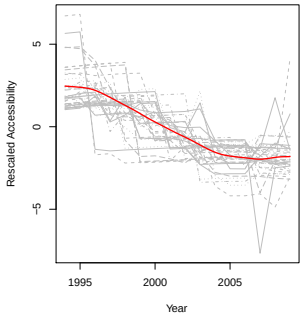
(d) Cluster 3 (119)



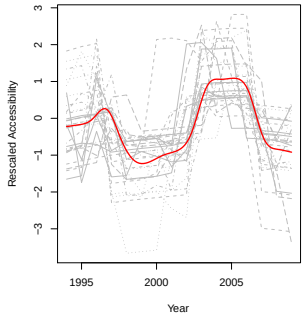
(e) Cluster 4 (216)



(f) Cluster 5 (111)

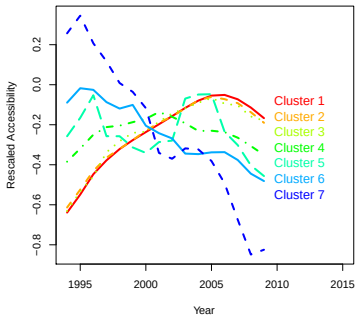


(g) Cluster 6 (37)

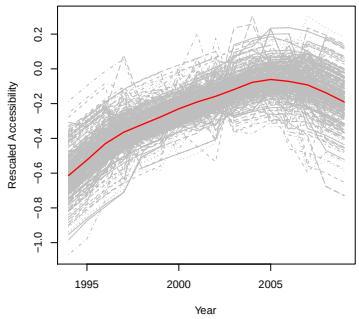


(h) Cluster 7 (29)

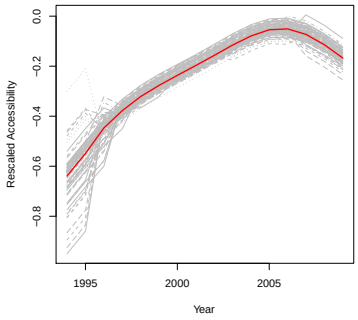
Figure 12: Georgia: $\mu_k(t)$ for 7 Clusters provided by FSCM



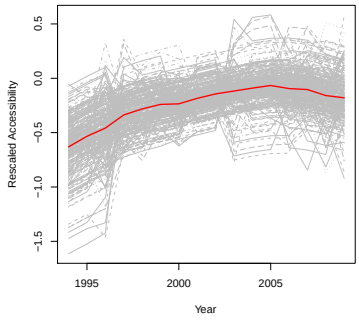
(a) Cluster Patterns



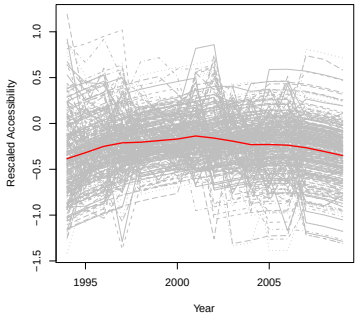
(b) Cluster 1 (294)



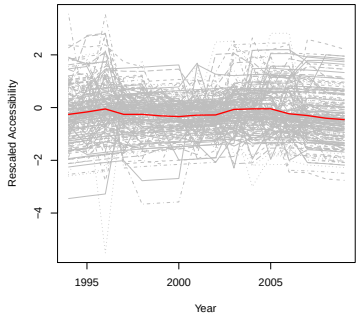
(c) Cluster 2 (266)



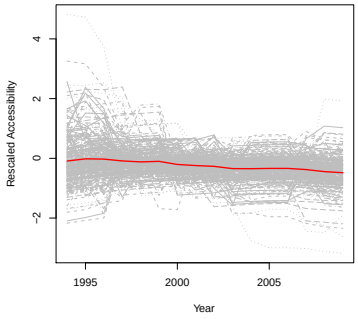
(d) Cluster 3 (172)



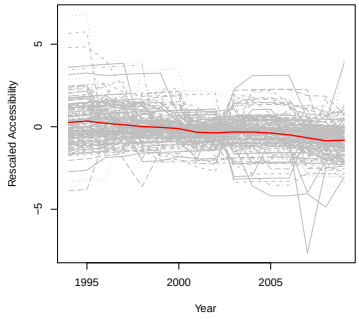
(e) Cluster 4 (309)



(f) Cluster 5 (176)

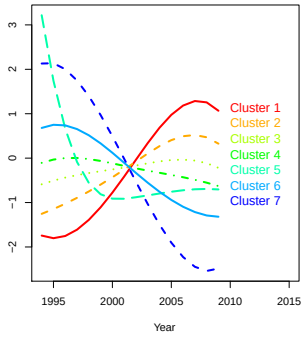


(g) Cluster 6 (294)

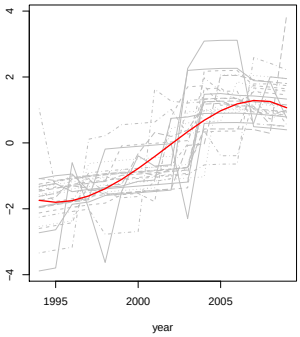


(h) Cluster 7 (113)

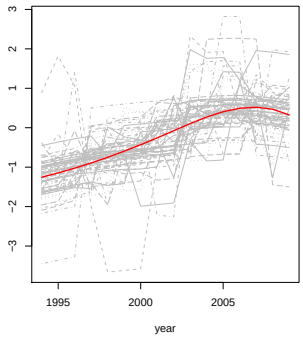
Figure 13: Georgia: $\mu_k(t)$ for 7 Clusters provided by MClust.



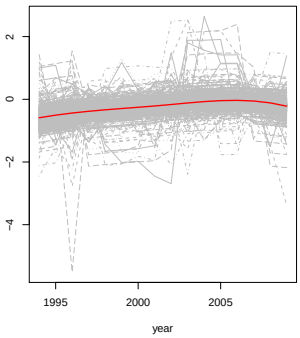
(a) Cluster Patterns



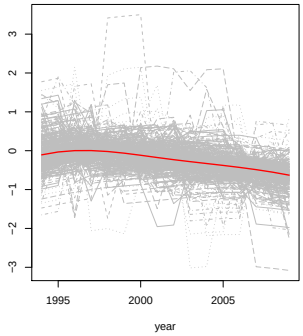
(b) Cluster 1 (25)



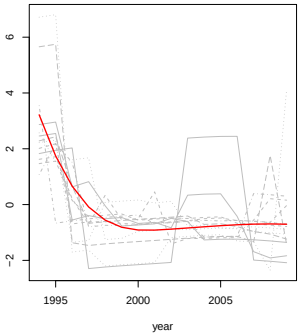
(c) Cluster 2 (66)



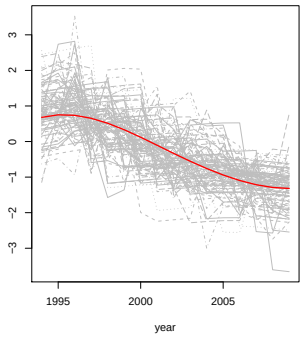
(d) Cluster 3 (1024)



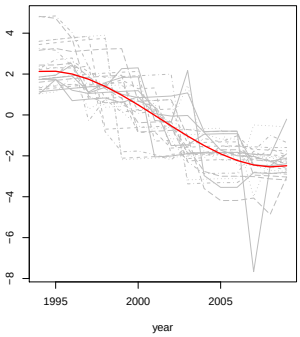
(e) Cluster 4 (362)



(f) Cluster 5 (13)



(g) Cluster 6 (110)



(h) Cluster 7 (24)

Figure 14: Georgia: $\mu_k(t)$ for 7 Clusters provided by Fclust.

APPENDIX C

PROOF OF THEOREM 1

Denote $M_t = \frac{1}{m}\Phi'\Phi$ and $M_s = \frac{1}{n}\Psi'\Psi$, then

$$\{M_t\}_{l,l'} = \frac{1}{m}\sum_{i=1}^{m}\phi_l(t_i)\phi_{l'}(t_i), \{M_s\}_{k,k'} = \frac{1}{n}\sum_{j=1}^{n}\psi_l(s_j)\psi_{k'}(s_j).$$

In order to show the asymptotic result in Theorem 1, we first need the following proposition.

PROPOSITION 1. *Under the assumption (A.1), (A.2) and (A.3), the elements of* M_T^{-1} *and* M_S^{-1} *are bounded.*

Proof:

We show the stated results for the matrix M_s. Applying the eigen-decomposition $M_s = Q\Lambda Q'$, then $M_s^{-1} = Q\Lambda^{-1}Q'$ where Λ is the diagonal matrix of the eigenvalues $\lambda_1, \ldots, \lambda_{n_S}$ and Q is the unitary matrix where its kth column q_k is the corresponding eigenvector of λ_k. Since the kk'th element of M_s is $m_{kk'}^S = \frac{1}{n}\sum_{j=1}^{n} K_S(\|s_j - \kappa_k^S\|)K_S(\|s_j - \kappa_{k'}^S\|)$, we find that the difference between two elements within the same row of matrix M_s is bounded

$$m_{kk_1}^S - m_{kk_2}^S = \frac{1}{n}\sum_{j=1}^{n} K_S(\|s_j - \kappa_k^S\|)[K_S(\|s_j - \kappa_{k_1}^S\|) - K_S(\|s_j - \kappa_{k_2}^S\|)]$$

Under (A.2), when n is large, we approximate the sum using the Riemann integral

$$m_{kk_1}^S - m_{kk_2}^S \approx \int_{\mathcal{S}} K_S(\|s - \kappa_k^S\|)[K_S(\|s - \kappa_{k_1}^S\|) - K_S(\|s - \kappa_{k_2}^S\|)]ds$$

We apply first order Taylor expansion on $K_S(\|s - \kappa_k^S\|)$ at 0 as the follows,

$$K_S(\|s - \kappa_k^S\|) = \sum_{p=0}^{p_s}\frac{1}{p!}K_S^{(p)}(0)(\|s - \kappa_k^S\|)^p + o((\|s - \kappa_k^S\|)^{p_s}).$$

Then

$$
\begin{aligned}
|m^S_{kk_1} - m^S_{kk_2}| &\approx |\int_{\mathcal{S}} K_S(\|s-\kappa^S_k\|)\sum_{p=1}^{p_s}\frac{1}{p!}K_S^{(p)}(0)[(\|s-\kappa^S_{k_1}\|)^p - (\|s-\kappa^S_{k_2}\|)^p]ds| \\
&\leq \int_{\mathcal{S}} |K_S(\|s-\kappa^S_k\|)\sum_{p=1}^{p_s}\frac{1}{p!}K_S^{(p)}(0)[(\|s-\kappa^S_{k_1}\|)^p - (\|s-\kappa^S_{k_2}\|)^p]|ds \\
&\leq \int_{\mathcal{S}} |K_S(\|s-\kappa^S_k\|)|\sum_{p=1}^{p_s}|\frac{1}{p!}K_S^{(p)}(0)[(\|s-\kappa^S_{k_1}\|)^p - (\|s-\kappa^S_{k_2}\|)^p]|ds \\
&\leq \int_{\mathcal{S}} |K_S(\|s-\kappa^S_k\|)|\sum_{p=1}^{p_s}\frac{1}{p!}K_S^{(p)}(0)\|\kappa^S_{k_1}-\kappa^S_{k_2}\|)^p|ds \\
&= \sum_{p=1}^{p_s}\frac{1}{p!}K_S^{(p)}(0)\|\kappa^S_{k_1}-\kappa^S_{k_2}\|)^p\int_{\mathcal{S}} K_S(\|s-\kappa^S_k\|)ds
\end{aligned}
$$

Therefore, as $\|\kappa^S_{k_1} - \kappa^S_{k_2}\| \to 0$, it follows that $m_{kk_1} - m_{kk_2} \to 0$. This implies that under $\|\kappa^S_{k_1} - \kappa^S_{k_2}\| \to 0$, the k_1th and k_2th columns are asymptotically linearly dependent, and therefore, the rank of the matrix M_s is reduced by 1. Furthermore, under $\|\kappa^S_{k_1} - \kappa^S_{k_2}\| \to 0$, the smallest eigenvalue of M_s goes to 0 and the largest eigenvalue of M_s^{-1} goes to infinity. Consequently, under (A.3) when $\|\kappa^S_{k_1} - \kappa^S_{k_2}\| > d^{(S)}$ with $d^{(S)}$ away from zero, the eigenvalues of M_s are finite.

Moreover, $|\{M_s^{-1}\}_{k,k'}| = |\sum_{j=1}^{n_S}\frac{q_{kj}q_{k'j}}{\lambda_j}| \leq \frac{1}{\min_j \lambda_j}|\sum_{j=1}^{n_S} q_{kj}q_{k'j}|$ and by Cauchy-Schwarz inequality, $|\sum_{j=1}^{n_S} q_{kj}q_{k'j}|^2 \leq \sum_{j=1}^{n_S} q^2_{kj}\sum_{j=1}^{n_S} q^2_{k'j} = 1$. The equality $\sum_{j=1}^{n_S} q^2_{kj}$ $\sum_{j=1}^{n_S} q^2_{k'j} = 1$ holds because $M_s = \frac{1}{n}\Psi'\Psi$ is a positive definite matrix, and thus Q is a unitary matrix. Furthermore, $|\{M_s^{-1}\}_{k,k'}| \leq \frac{1}{\min_j \lambda_j}$. Under the assumption (A.1), (A.2) and (A.3), since the smallest eigenvalue of M_s, $\min_j \lambda_j$ is finite, there exists a constant $M_1 > 0$ such that $|\{M_s^{-1}\}_{k,k'}| \leq M_1$.

Similar arguments apply to $\{M_t^{-1}\}_{l,l'}$, i.e., there exists a constant $M_2 > 0$ such that $|\{M_t^{-1}\}_{l,l'}| \leq M_2$ This concludes the proof of this proposition.

Proof of Theorem 1:

The mean of the penalized estimates is

$$\begin{aligned}
\mathbb{E}(\hat{\gamma}) &= [(\Psi'\Psi \otimes \Phi'\Phi + \lambda_s(I_{n_S} \otimes \Phi'\Phi) + \lambda_t(\Psi'\Psi \otimes I_{m_T})]^{-1} (\Psi \otimes \Phi)'\mathbb{E}(Y) \\
&= [(nM_s \otimes mM_t + \lambda_s(I_{n_S} \otimes mM_t) + \lambda_t(nM_s \otimes I_{m_T})]^{-1} (\Psi \otimes \Phi)'(\Psi \otimes \Phi)\gamma \\
&= [(nM_s \otimes mM_t + \lambda_s(I_{n_S} \otimes mM_t) + \lambda_t(nM_s \otimes I_{m_T})]^{-1} (nM_s \otimes mM_t)\gamma \\
&= \left[M_s \otimes M_t + \frac{1}{n}\lambda_s(I_{n_S} \otimes M_t) + \frac{1}{m}\lambda_t(M_s \otimes I_{m_T})\right]^{-1} (M_s \otimes M_t)\gamma \\
&= \left[I + \frac{1}{n}\lambda_s(I_{n_S} \otimes M_t)(M_s \otimes M_t)^{-1} + \frac{1}{m}\lambda_t(M_s \otimes I_{m_T})(M_s \otimes M_t)^{-1}\right]^{-1} \gamma \\
&= \left[I + \frac{1}{n}\lambda_s(M_s^{-1} \otimes I_{m_T}) + \frac{1}{m}\lambda_t(I_{n_S} \otimes M_t^{-1})\right]^{-1} \gamma
\end{aligned}$$

Therefore, the mean of $\hat{\gamma}$ is $\mathbb{E}(\hat{\gamma}) = (I + \lambda_s\tilde{B}_s^{-1} + \lambda_t\tilde{B}_t^{-1})^{-1}\gamma$ and the bias is $\mathbb{B}(\hat{\gamma}) = \mathbb{E}(\hat{\gamma}) - \gamma = [(I + \lambda_s\tilde{B}_s^{-1} + \lambda_t\tilde{B}_t^{-1})^{-1} - I]\gamma$.

Applying Sherman-Morrison-Woodbury formula, we have

$$\begin{aligned}
\mathbb{B}(\hat{\gamma}) &= -[I + (\lambda_s\tilde{B}_s^{-1} + \lambda_t\tilde{B}_t^{-1})^{-1}]^{-1}\gamma \\
&= -\{I + [\frac{1}{n}\lambda_s(M_s^{-1} \otimes I_{m_T}) + \frac{1}{m}\lambda_t(I_{n_S} \otimes M_t^{-1})]^{-1}\}^{-1}\gamma
\end{aligned}$$

When the temporal sample size m goes to infinity, we get

$$\begin{aligned}
\mathbb{B}(\hat{\gamma}) &= -\{I + m[\frac{m}{n}\lambda_s(M_s^{-1} \otimes I_{m_T}) + \lambda_t(I_{n_S} \otimes M_t^{-1})]^{-1}\}^{-1}\gamma \\
&= -\frac{1}{m}\{\frac{1}{m}I + [\frac{m}{n}\lambda_s(M_s^{-1} \otimes I_{m_T}) + \lambda_t(I_{n_S} \otimes M_t^{-1})]^{-1}\}^{-1}\gamma \\
&\to -\frac{1}{m}\{[\frac{m}{n}\lambda_s(M_s^{-1} \otimes I_{m_T}) + \lambda_t(I_{n_S} \otimes M_t^{-1})]^{-1}\}^{-1}\gamma \\
&= -\frac{1}{m}[\frac{m}{n}\lambda_s(M_s^{-1} \otimes I_{m_T}) + \lambda_t(I_{n_S} \otimes M_t^{-1})]\gamma \\
&= -[\frac{1}{n}\lambda_s(M_s^{-1} \otimes I_{m_T}) + \frac{1}{m}\lambda_t(I_{n_S} \otimes M_t^{-1})]\gamma \\
&\to -\frac{1}{n}\lambda_s(M_s^{-1} \otimes I_{m_T})\gamma = -\lambda_s\tilde{B}_s\gamma
\end{aligned}$$

Similarly, as the spatial sample size n goes to infinity, we get

$$\mathbb{B}(\hat{\gamma}) \to -\frac{1}{m}\lambda_t(I_{n_S} \otimes M_t^{-1})\gamma = -\lambda_t\tilde{B}_t\gamma$$

When both m and n go to infinity, we get

$$\begin{aligned}
\mathbb{B}(\hat{\gamma}) &= -\{I + mn[m\lambda_s(M_s^{-1} \otimes I_{m_T}) + n\lambda_t(I_{n_S} \otimes M_t^{-1})]^{-1}\}^{-1}\gamma \\
&= -\frac{1}{mn}\{\frac{1}{mn}I + [m\lambda_s(M_s^{-1} \otimes I_{m_T}) + n\lambda_t(I_{n_S} \otimes M_t^{-1})]^{-1}\}^{-1}\gamma \\
&\to -\frac{1}{mn}[m\lambda_s(M_s^{-1} \otimes I_{m_T}) + n\lambda_t(I_{n_S} \otimes M_t^{-1})]\gamma \\
&= -[\frac{1}{n}\lambda_s(M_s^{-1} \otimes I_{m_T}) + \frac{1}{m}\lambda_t(I_{n_S} \otimes M_t^{-1})]\gamma \to 0
\end{aligned}$$

The variance of the penalized estimates γ are

$$\begin{aligned}
\mathbb{V}(\hat{\gamma}) &= \sigma_\epsilon^2(B'B + \lambda_s\tilde{B}_s + \lambda_t\tilde{B}_t)^{-1}B'B(B'B + \lambda_s\tilde{B}_s + \lambda_t\tilde{B}_t)^{-1} \\
&= \sigma_\epsilon^2\left[(mnM_s \otimes M_t + \lambda_s m(I_K \otimes M_t) + \lambda_t(nM_s \otimes I_L)\right]^{-1}(mnM_s \otimes M_t) \\
&[(mnM_s \otimes M_t + \lambda_s m(I_K \otimes M_t) + \lambda_t(nM_s \otimes I_L)]^{-1} \\
&= \sigma_\epsilon^2\left[I_{KL} + \frac{1}{n}\lambda_s M_s^{-1} \otimes I_L + \frac{1}{m}\lambda_t I_K \otimes M_t^{-1}\right]^{-1} \\
&[(mnM_s \otimes M_t + \lambda_s m(I_K \otimes M_t) + \lambda_t(nM_s \otimes I_L)]^{-1} \\
&= \sigma_\epsilon^2[mnM_s \otimes M_t + 2m\lambda_s I_K \otimes M_t + 2n\lambda_t M_s \otimes I_l + \frac{m}{n}\lambda_s^2 M_s^{-1} \otimes M_t \\
&+ 2\lambda_s\lambda_t I_{KL} + \frac{n}{m}\lambda_t^2 M_s \otimes M_t^{-1}]^{-1}
\end{aligned}$$

As the temporal sample size m goes to infinity, we get

$$\begin{aligned}
\mathbb{V}(\hat{\gamma}) &= \sigma_\epsilon^2\frac{1}{m}[nM_s \otimes M_t + 2\lambda_s I_{n_S} \otimes M_t + 2\frac{n}{m}\lambda_t M_s \otimes I_{m_T} \\
&+ \frac{1}{n}\lambda_s^2 M_s^{-1} \otimes M_t + \frac{2}{m}\lambda_s\lambda_t I_{n_S m_T} + \frac{n}{m^2}\lambda_t^2 M_s \otimes M_t^{-1}]^{-1} \\
&\to \sigma_\epsilon^2\frac{1}{m}[nM_s \otimes M_t + 2\lambda_s I_{n_S} \otimes M_t + \frac{1}{n}\lambda_s^2 M_s^{-1} \otimes M_t]^{-1} \\
&= \sigma_\epsilon^2[nmM_s \otimes M_t + 2m\lambda_s I_{n_S} \otimes M_t + \frac{m}{n}\lambda_s^2 M_s^{-1} \otimes M_t]^{-1} \\
&= \sigma_\epsilon^2[B'B + 2\lambda_s\tilde{B}_s + \lambda_s^2\tilde{B}_s\tilde{B}_t^{-1}]^{-1}
\end{aligned}$$

Similarly, as the spatial size n goes to infinity, we have

$$\mathbb{V}(\hat{\gamma}) \to \sigma_\epsilon^2[B'B + 2\lambda_t\tilde{B}_t + \lambda_t^2\tilde{B}_t\tilde{B}_s^{-1}]^{-1}$$

As both m and n go to infinity, we have

$$\begin{aligned}\mathbb{V}(\hat{\gamma}) &= \sigma_\epsilon^2 \frac{1}{mn}[M_s \otimes M_t + \frac{2}{n}\lambda_s I_K \otimes M_t + \frac{2}{m}\lambda_t M_s \otimes I_l \\ &\quad + \frac{1}{n^2}\lambda_s^2 M_s^{-1} \otimes M_t + \frac{2}{mn}\lambda_s\lambda_t I_{KL} + \frac{1}{m^2}\lambda_t^2 M_s \otimes M_t^{-1}]^{-1} \\ &\quad \to \sigma_\epsilon^2 \frac{1}{mn}[M_s \otimes M_t]^{-1} = \sigma_\epsilon^2 (B'B)^{-1}\end{aligned}$$

APPENDIX D

PROOF OF THEOREM 2

Proof: We apply first order Taylor expansion on $\hat{\rho}(t, t+\eta)$ at $\gamma^x_{k,l}$ and $\gamma^y_{k',l'}$ as follows

$$\hat{\rho}(t,t+\eta) = \rho(t,t+\eta) + \sum_k \sum_l \frac{\partial \rho(t,t+\eta)}{\partial \gamma^x_{kl}} (\hat{\gamma_{kl}}^x - \gamma^x_{kl}) + \sum_{k'} \sum_{l'} \frac{\partial \rho(t,t+\eta)}{\partial \gamma^y_{k'l'}} (\hat{\gamma}^y_{k'l'} - \gamma^y_{k'l'}) \tag{37}$$

$$\mathbb{E}[\hat{\rho}(t,t+\eta)] - \rho(t,t+\eta) = \sum_k \sum_l \frac{\partial \rho(t,t+\eta)}{\partial \gamma^x_{kl}} \mathbb{B}[\hat{\gamma}^x_{kl}] + \sum_{k'} \sum_{l'} \frac{\partial \rho(t,t+\eta)}{\partial \gamma^y_{k'l'}} \mathbb{B}[\hat{\gamma}^y_{k'l'}]$$

From Theorem 1, we know that as $m \to \infty$ and $n \to \infty$, $\mathbb{E}[\hat{\gamma}_{k,l}] - \gamma_{k,l} \to 0$ and thus

$$\mathbb{E}[\hat{\rho}(t,t+\eta)] \to \rho(t,t+\eta).$$

Similarly, we can derive $\mathbb{E}[\hat{\rho}(s,s+\delta)] \to \rho(s,s+\delta)$.

Given the equations (37), the variances of $\hat{\rho}(t,t+\eta)$ is approximately

$$\begin{aligned}
\mathbb{V}[\hat{\rho}(t,t+\eta)] &\approx \mathbb{V}[\rho(t,t+\eta) + \sum_k \sum_l \frac{\partial \rho(t,t+\eta)}{\partial \gamma^x_{kl}} (\hat{\gamma}^x_{kl} - \gamma^x_{kl}) + \sum_{k'} \sum_{l'} \frac{\partial \rho(t,t+\eta)}{\partial \gamma^y_{k'l'}} (\hat{\gamma}^y_{k'l'} - \gamma^y_{k'l'})] \\
&= \mathbb{V}[\sum_k \sum_l \frac{\partial \rho(t,t+\eta)}{\partial \gamma^x_{kl}} \hat{\gamma}^x_{kl} + \sum_{k'} \sum_{l'} \frac{\partial \rho(t,t+\eta)}{\partial \gamma^y_{k'l'}} \hat{\gamma}^y_{k'l'}] \\
&= \mathbb{V}[\sum_k \sum_l \frac{\partial \rho(t,t+\eta)}{\partial \gamma^x_{kl}} \hat{\gamma}^x_{kl}] + \mathbb{V}[\sum_{k'} \sum_{l'} \frac{\partial \rho(t,t+\eta)}{\partial \gamma^y_{k'l'}} \hat{\gamma}^y_{k'l'}] \\
&= D_t^{x'} \mathbb{V}(\hat{\gamma}^x) D_t^x + D_t^{y'} \mathbb{V}(\hat{\gamma}^y) D_t^y
\end{aligned}$$

where $D_t = (\frac{\partial \rho(t,t+\eta)}{\partial \gamma_{kl}}, k = 0, \ldots, n_S; l = 0, \ldots, m_T)'$. As $m \to \infty$ and $n \to \infty$,

$$\mathbb{V}[\hat{\rho}(t,t+\eta)] \to D_t^{x'} \sigma^2_{\epsilon,x} (B'B)^{-1} D_t^x + D_t^{y'} \sigma^2_{\epsilon,y} (B'B)^{-1} D_t^y.$$

Similarly, $\mathbb{V}[\hat{\rho}(s,s+\delta)] \to D_s^{x'} \sigma^2_{\epsilon,x} (B'B)^{-1} D_s^x + D_s^{y'} \sigma^2_{\epsilon,y} (B'B)^{-1} D_s^y$ where $D_s = (\frac{\partial \rho(s,s+\delta)}{\partial \gamma_{kl}}, k = 0, \ldots, n_S; l = 0, \ldots, m_T)'$.

APPENDIX E

ESTIMATION ALGORITHM FOR MULTI-LEVEL FUNCTIONAL CLUSTERING MODEL

In this appendix, we describe the estimation algorithm for the clustering model parameters in *Model 1* and *Model 2*. For this we join the two models in a more general model which assumes that there is clustering at both levels. Based on the derivations under level-1 clustering and level-2 clustering models, we generalize to allow for simultaneous clustering at level 1 and 2. The general clustering model becomes:

$$\begin{cases} X_{ij}(t) = \sum_{s=1}^{N_1} \xi_{i,s}\phi_s^{(1)}(t) + \sum_{r=1}^{N_2} \zeta_{ij,r}\phi_r^{(2)}(t) + \varepsilon_{ij}(t) \\ \nu_{i,s,k} = \xi_{i,s}|(Z_i^{(1)} = k) \sim N(\mu_{s,k}, \lambda_{s,k}^{(1)}) \text{ and } Z_i^{(1)} \sim \text{ Multinomial}(1; \pi_1^{(1)}, \ldots, \pi_{C_1}^{(1)}) \\ \delta_{ij,r,k} = \zeta_{ij,r}|(Z_i^{(2)} = k) \sim N(\eta_{j,k}, \Lambda_{j,k}^{(1)}) \text{ and } Z_i^{(2)} \sim \text{ Multinomial}(1; \pi_1^{(2)}, \ldots, \pi_{C_2}^{(2)}) \end{cases} \tag{38}$$

under two constraints

$$\begin{cases} \sum_{k=1}^{C_1} \pi_k^{(1)} \mu_{s,k} = 0 \\ \sum_{k=1}^{C_2} \pi_k^{(2)} \eta_{j,r,k} = 0 \end{cases}$$

We denote the following set of parameters

$$\theta_{Z^{(1)}} = (\pi_1^{(1)}, \ldots, \pi_{C_1}^{(1)}) \text{ specifying the distr. of } Z^{(1)}$$

$$\theta_{Z^{(2)}} = (\pi_1^{(2)}, \ldots, \pi_{C_2}^{(2)}) \text{ specifying the distr. of } Z^{(2)}$$

$$\theta_\xi = \{\lambda_{s,k}^{(1)}, \mu_{s,k}\}_{s=1,\ldots,N_1} \text{ specifying the distr. of } \xi_{i,s}|(Z^{(1)} = k) \text{ for } k = 1, \ldots, C_1$$

$$\theta_\zeta = \{\lambda_{j,s,k}^{(2)}, \eta_{j,r,k}\}_{r=1,\ldots,N_2} \text{ specifying the distr. of } \zeta_{ij,r}|(Z^{(1)} = k)$$

$$\text{for } k = 1, \ldots, C_1, \ j = 1, \ldots, J$$

Remark: For the particular cases $C_1 = 1$ or $C_2 = 1$, the general model reduces to the clustering models for level 1 ($C_2 = 1$) and to the clustering models for level 2

($C_1 = 1$) because of the two constrains. Specifically, when $C_1 = 1$, the constraint becomes $\pi_1^{(1)}\mu_{1,k} = 0$ and since $\pi_1^{(1)} = 1$ then $\mu_{1,k} = 0$ and $\lambda_{s,k}^{(1)} = \tau_s^{(1)}$. Similarly, for $C_2 = 1$, the constraints imply that $\eta_{1,k,j} = 0$ for $j = 1, \ldots, J$ and $\lambda_{j,r,k}^{(2)} = \tau_{r,j}^{(2)}$.

The estimation algorithm for the general model is a two-step (EM) iterative procedure to maximize the likelihood of the observed likelihood

$$L(\theta_{Z^{(1)}}, \theta_{Z^{(2)}}, \theta_\xi, \theta_\zeta, \sigma^2) = \prod_{i=1}^{I} \sum_{k=1}^{C_1} \pi_k^{(1)} \sum_{k'=1}^{C_2} \pi_{k'}^{(2)} f(X_i; \theta_{\xi_k}, \theta_{\zeta_{k'}}, \sigma^2)$$

where $X_i \sim N(\Phi_i^{(1)}\mu_k + \Phi_i^{(2)}\eta_{k'}, \sigma^2 I_{JN})$.

The EM algorithm converges to the global maximum of the observed likelihood $L(\theta_{Z^{(1)}}, \theta_{Z^{(2)}}, \theta_\xi, \theta_\zeta, \sigma^2)$ by iteratively imputing the latent variables in the E-step and maximizing the expectation of the likelihood of the complete data conditional on the observed data in the M-step. Briefly, the EM algorithm for our clustering model is

• At the E-step, impute the latent variables $Z^{(1)}$, $Z^{(2)}$, $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$ given the parameter estimates based on the conditional expectation of the complete likelihood

$$L_C(\theta_{Z^{(1)}}, \theta_{Z^{(2)}}, \theta_\xi, \theta_\zeta, \sigma^2) =$$

$$f(X|Z^{(1)}, Z^{(2)}, \boldsymbol{\xi}, \boldsymbol{\zeta}) f(\boldsymbol{\xi}|Z^{(1)}) f(\boldsymbol{\zeta}|Z^{(2)}) f(Z^{(1)}) f(Z^{(2)}) =$$

$$\prod_i^I [f(X_i|\boldsymbol{\xi}_i, \boldsymbol{\zeta}_i, Z_i^{(1)}, Z_i^{(2)}; \sigma^2) f(\boldsymbol{\xi}_i|Z_i^{(1)}; \theta_\xi) f(\boldsymbol{\zeta}_i|Z_i^{(2)}; \theta_\zeta) f(Z_i^{(1)}; \theta_{Z^{(1)}}) f(Z_i^{(2)}; \theta_{Z^{(2)}})].$$

• At the M-step, estimate the model parameters by maximizing the conditional expectation of the complete likelihood $E[L_C(\theta_{Z^{(1)}}, \theta_{Z^{(2)}}, \theta_\xi, \theta_\zeta, \sigma^2|X_1, \ldots, X_I]$.

Initialization. Because this in an iterative algorithm, we need first to input initial estimates for the model parameters. Using MFPCA, we obtain initial estimates for the unconditional scores $\boldsymbol{\xi}_{i,s}$ and $\boldsymbol{\zeta}_{ij,r}$ and initial estimates for their variances $\tau_s^{(1)}$ and $\tau_{j,r}^{(1)}$, respectively. The MFPCA will also provide the set of eigenfunctions describing the spectral decomposition of the between and within covariance functions. In the estimation algorithm, we assume the eigenfunctions at levels 1 and 2 are fixed and

denote

$$\begin{aligned}
\Phi_{ij,m}^{(1)} &= (\phi_1^{(1)}(t_{ij,m}), \ldots, \phi_{N_1}^{(1)}(t_{ij,m}))' \text{ for } m = 1, \ldots, n_{ij}(N_1 \times 1) \\
\Phi_{ij}^{(1)} &= (\Phi_{ij,1}^{(1)}, \ldots, \Phi_{ij,n_{ij}}^{(1)})' \text{ for } j = 1, \ldots, J(n_{ij} \times N_1) \\
\Phi_{i}^{(1)} &= (\Phi_{i1}^{(1)'}, \ldots, \Phi_{iJ_i}^{(1)'})' \text{ for } i = 1, \ldots, I(Jn_{ij} \times N_1) \\
\Phi_{ij,m}^{(2)} &= (\phi_1^{(2)}(t_{ij,m}), \ldots, \phi_{N_2}^{(2)}(t_{ij,m}))' \text{ for } m = 1, \ldots, n_{ij}(N_2 \times 1) \\
\Phi_{ij}^{(2)} &= (\Phi_{ij,1}^{(2)}, \ldots, \Phi_{ij,n_{ij}}^{(2)})' \text{ for } j = 1, \ldots, J(n_{ij} \times N_2) \\
\Phi_{i}^{(2)} &= diag(\Phi_{i1}^{(2)}, \ldots, \Phi_{iJ_i}^{(2)}) \text{ for } i = 1, \ldots, I(Jn_{ij} \times JN_2)
\end{aligned}$$

where $t_{ij,1}, \ldots, t_{ij,m}$ are the observation time points of case i and measurement j for $i = 1, \ldots, I$ and $j = 1, \ldots, J$.

Estimation. The estimation algorithm is described using the following data-vector and unconditional model parameters notations:

- Observation for case i and measurement j at time $t_{ij,m}$: $X_{ij,m} = X_{ij}(t_{ij,m})$ for $m = 1, \ldots, n_{ij}$.
- Vector of observations for case i and measurement j: $X_{ij} = (X_{ij,1}, \ldots, X_{ij,n_{ij}})'$ for $j = 1, \ldots, J$.
- Vector of all observations for case i: $X_i = (X_{i1}', \ldots, X_{iJ})'$ for $i = 1, \ldots, I$.
- Vectors of unconditional scores $\boldsymbol{\xi}_i = (\xi_{i,1}, \ldots, \xi_{i,N_1})'$ and $\boldsymbol{\zeta}_i = (\zeta_{i1,1}, \ldots, \zeta_{i1,N_2}, \ldots, \zeta_{iJ,1}, \ldots, \zeta_{iJ,N_2})'$.

Following these notations, the unconditional multi-level model becomes

$$X_i = \Phi_i^{(1)} \boldsymbol{\xi}_i + \Phi_i^{(2)} \boldsymbol{\zeta}_i + \epsilon_i.$$

We estimate the model parameters by maximizing the expectation of the log-likelihood for $(X_i, Z_i^{(1)}, Z_i^{(2)}, \boldsymbol{\xi}_i, \boldsymbol{\zeta}_i)$ with $i = 1, \ldots, I$.

$$
\begin{aligned}
&l_C(\theta_{Z^{(1)}}, \theta_{Z^{(2)}}, \theta_\xi, \theta_\zeta, \sigma^2) = -2 \log L_C(\theta_{Z^{(1)}}, \theta_{Z^{(2)}}, \theta_\xi, \theta_\zeta, \sigma^2) \\
&= -2 \sum_{i=1}^{I} [\log f(X_i | Z_i^{(1)}, Z_i^{(2)}, \boldsymbol{\xi}_i, \boldsymbol{\zeta}_i; \sigma^2) + \log f(\boldsymbol{\xi}_i | Z_i^{(1)}; \theta_\xi) \\
&+ \log f(\boldsymbol{\zeta}_i | Z_i^{(2)}; \theta_\zeta) + \log f(Z^{(1)_i}; \theta_{Z^{(1)}}) + \log f(Z_i^{(2)}; \theta_{Z^{(2)}})] \\
&= \sum_{i=1}^{I} \sum_{k=1}^{C_1} Z_{ik}^{(1)} \sum_{k'=1}^{C_2} Z_{ik'}^{(2)} [n_{ij} J \log(\sigma^2) + \| X_i - \Phi_i^{(1)} \boldsymbol{\nu}_{ik} - \Phi_i^{(2)} \boldsymbol{\delta}_{ik'} \|^2 / \sigma^2] \\
&+ \sum_{i=1}^{I} \sum_{k=1}^{C_1} Z_{ik}^{(1)} [\log |\Lambda_k^{(1)}| + (\boldsymbol{\xi}_i - \mu_k)' \Lambda_k^{(1),-1} (\boldsymbol{\xi}_i - \mu_k)] \\
&+ \sum_{i=1}^{I} \sum_{k'=1}^{C_2} Z_{ik'}^{(2)} [\sum_{j=1}^{J_i} (\log |\Lambda_{jk'}^{(2)}| + (\boldsymbol{\zeta}_{ij} - \eta_{jk'})' \Lambda_{jk'}^{(2),-1} (\boldsymbol{\zeta}_{ij} - \eta_{jk'}))] \\
&-2 \sum_{i=1}^{I} [\sum_{k=1}^{C_1} Z_{ik}^{(1)} \log(\pi_k^{(1)}) + \sum_{k'=1}^{C_2} Z_{ik'}^{(2)} \log(\pi_{k'}^{(2)})]
\end{aligned}
$$

We estimate based on the complete likelihood using the EM algorithm.

E-step Compute the conditional expectation of the complete likelihood given $Q(\theta_{Z^{(1)}}, \theta_{Z^{(2)}}, \theta_\xi, \theta_\zeta) = E[l(\theta_{Z^{(1)}}, \theta_{Z^{(2)}}, \theta_\xi, \theta_\zeta, \sigma^2) | X_i]$:

- Compute $E[z_{ik}^{(1)} | X_i]$ and $E[Z_{ik}^{(2)} | X_i]$.

$$
\begin{aligned}
&\hat{z}_{ik}^{(1)} = E[Z_{ik}^{(1)} | X_i] = Pr(Z_{ik}^{(1)} = 1 | X_i) \\
&(*) = \frac{f(Z_{ik}^{(1)} = 1, X_i)}{f(X_i)} = \frac{\sum_{k'=1}^{C_2} f(X_i, Z_{ik}^{(1)} = 1, Z_{ik'}^{(2)} = 1)}{\sum_{k=1}^{C_1} \sum_{k'=1}^{C_2} f(X_i, Z_{ik}^{(1)} = 1, Z_{ik'}^{(2)} = 1)} \\
&= \frac{\sum_{k'=1}^{C_2} f(X_i | Z_{ik}^{(1)} = 1, Z_{ik'}^{(2)} = 1) Pr(Z_{ik}^{(1)} = 1) Pr(Z_{ik'}^{(2)} = 1)}{\sum_{k=1}^{C_1} \sum_{k'=1}^{C_2} f(X_i | Z_{ik}^{(1)} = 1, Z_{ik'}^{(2)} = 1) Pr(Z_{ik}^{(1)} = 1) Pr(Z_{ik'}^{(2)} = 1)} \\
&= \frac{\pi_k^{(1)} \sum_{k'=1}^{C_2} \pi_{k'}^{(2)} f(X_i | Z_{ik}^{(1)} = 1, Z_{ik'}^{(2)} = 1)}{\sum_{k=1}^{C_1} \pi_k^{(1)} \sum_{k'=1}^{C_2} \pi_{k'}^{(2)} f(X_i | Z_{ik}^{(1)} = 1, Z_{ik'}^{(2)} = 1)}
\end{aligned}
$$

where $X_i | (Z_{ik}^{(1)} = 1, Z_{ik'}^{(2)} = 1) \sim N(\Phi_i^{(1)} \mu_k + \Phi_i^{(2)} \eta_{jk'}, \Phi_i^{(1)} \Lambda_k^{(1)} \Phi_i^{(1)'} + \Phi_i^{(2)} \Lambda_{jk'}^{(2)} \Phi_i^{(2)'} + \sigma^2 I_{J n_{ij}})$. The equality (*) is because of the assumption that $(z_{i1}^{(1)}, \ldots z_{iC_1}^{(1)}) \sim$

Multinomial$(1, \pi_1^{(1)}, \ldots, \pi_{C_1}^{(1)})$ and $(z_{i1}^{(2)}, \ldots z_{iC_2}^{(2)}) \sim$ Multinomial$(1, \pi_1^{(2)}, \ldots, \pi_{C_2}^{(2)})$. Similarly, we derive

$$\begin{aligned}\hat{z}_{ik'}^{(2)} &= E(Z_{ik'}^{(2)}|X_i) = E(Z_{ik'}^{(2)} = 1|X_i) = Pr(Z_{ik'}^{(2)} = 1|X_i) \\ &= \frac{\pi_{k'}^{(2)} \sum_{k=1}^{C_1} \pi_k^{(1)} f(X_i|Z_{ik}^{(1)} = 1, Z_{ik'}^{(2)} = 1)}{\sum_{k=2}^{C_2} \pi_{k'}^{(2)} \sum_{k=1}^{C_1} \pi_k^{(1)} f(X_i|Z_{ik}^{(1)} = 1, Z_{ik'}^{(2)} = 1)}\end{aligned}$$

• Compute the first and second moments of $\boldsymbol{\nu}_{ik} \equiv (\boldsymbol{\xi}_i|Z_{ik}^{(1)} = 1)$ and $\boldsymbol{\delta}_{ik'} \equiv (\boldsymbol{\zeta}_i|Z_{ik'}^{(2)} = 1)$ conditional on the observed data X_i.

The first conditional moments are

$$\hat{\boldsymbol{\nu}}_{ik} = E[\boldsymbol{\xi}_i|Z_{ik}^{(1)} = 1, X_i] = \sum_{k'=1}^{C_2} \hat{z}_{ik'}^{(2)} E[\boldsymbol{\xi}_i|Z_{ik}^{(1)} = 1, Z_{ik'}^{(2)} = 1, X_i],$$

$$\hat{\boldsymbol{\delta}}_{ik'} = E[\boldsymbol{\zeta}_i|Z_{ik'}^{(2)} = 1, X_i] = \sum_{k=1}^{C_1} \hat{z}_{ik}^{(1)} E[\boldsymbol{\zeta}_i|Z_{ik}^{(1)} = 1, Z_{ik'}^{(2)} = 1, X_i],$$

and the second conditional moments are

$$\hat{\boldsymbol{\nu}\boldsymbol{\nu}'}_{ik} = E[\boldsymbol{\xi}_i\boldsymbol{\xi}_i'|Z_{ik}^{(1)} = 1, X_i] = \sum_{k'=1}^{C_2} \hat{z}_{ik'}^{(2)} E[\boldsymbol{\xi}_i\boldsymbol{\xi}_i'|Z_{ik}^{(1)} = 1, Z_{ik'}^{(2)} = 1, X_i],$$

$$\hat{\boldsymbol{\delta}\boldsymbol{\delta}'}_{ik'} = E[\boldsymbol{\zeta}_i\boldsymbol{\zeta}_i'|Z_{ik'}^{(2)} = 1, X_i] = \sum_{k=1}^{C_1} \hat{z}_{ik}^{(1)} E[\boldsymbol{\zeta}_i\boldsymbol{\zeta}_i'|Z_{ik}^{(1)} = 1, Z_{ik'}^{(2)} = 1, X_i].$$

These moments can be calculated using the distributions of the conditional scores

$$\boldsymbol{\xi}_i|(Z_{ik}^{(1)} = 1, Z_{ik'}^{(2)} = 1, X_i) \sim$$
$$N(\mu_k + (\Lambda_k^{(1),-1} + \Phi_i^{(1)'} D^{-1} \Phi_i^{(1)})^{-1} D^{-1}(X_i - \Phi_i^{(1)}\mu_k - \Phi_i^{(2)}\eta_{k'}), (\Lambda_k^{(1),-1} + \Phi_i^{(1)'} D^{-1} \Phi_i^{(1)})^{-1}),$$
$$\boldsymbol{\zeta}_i|(Z_{ik}^{(1)} = 1, Z_{ik'}^{(2)} = 1, X_i) \sim$$
$$N(\eta_{k'} + (\Lambda_{k'}^{(2),-1} + \Phi_i^{(2)'} F^{-1} \Phi_i^{(2)})^{-1} F^{-1}(X_i - \Phi_i^{(1)}\mu_k - \Phi_i^{(2)}\eta_{k'}), (\Lambda_{k'}^{(2),-1} + \Phi_i^{(2)'} F^{-1} \Phi_i^{(2)})^{-1})$$

where $D = \Phi_i^{(2)} \Lambda_{k'}^{(2)} \Phi_i^{(2)'} + \sigma^2 I_{n_{ij}}$ and $F = \Phi_i^{(1)} \Lambda_k^{(1)} \Phi_i^{(1)'} + \sigma^2 I_{n_{ij}}$. $\eta_k = (\eta_{1k'}', \ldots, \eta_{Jk'}')'$ and $\Lambda_k^{(2)} = diag(\Lambda_{1k'}^{(2)}, \ldots, \Lambda_{Jk'}^{(2)})$.

M-step Estimate the parameters $\theta_{Z^{(1)}}$, $\theta_{Z^{(2)}}$, θ_ξ, θ_ζ and σ^2 by maximizing the expectation of the complete likelihood given X_i.

• Estimate the parameters $\theta_{Z^{(1)}}$ and $\theta_{Z^{(1)}}$ by maximizing $E[l_C(\theta_{Z^{(1)}}; Z^{(1)})|X]$ and $E[l_C(\theta_{Z^{(2)}}; Z^{(2)})|X]$ subject to the constrains $\sum_{k=1}^{C_1} \pi_k^{(1)} = 1$ and $\sum_{k'=1}^{C_2} \pi_{k'}^{(2)} = 1$. The estimates are

$$\hat{\pi}_k^{(1)} = \frac{1}{I}\sum_{i=1}^{I} \hat{z}_{ik}^{(1)} \text{ and } \hat{\pi}_{k'}^{(2)} = \frac{1}{I}\sum_{i=1}^{I} \hat{z}_{ik'}^{(2)}.$$

• Estimate the parameters θ_ξ by maximizing $E[l_C(\theta_\xi; \xi)|Z^{(1)}, X]$ subject to the constrains $\sum_{k=1}^{C_1} \pi_k^{(1)}\mu_k = 0$. We solve this optimization problem using Lagrange multiplier

$$\min_{\mu_k, \lambda_\xi} \sum_{k=1}^{C_1}\sum_{i=1}^{I} \hat{z}_{ik}^{(1)} E[\log|\Lambda_k^{(1)}| + (\boldsymbol{\xi}_i - \mu_k)'\Lambda_k^{(1),-1}(\boldsymbol{\xi}_i - \mu_k)|Z_{ik}^{(1)} = 1, X_i] + \lambda_\mu(\sum_{k=1}^{C_1} \pi_k^{(1)}\mu_k)$$

where λ_μ is the Lagrange multiplier. The location estimates are

$$\hat{\mu}_k = \frac{1}{n_k^{(1)}}(\sum_{i=1}^{I} \hat{z}_{ik}^{(1)}\hat{\boldsymbol{\nu}}_{ik} - \pi_k^{(1)}\Lambda_k^{(1)}\lambda_\mu)$$

where $n_k^{(1)} = \sum_{i=1}^{I} Z_{ik}^{(1)}$ and $\lambda_\mu = (\sum_{k=1}^{C_1} \pi_k^{(1)}\Lambda_k^{(1)})^{-1}\sum_{k=1}^{C_1}\sum_{i=1}^{I} Z_{ik}^{(1)}\xi_{ik}$. The variance components are

$$\begin{aligned}
\hat{\lambda}_k^{(1)} &= \frac{1}{n_k^{(1)}}\sum_{i=1}^{I} \hat{z}_{ik}^{(1)} diag(E[(\boldsymbol{\xi}_i - \mu_k)(\boldsymbol{\xi}_i - \mu_k)'|Z_{ik}^{(1)} = 1, X_i]) \\
&= \frac{1}{n_k^{(1)}}\sum_{i=1}^{I} \hat{z}_{ik}^{(1)} diag(E[\boldsymbol{\xi}_i - \mu_k|Z_{ik}^{(1)} = 1, X_i]E[\boldsymbol{\xi}_i - \mu_k|Z_{ik}^{(1)} = 1, X_i]' \\
&+ Cov[\boldsymbol{\xi}_i - \mu_k|Z_{ik}^{(1)} = 1, X_i]) \\
&= \frac{1}{n_k^{(1)}}\sum_{i=1}^{I} \hat{z}_{ik}^{(1)} diag((\hat{\boldsymbol{\nu}}_{ik} - \mu_k)(\hat{\boldsymbol{\nu}}_{ik} - \mu_k)' + Cov[\boldsymbol{\xi}_i|Z_{ik}^{(1)} = 1, X_i])
\end{aligned}$$

• Similarly, we estimate the parameters θ_ζ by maximizing $E[l_C(\theta_\zeta; \zeta)|Z^{(2)}, X]$ subject to the constrains $\sum_{k'=1}^{C_2} \pi_{k'}^{(2)}\eta_{jk'} = 0$. The Lagrange multiplier problem is as follows

$$\min_{\eta_{jk'}, \lambda_\zeta} \sum_{k'=1}^{C_2}\sum_{i=1}^{I} \hat{z}_{ik'}^{(2)} E[\log|\Lambda_{jk'}^{(2)}| + (\boldsymbol{\zeta}_{ij} - \eta_{jk'})'\Lambda_{jk'}^{(2),-1}(\boldsymbol{\zeta}_{ij} - \eta_{jk'})|Z_{ik'}^{(2)} = 1, X_i] + \lambda_\eta(\sum_{k'=1}^{C_2} \pi_{k'}^{(2)}\eta_{jk'})$$

where λ_η is the Lagrange multiplier. The estimates are

$$\hat{\eta}_{jk'} = \frac{1}{n_{k'}^{(2)}}(\sum_{i=1}^{I} \hat{z}_{ik'}^{(2)}\hat{\boldsymbol{\delta}}_{ij,k'} - \pi_{k'}^{(2)}\Lambda_{jk}^{(2)}\eta)$$

where $n_k^{(2)} = \sum_{i=1}^{I} \hat{z}_{ik'}^{(2)}$ and $\lambda_\eta = (\sum_{k'=1}^{C_2} \pi_{k'}^{(2)} \Lambda_{jk'}^{(2)})^{-1} \sum_{k'=1}^{C_2} \sum_{i=1}^{I} \hat{z}_{ik'}^{(2)} \hat{\boldsymbol{\delta}}_{ij,k'}$. The variance components are

$$\begin{aligned}
&\hat{\lambda}_{k'}^{(2)} = \frac{1}{n_{k'}^{(2)}} \sum_{i=1}^{I} \hat{z}_{ik'}^{(2)} diag(E[(\boldsymbol{\zeta}_i - \eta_{k'})(\boldsymbol{\zeta}_i - \eta_{k'})' | Z_{ik'}^{(2)} = 1, X_i]) \\
&= \frac{1}{n_k^{(2)}} \sum_{i=1}^{I} \hat{z}_{ik}^{(2)} diag(E[\boldsymbol{\zeta}_i - \eta_{k'} | Z_{ik'}^{(2)} = 1, X_i] E[\boldsymbol{\zeta}_i - \eta_{k'} | Z_{ik}^{(2)} = 1, X_i] \\
&+ Cov[\boldsymbol{\zeta}_i - \eta_k | Z_{ik'}^{(2)} = 1, X_i]) \\
&= \frac{1}{n_k^{(2)}} \sum_{i=1}^{I} \hat{z}_{ik}^{(2)} diag((\hat{\boldsymbol{\delta}}_{ik'} - \hat{\eta}_{k'}])(\hat{\boldsymbol{\delta}}_{ik'} - \hat{\eta}_{k'})' + Cov(\boldsymbol{\zeta}_i | Z_{ik'}^{(2)} = 1, X_i)]
\end{aligned}$$

where $Cov[\boldsymbol{\xi}_i | Z_{ik}^{(1)} = 1, X_i]) = \hat{\boldsymbol{\nu}\boldsymbol{\nu}'}_{ik} - \hat{\boldsymbol{\nu}}_{ik} \hat{\boldsymbol{\nu}}_{ik}'$ and $Cov[\boldsymbol{\zeta}_i | Z_{ik'}^{(2)} = 1, X_i]) = \hat{\boldsymbol{\delta}\boldsymbol{\delta}'}_{ik'} - \hat{\boldsymbol{\delta}}_{ik'} \hat{\boldsymbol{\delta}}_{ik'}'$.

• The final step is to estimate the variance of random error σ by maximizing $E[l_C(\sigma^2) | Z^{(1)}, Z^{(2)}, \boldsymbol{\xi}, \boldsymbol{\zeta}, X]$. Denote the sample size $N = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$, then the estimate is

$$\begin{aligned}
&\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{I} \sum_{k=1}^{C_1} \hat{z}_{ik}^{(1)} \sum_{k'=1}^{C_2} \hat{z}_{ik'}^{(2)} E[(X_i - \Phi_i^{(1)} \boldsymbol{\nu}_{ik} - \Phi_i^{(2)} \boldsymbol{\delta}_{ik'})' (X_i - \Phi_i^{(1)} \boldsymbol{\nu}_{ik} - \Phi_i^{(2)} \boldsymbol{\delta}_{ik'}) \\
&| X_i, Z_{ik}^{(1)} = 1, Z_{ik'}^{(2)} = 1] \\
&= \frac{1}{N} \sum_{i=1}^{I} \sum_{k=1}^{C_1} \hat{z}_{ik}^{(1)} \sum_{k'=1}^{C_2} \hat{z}_{ik'}^{(2)} \{(X_i - \Phi_i^{(1)} \hat{\boldsymbol{\nu}}_{i,kk'} - \Phi_i^{(2)} \hat{\boldsymbol{\delta}}_{i,kk'})' (X_i - \Phi_i^{(1)} \hat{\boldsymbol{\nu}}_{i,kk'} - \Phi_i^{(2)} \hat{\boldsymbol{\delta}}_{i,kk'}) \\
&+ \text{trace}(Cov[X_i - \Phi_i^{(1)} \boldsymbol{\nu}_{ik} - \Phi_i^{(2)} \boldsymbol{\delta}_{ik'} | Z_k^{(1)} = 1, Z_{k'}^{(2)} = 1, X_i])\} \\
&= \frac{1}{N} \sum_{i=1}^{I} \sum_{k=1}^{C_1} \hat{z}_{ik}^{(1)} \sum_{k'=1}^{C_2} \hat{z}_{ik'}^{(2)} \{(X_i - \Phi_i^{(1)} \hat{\boldsymbol{\nu}}_{i,kk'} - \Phi_i^{(2)} \hat{\boldsymbol{\delta}}_{i,kk'})' (X_i - \Phi_i^{(1)} \hat{\boldsymbol{\nu}}_{i,kk'} - \Phi_i^{(2)} \hat{\boldsymbol{\delta}}_{i,kk'}) \\
&+ \text{trace}(\Phi_i^{(1)} Cov[\boldsymbol{\xi}_i | Z_k^{(1)} = 1, Z_{k'}^{(2)} = 1, X_i] \Phi_i^{(1)'} + \Phi_i^{(2)} Cov[\boldsymbol{\zeta}_{ik} | Z_k^{(1)} = 1, Z_{k'}^{(2)} = 1, X_i] \Phi_i^{(2)'} \\
&+ 2\Phi_i^{(1)} Cov[\boldsymbol{\xi}_i, \boldsymbol{\zeta}_i | Z_k^{(1)} = 1, Z_{k'}^{(2)} = 1, X_i] \Phi_i^{(2)'})\}
\end{aligned}$$

where $\hat{\boldsymbol{\nu}}_{i,kk'} = E[\xi_i | Z_{ik}^{(1)} = 1, Z_{ik'}^{(1)} = 1, X_i]$, $\hat{\boldsymbol{\delta}}_{i,kk'} = E[\zeta_i | Z_{ik}^{(1)} = 1, Z_{ik'}^{(1)} = 1, X_i]$ and $Cov[\boldsymbol{\xi}_i, \boldsymbol{\zeta}_i | Z_k^{(1)} = 1, Z_{k'}^{(2)} = 1, X_i] = -\Lambda_k^{(1)} \Phi^{(1)'} (\Phi^{(1)} \Lambda_k^{(1)} \Phi^{(1)'} + \Phi^{(2)} \Lambda_{k'}^{(2)} \Phi^{(2)'} + \sigma^2 I_{n_{ij}})^{-1} \Phi^{(2)} \Lambda_{k'}^{(2)}$.

REFERENCES

[1] Archer, G.E.B., Titterington, D.M. (2002), "Parameter estimation for hidden Markov chains", *Journal of Statistical Planning and Inference*, 108, 365.

[2] Bar-Joseph, Z., Gerber, G., Gifford, D.K., Jaakkola, T.S. (2002). "A new approach to analyzing gene expression time series data", *Proceedings of the 6th Annual International Conference on RECOMB*, 39-48.

[3] Besag, J. (1986), "On the statistical analysis of dirty pictures"(with discussion), *Journal of the Royal Statistical Society, B*, 48(3),259.

[4] Blackwell, A.G. and Treuhaft, S. (2008), "Regional Equity and the quest for Full Inclusion". *PolicyLink*

[5] Blekas, K., Nikou, C., Galatsanos, N., Tsekos, N.V.(2007), "Curve Clustering with Spatial Constrains for analysis of Spatiotemporal Data", *19th IEEE International Conference on Tools with Artificial Intelligence*, 529.

[6] Booth, J.G., Casella, G., Hobert, J.P. (2008), "Clustering Using Objective Functions and Stochastic Search", *Journal of the Royal Statistical Society, B* 70(1), 119-140.

[7] Bretherton, C.S., Smith, C., Wallance, J.M. (1992), "An Intercomparison of methods for coupled patterns in climate data", *Journal of Climate*, 5, 54.

[8] Bugli, C., Lambert, P. (2006) "Functional ANOVA with random functional effects: an application to event-related potentials modelling for electroencephalograms analysis", *Statistics in Medicine*, 25, 3718C3739.

[9] Cardot, H. (2007), "Conditional functional principal components analysis", *Scandinavian Journal of Statistics*, 34, 317-335.

[10] Celeux, G., Soromenho, G. (1996), "An entropy critierion for assessing the number of clusters in a mixture model", *Classification Journal*, 13, 195-212.

[11] Clarke, E.D., Speirs, D.C., Heath, M.R., Wood, S.N., Gurney, W.S.C., Holmes, S.L. (2006), "Calibrating remotely sensed chlorophyll-a data by using penalized regression splines", *Applied Statistics*, 55(3), p331.

[12] Chiou, J.M., Li, P.L. (2007), "Functional clustering and identifying substructures of longitudinal data," *Journal of the Royal Statistical Society*, Series B, 69, 679-699.

[13] Diggle, P. J. (1985). "A kernel method for smoothing point process data." *Applied Statistics*s 34(2), 138-147.

[14] Dubin, J.A., Müller, H.(2005), "Dynamical Correlation for Multivariate Longitudinal Data", *Journal of the American Statistical Association*, 100,872.

[15] Fahrmeir, L., Kneib, T., and Lang, S. (2004) , "Penalized structured additive regression for space-time data: A Bayesian perspective", *Statistica Sinica*, 14, 731.

[16] Fraley, C, Raftery, A.E. (1998), "How many clusters? Which clustering method? Answers visa model-based cluster analysis", *Computer Journal*, 41, 561-588

[17] Fraley, C., Raftery, A. E. (2002), "Model-Based Clustering, Discriminant Analysis, and Density Estimation", *Journal of the American Statistical Association*, 97, 611.

[18] Frumkin, H., Frank, L., and Jackson, R. (2004), *Urban Sprawl and Public Health Designing, Planning, and Building for Healthy Communities.* Island Press.

[19] Graves, S. (2003), "Landscapes of Predation, Landscapes of Neglect: A Location Analysis of Payday Lenders and Banks," *The Professional Geographer*, 55(3).

[20] Hastie, T., Tibshirani, R., Eisen, M., Alizadeh, A., Levy, R., Staudt, L., Chan, W., Botstein, D., Brown, P. (2000), "'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns", *Genome Biology*, I(2).

[21] He, G., Muller, H.G., Wang, J.L. (2004), "Methods for Canonical Analysis for Functional Data", *Journal of Statistical Planning and Inference*, 122, 141.

[22] Heckman, N.E., Zamar, R.H. (2000), "Comparing the shapes of regression functions", *Biometrika*, 87(1), 135.

[23] Huang, Y., Zhang, P. (2006), "On The Relationships Between Clustering and Spatial Co-location Pattern Mining", *Proc. of IEEE International Conference on Tools with Artificial Intelligence (ICTAI).*

[24] James, G., Hastie, T., Sugar, C. (2000) "Principal Component Models for Sparse Functional Data", *Biometrika*, 87, 587-602.

[25] James, G.M., Sugar,C.A. (2003), "Clustering for sparsely sampled functional data", *Journal of the American Statistical Association*, 98, 397.

[26] Kaufman, C., Sain, S.R. (2010) "Bayesian Functional ANOVA Modeling Using Gaussian Process Prior Distributions", *Bayesian Analysis*, 5, Number 1, 123-150.

[27] Koehler, A.B., Murphee, E.H. (1988), "A comparison of the Akaike and Schwarz criteria for selecting model order", *Applied Statistics*, 37, 187-195

[28] Kyriakidis, P.C., Journel, A.G. (1999), "Geostatistical Space-time Models: A Review", *Mathematical Geology*, Vol. 31, No. 6, pp 651-684.

[29] Lee, S-I. (2001), "Developing a bivariate spatial association measure: an integration of Pearson's and Morran'a I", *Journal of Geographical Systems*, 3, 369-386.

[30] Matern, B. (1986), "Spatial Variation", 2nd ed. *Lecture Notes in Statistics*, Springer Verlag, New York.

[31] Leroux, B.G.(1992), "Consistent estimation of a mixing distribution", *Annals of Statistics*, 20, 1350-1360.

[32] Li, Y., and Ruppert, D.(2008), "On the asymptotics of penalized splines", *Biometrika*, 95(2), 415.

[33] Larson, T. (2003), "Why There Will Be No Chain Supermarkets In Poor Inner-City Neighborhoods," *California Politics and Policy*, 7(1).

[34] Lee, M. and Rubin, V. (2007), "The Impact of the Built Environment on Community Health: The State of Current Practice and Next Steps for a Growing Movement". *PolicyLink*.

[35] Lovett, A., Haynes, R., Sunnenberg, G., and Gale, S. (2002), "Car Travel Time and Accessibility by Bus to General Practitioner Services: A Study Using Patient Registers and GIS," *Pergamon-Elsevier Science Ltd*, 97-111.

[36] Mandela, M., Betenskyb, R.A.(2008), "Simultaneous confidence intervals based on the percentile bootstrap approach", *Computational Statistics and Data Analysis*, 52, 2158.

[37] Marsh, M. T., and Schilling, D. A. (1994), "Equity Measurement in Facility Location Analysis - a Review and Framework," *European Journal of Operational Research*, 74, 1-17.

[38] Maruca, S. L., Jacquez, G.M. (2002), "Area-based tests for association between spatial patterns", *Journal of Geographical Systems*, 4, 69.

[39] Moore, L. V., and Diez Roux, A V. (2006), "Associations of Neighborhood Characteristics With the Location and Type of Food Stores," *American Journal of Public Health*, 96(2), 325-331.

[40] Morland, K., Wing, S., Diez Roux, A., Poole, C. (2002), "Neighborhood Characteristics Associated with the Location of Food Stores and Food Service Places," *American Journal of Preventive Medicine*, 22(1).

[41] PolicyLink and UC Berkeley School of Public Health (2008), "Promoting Healthy Public Policy through Community-Based Participatory Research: Ten Case Studies," PolicyLink report.

[42] Powell, J. A. and Graham, K. M. (2002). "Urban fragmentation as a barrier to equal opportunity". In D. M. Piche, W. L. Taylor, R. A. Reed (Eds.), Rights at risk: Equality in an age of terrorism. Report of the Citizens Commission on Civil Rights, Washington, DC.

[43] Ramsay, J.O., Silverman, B.W. (1997,2005), *Functional Data Analysis*, Springer, New York.

[44] Ramsay, J.O., Silverman, B.W (2002) *Applied Functional Data Analysis*, Springer, New York.

[45] Rand, W.M. (1971),"Objective Criteria for the Evaluation of Clusterings Methods", *J. of American Statistical Association*, 66, 846-850.

[46] Rice, J.A.(2004),"Functional and Longitudinal Data Analysis: Perspectives on Smoothing"*Statistica Sinica*,14, 631.

[47] Ruppert,D.(2002), "Selecting the number of knots for penalized splines", Journal of Computational and Graphical Statistics , 11, 735-757.

[48] Ruppert, D., Wand, M. P. & Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press.

[49] Salim, A., Pawitan, Y., Bond, K.(2005), "Modelling association between two irregularly observed spatiotemporal processes by using maximum covariance analysis", *Applied Statistics*, 54(3), p555.

[50] Serban, N. (2008), "Clustering in the Presence of Heteroscedastic Errors", *Journal of Nonparametric Statistics*, 20, 7 , 553 - 571.

[51] N. Serban (2009), "Clustering Confidence Sets" , *Journal of Statistical Planning and Inference*, 139,109 -124.

[52] N. Serban, L. Wasserman (2005), "CATS: Cluster Analysis by Transformation and Smoothing", *J. of the American Statistical Association*, 100, 471-481.

[53] Shi, J.Q., Wang, B. (2008), "Curve Prediction and Clustering with Mixtures of Gaussian Process Functional Regression Model", *Statistical Computing*, 18, 267.

[54] Silverman, B.W.(1985), "Some aspects of the spline smoothing approach to nonparametric regression curve fitting", *Journal of the Royal Statistical Society, B*, 47(1), 1.

[55] Small, M.L. and McDermott, M. (2006), "The Presence of Organizational Resources in Poor Urban Neighborhoods: An Analysis of Average and Contextual Effects," *Social Forces*; 84(3),1697.

[56] Soromenho, G. (1933), "Comparing approaches for testing the number of components in a finite mixture model", *Computational Statistics*, 9, 65-78.

[57] Storch, H., Zwiers, F.W.(2002), "Statistical Analysis in Climate Research", Cambridge University Press.

[58] Sugar, C., James, G. (2003), "Finding the Number of Clusters in a Data Set: An Information Theoretic Approach", *Journal of the American Statistical Association*, 98, (2003), 750-763.

[59] Talen, E. (1996), "Visualizing Fairness: Equity Maps for Planners." *Journal of the American Planning Association*, 64(1), 22-38.

[60] Talen, E. (2001), "School, Community, and Spatial Equity: An Empirical Investigation of Access to Elementary Schools in West Virginia," *Annals of the Association of American Geographers*. 91(3), 465-486.

[61] Talen, E. and Anselin, L. (1998), "Assessing spatial equity: an evaluation of measures of accessibility to public playgrounds," *Environment and Planning, A*, 30, 595-613.

[62] Tibshirani, R., Walther, G., Hastie, T. (2001), "Estimating the number of clusters in a dataset via the gap statistic", *Journal of the Royal Statististical Society, B* 63, 411–423.

[63] Vaida, F., Blanchard, S.(2005),"Conditional Akaike information for mixed-effects models", *Biometrika*, 92(2), 351-370.

[64] Wahba G., (1990), "Spline Models for Observational Data", SIAM, Philadelphia. CBMS-NSF Regional Conference SEries in Applied Mathematics, Vol. 59.

[65] Wakefield, J., Zhou, C., Self, S. (2002), "Modelling Gene Expression Data over Time: Curve Clustering with Informative Prior Distributions", *Bayesian Statistics 7, Proceedings of the Seventh Valencia International Meeting, 2003.*

[66] Wang. Y., Guo, W., Brown, M.B. (2000), "Spline Smoothing for Bivariate Data with Applications to Association Between Hormones", *Statistica Sinic*, 10, p377.

[67] Wikle, C.K., Cressie, N.(1999),"A dimension-reduced approach to space-time Kalman filtering ", *Biometrika*, 86(4), 815.

[68] Wood, S., (2003), "Thin Plate Regression Splines", *Journal of the Royal Statistical Society, B*, 65(1), 95-114.

[69] Wood, S.N. (2006), "Generalized Additive Models, An Introduction with R", Chapman& Hall.

[70] Yao, F., Müller, H.G., Wang, J.L. (2005), "Functional data analysis for sparse longitudinal data", *Journal of the American Statistical Association*, 100, 577-590.

[71] Zhou, L., Huang, J. Z., Carroll, R. (2008), "Joint modelling of paired sparse functional data using principal components", *Biometrika*, 95(3), p601.