

# Large Scale Clustering of Dependent Curves

Huijing Jiang, Nicoleta Serban  
Industrial Systems and Engineering School  
Georgia Institute of Technology

Abstract: In this paper, we introduce a model-based method for clustering multiple curves or functionals under spatial dependence specified up to a set of unknown parameters. The functionals are decomposed using a semi-parametric model where the fixed effects account for the large-scale clustering association and the random effects for the small scale spatial-dependence variability. The clustering model assumes the clustering membership as a realization from a Markov random field. Within our estimation framework, the emphasis is on a large number of functionals/spatial units with sparsely sampled time points. To overcome the computational cost resulting from large dependence matrix operations, the estimation algorithm includes a two-stage approximation: low-ranked kernel-based decomposition of the dependence matrix and Incomplete Cholesky Factorization of the kernel matrix. We assess the performance of our clustering approach within a simulation study. The simulation results show enhanced clustering estimation accuracy of our method compared with other existing model-based clustering methods under a series of settings: small number of time points, low signal-to-noise ratio and different spatial dependence structures. Many case studies will fall within our clustering framework, but we focus on obtaining fine-grid spatial clusters for demographics trends including ethnicity and income for five southern states of US over the past 11 years.

# 1 Introduction

The emergence of time-dependent or functional data in many research areas has been followed by exploratory and inferential methodology coined Functional Data Analysis (FDA) - a central book in this field is by Ramsay and Silverman, 1997. Due to the increasing number of applications with very large number of observed functionals, exploratory tools such as clustering and classification play an important role in FDA. One common approach to clustering functional data is to cluster functionals observed over the same discrete fine grid of points with or without accounting for the functional (temporal) dependence within each curve; this approach is called regularization clustering. A second approach accounts for the curve functionality by transforming the functionals using an orthonormal basis of functions and clustering the transform (estimated) coefficients; this approach is called filtering clustering. Both regularization and filtering clustering methods further divide into hard and soft clustering. Examples of hard clustering methods are Hastie et al, 2000, Bar-Joseph, Gerber, Gifford and Jaakkola, 2002; Serban and Wasserman, 2005; and Serban, 2007. Some examples of model-based clustering are James and Sugar, 2003 for a filtering method; Fraley and Raftery, 2002; and Wakefield, Zhou and Self, 2002 for regularization methods.

Even though the field of clustering functional data has already provided a series of competitive approaches, they are generally limited to the assumption of independence between functionals. An extension from the independence assumption to within-cluster dependence is provided by Booth, Casella & Hobert, 2007 which is a regularization method based on a Bayesian model formulation. Even though the independence assumption is common practice, in many case studies, it is rather restrictive. Accounting for dependence between functionals will not only enhance the estimation accuracy of the patterns and cluster memberships by borrowing information across dependent curves, but will also allow estimation of the underlying dependence, which may be used in the inference analysis and cluster interpretation. Importantly, the accuracy of the clustering membership greatly improves under low signal-to-noise ratio and small number of sampled time points.

Without digressing much from the overall functional approach, we highlight here a

second thrust of research that relates to the clustering spatial dependent functionals: spatio-temporal data analysis. The research on spatio-temporal estimation is rather mature with both Bayesian and frequentist approaches. One related method is extending spatial smoothing (Nychka, 2000) to incorporate temporal dependence (Luo et al., 1998; Kamman and Wand, 2003; Clark et al., 2006). Even though the literature on spatial-temporal data analysis is very rich, to the best of our knowledge, there have been few developed time-dependent clustering methods incorporating spatial dependence information between variables. Existing work on clustering spatiotemporal data has been mostly studied by computer scientists (Kakka, 2004 and the references herein), most often offering ad-hoc rather than model-based solutions to clustering under nuisance spatial dependence.

Motivated by the need of overcoming the restrictive independence assumption between functionals in many real life applications, in this paper, we introduce a novel model-based method for clustering multiple curves or functionals under a dependence structure specified up to a set of unknown parameters. The general method allows for any type of correlation structure between functionals, but to keep our presentation focused, we discuss the clustering approach under spatial dependence. The dependence does not extend only to within-cluster functionals as provided by Booth, Casella & Hobert, 2007 but to both within- and between-cluster functionals as the spatial dependence extends beyond the cluster membership. In the followings, we refer to our modeling procedure as *Functional Spatial Clustering Model (FSCM)*.

In multiple curve/functional studies, a series of functionals are observed with error:

$$Y_{si} = f_s(t_{is}) + \sigma_s^2 \epsilon_{is}, \quad s = 1, \dots, S, \quad (1)$$

where  $t_s = (t_{1,s}, \dots, t_{T,s})$  are the observed time points for functional  $s$ . The overarching objective of this paper is to estimate a clustering of the functionals  $f_s(t)$  for  $s = 1, \dots, S$  where a cluster consists of similar time-varying functional patterns. In our clustering framework, the complete data are  $(Y_s, Z_s)$ ,  $s = 1, \dots, S$  where  $Y_s$ 's are observed functionals as provided in the functional model (1) and  $Z_s$ 's are missing latent variables defining the cluster membership.

The primary contribution of this paper is to allow  $(Y_s, Z_s)$ ,  $s = 1, \dots, S$  be spatially correlated; that is, we assume that both the latent variables  $Z_s$ ,  $s = 1, \dots, S$  and the conditional observations  $Y_s|Z_s$ ,  $s = 1, \dots, S$  are spatially dependent, which will further provide the spatial dependence for the joint data  $(Y_s, Z_s)$ ,  $s = 1, \dots, S$ . To account for the correlation in the latent variables  $Z_s$ ,  $s = 1, \dots, S$ , we consider a Hidden Markov model (HMM) with the spatial dependence of  $Z$  modelled as a realization of a Markov random field. In the literature of HMM's, the random variables  $Y_1, \dots, Y_S$  are assumed conditional independent given a cluster membership  $Z$ . Generally, under spatial dependence, this conditional independence assumption does not hold since nearby locations will be more highly associated than locations that are far apart regardless of the clustering membership. Consequently, in our model formulation, the assumption that  $Y$  is conditional independent on  $Z$  is relaxed to spatial dependence. In Section 2, we expand on this model formulation and the corresponding assumptions followed by the general clustering procedure.

The secondary contribution of this paper is an estimation procedure which allows for large number of functionals/spatial units ( $S$  large). Since the estimation algorithm involves operations with a large dependence matrix ( $ST \times ST$  where  $ST \gg 50,000$ ), we employ a two-stage approximation: at the first stage, we reduce the size of the full-ranked dependence matrix using a low-rank approximation; at the second stage (for inversion operation only), we apply advanced matrix factorization techniques to reduce the computational cost of its inversion. These two approximations are discussed in Section 3.2, which are implemented within the fitting algorithm presented in Section 3.1.

We illustrate our methodology with a series of synthetic examples (Section 4) and an empirical example using demographic variables (Section 5). Within the simulation study, we investigate a range of model scenarios with different levels of signal-to-noise ratio and spatial correlation structures to provide insights into the primary settings under which the estimation accuracy of the clustering membership improves when allowing for spatial dependence. Within the empirical study, we apply our clustering method to the US demographics data where each curve is the time-dependent behav-

ior of a demographical variable in a specific spatial area; therefore, there is an intrinsic spatial dependence between the demographics curves. We investigate the time dependent changes and spatial clustering for income and four population variables, the proportion of Asian, Black, Hispanic and White population. Nonetheless, our clustering algorithm may apply to other challenging examples arising in 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.

## 2 The Model

### 2.1 Functional-Spatial Model

In our clustering model, we estimate the functionals  $f_s(t)$  in model (1) using a semi-parametric model extending on the formulations introduced by Ruppert, Wand and Carrol 2003,

$$f_s(t) = \beta_{s0} + \beta_{s1}t + \dots + \beta_{s,p-1}t^{p-1} + \sum_{i=1}^T \gamma_{s,i} K_{temp}(|t - t_i|) \quad (2)$$

where  $t_i$ ,  $i = 1, \dots, T$  are the observed time points, and  $K_{temp}$  is a temporal smoothing kernel. Under this decomposition,  $\beta_s = (\beta_{s0}, \dots, \beta_{s,p-1})$  is the vector of space-varying fixed effects and  $\gamma_s = (\gamma_{s1}, \dots, \gamma_{sT})$  is the vector of space-varying random effects. In the semiparametric literature,  $\gamma_s$ ,  $s = 1, \dots, S$  are assumed independent. To model the spatial dependence in  $\gamma_s$ ,  $s = 1, \dots, S$ , we further decompose them according to

$$\gamma_{s,i} = \sum_{j=1}^S \gamma_{i,j} K_{sp}(\|s - s_j\|) \quad (3)$$

where  $s_j$ ,  $j = 1, \dots, S$  are the observed spatial locations and  $K_{sp}$  is a spatial smoothing kernel. With this second decomposition level, the random effects are  $\gamma = \{\gamma_{i,j}\}_{i=1, \dots, T, j=1, \dots, S}$ .

Because we are interested in clustering under densely sampled spatial domains, we will reduce the set of random effects using a low-rank kernel approximation as we will discuss in Section 3.2.

Under this decomposition, the model in (1) extends to the general form

$$Y_s(t) = \beta_s X_s(t) + \gamma_s K(s, t) + \varepsilon_s(t) \quad (4)$$

where  $X_s(t)$  is a vector of  $p$  covariates, which in our model formulation, are polynomial functions  $X_s(t) = (1, t, \dots, t^{p-1})$  and  $K(s, t)$  is a spatio-temporal smoother, in our model,  $K(s, t) = K_{sp}(s)K_{temp}(t)$ . In model (2), the spatio-temporal smoother is separable as discussed in Section 2.3. We assume that the measurement errors  $\varepsilon_s(t)$  are independent and identically normal distributed with mean zero.

Since we observe  $Y_s(t)$  at  $T$  discrete time points and  $S$  spatial units/areas, the functional mixed model in (2) and (4) can be formalized as a linear mixed model

$$Y = X\beta + B\gamma + \varepsilon, \quad \text{cov} \begin{bmatrix} u \\ \varepsilon \end{bmatrix} = \begin{bmatrix} (B^{-1/2})\Gamma(B^{-1/2})^T & 0 \\ 0 & R \end{bmatrix} \quad (5)$$

where  $Y = (Y_{11}, \dots, Y_{1T}, \dots, Y_{s1}, \dots, Y_{sT})'$  is the vector of length  $ST$  containing all observations,  $X = 1_S \otimes X_T$  is the covariate matrix with  $X_T = \{(1 \ t_i \dots t_i^{p-1})\}_{i=1, \dots, T}$  and  $B = (B_1, \dots, B_S)'$  is the basis matrix function of spatiotemporal smoother  $K = K_{sp} \otimes K_{temp}$ , which will be discussed in detail in Section 2.3. The basis matrix element  $B_s$  corresponds to the spatiotemporal smoother evaluated at  $s$ ,  $K_{sp}(s) \otimes K_{temp}$  for  $s = 1, \dots, S$ .

In the estimation method, we assume that the random effects  $\gamma = \{\gamma_{ij}\}_{i=1, \dots, T, j=1, \dots, S}$  are normally distributed with mean zero and covariance matrix  $\Gamma = V \otimes C$  with  $V = \text{Diag}(\sigma_{s,1}^2, \dots, \sigma_{s,S}^2)$  and  $C = \text{Diag}(\sigma_{t,1}^2, \dots, \sigma_{t,T}^2)$ ; that is,  $\mathbb{V}(\gamma_{ij}) = \sigma_{ti}^2 \sigma_{sj}^2$  but otherwise the random effects are independent. Allowing for non-constant variance across random effects implies that the variance of  $Y_s(t)$  varies with space and time.

## 2.2 Functional-Spatial Clustering Model

For our clustering procedure, we extend the model in equation (5) by decomposing the mean function into global trend and cluster trend, both of which are polynomial functions of time. The overall fixed effect is  $\beta_0$  and the cluster fixed effect is  $\beta_Z$  where  $Z = (Z_1, \dots, Z_S)$  is a hidden variable specifying the clustering partition. Given  $Z$ , the model becomes

$$Y = X(\beta_0 + \beta_Z) + B\gamma + \varepsilon.$$

Following the current literature on HMRF (Hidden Markov Random Field) modelling, we assume that the clustering configuration  $Z_1, \dots, Z_S$  is a realization of a locally dependent Markov random field with a prior distribution – Gibbs distribution, which depends on a parameter  $\psi$ ,

$$p(z_{s_i}) = \frac{1}{Z_{s_i}(\psi)} \exp(U_{s_i}(\psi)),$$

where  $U_{s_i}(\psi) = \sum_{s_j \in \partial s_i} \psi I(z_{s_j} = z_{s_i})$  is called the energy function. Large values of  $U_{s_i}(\psi)$  correspond to spatial patterns with large spatial connected sub-areas belonging to the same cluster. Small values of  $U_{s_i}(\psi)$  correspond to patterns that do not display any sort of spatial organization.  $Z_{s_i}(\psi)$  is a normalizing constant called the partition function and  $\partial s_i$  is a prescribed neighborhood of the  $s_i$ th spatial unit;  $\psi$  is called the interaction parameter. The value  $\psi = 0$  corresponds to the uniform distribution on the configuration space.

One difficulty in this formulation is that the normalization constant depends on the scale parameter  $\psi$ ; there is not a close-form estimation of  $\psi$  using the likelihood function approach due to the spatial dependence among  $Z_1, \dots, Z_S$ . In the HMRF literature (Besag, 1986, Archer and Titterton, 2002), this difficulty is overcome by assuming local dependence on each spatial unit  $s_i$ , i.e.,  $s_i$  only depends on its neighbors  $\partial_i$ . Thus the joint distribution of  $Z_1, \dots, Z_S$  can be approximated by a

pseudo-likelihood function,

$$f(z_1, \dots, z_S) \approx \prod_{s_i=1}^S f(z_{s_i} | z_{\partial s_i}; \psi)$$

In addition to the difficulty of estimating  $\psi$ , computational challenges arise in recovering the cluster membership  $Z_1, \dots, Z_S$  because of the spatial dependence between  $Y_s | Z$ . To our best knowledge, in all relevant work,  $Y_s | Z$  are assumed conditional 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). In the fitting algorithm discussed in Section 3.1, the conditional independence assumption is relaxed.

## 2.3 Spatio-temporal Dependence

In the model definition in (4), we use a spatiotemporal kernel decomposition of the spatial and temporal dependence. A spatiotemporal kernel is defined as a function of the distances between the points in both time and space domains,  $K_{st}((s, t), (s', t')) = K_{st}(s - s', t - t')$ . For computational feasibility under densely sampled spatial domain, we impose two constraints on the spatiotemporal kernel.

*Assumption 1:* The space-time dependence is multiplicative separable, i.e. the spatial dependence dies out across time. This translates into  $K_{st} = K_{sp} \otimes K_{temp}$  or  $K_{st}((s, t), (s', t')) = K_{sp}(s - s')K_{temp}(t - t')$  where  $K_{sp}$  is a spatial kernel and  $K_{temp}$  is a temporal kernel. Methods for constructing valid non-separable covariance functions with spatiotemporal interactions include the monotone function approach of Gneiting, 2002 and the spectral method of Cressie and Huang, 1999 and Chen, Fuentes and Davis, 2006. However, non-separability involves complex interaction between space and time, and therefore, will require expensive computation.

*Assumption 2:* The spatial dependence is isotropic, i.e. the covariance function depends only on the absolute distance between points. This translates into  $K_{st}((s, t), (s', t')) = K_{st}(\|s - s'\|, \|t - t'\|)$ , which together with the separability as-

sumption, it results in

$$K_{st}((s, t), (s', t')) = K_{sp}(\|s - s'\|)K_{temp}(\|t - t'\|).$$

For modeling the isotropic temporal dependence, we use a temporal smoother kernel derived from the thin plates family, which for one-dimension is expressed as

$$K_{temp}(\|t_i - t_j\|) = \|t_i - t_j\|^{2p-1}.$$

For modeling the isotropic spatial dependence, we use the Matérn class of functions (see Matérn, 1986; Stein 1999) since this class provides correlation surfaces with a wide range of smoothness levels controlled through a parameter  $\nu$ . The range parameter  $\phi$  defines the extent of spatial dependence.

$$K_{sp}(\|s_i - s_j\|) = \sigma_s^2 \frac{1}{\Gamma(\nu)} \left( \frac{\phi \|s_i - s_j\|}{2} \right)^\nu 2B_\nu(\phi \|s_i - s_j\|)$$

where  $B_\nu$  is the modified Bessel function of the second kind of order  $\nu > 0$ . Both  $\phi$  and  $\nu$  are unknown hyperparameters. The classical approach of estimating these two parameters is empirical variogram (Cressie, 1993). Stein, 1999 recommends likelihood-based estimation - Matérn family kriging. On the other hand, Zhang, 2004 proved that for  $d \leq 3$  and under in-fill asymptotics, the parameters  $\phi$  and  $\nu$  cannot be consistently estimated if considered jointly. Therefore, at least one of the parameters need to be fixed. Kammann and Wand, 2003 suggest choosing  $\phi$  equal to the maximum distance between the points in the space of interest to ensure scale invariance and numerical stability and  $\nu = \frac{3}{2}$  because it is the simplest member that still results in differentiable surface estimates. In the examples investigated in this paper, we follow the recommendation of Kammann and Wand, 2003 using the range parameter  $\phi$  equal to the maximum distance between the spatial knots and holding  $\nu$  fixed.

The isotropy assumption may be relaxed to geometric anisotropy, for example, which gives rise to elliptical contours which can be corrected by a linear transformation of the initial coordinate system. An extension to the geometric anisotropy is

zonal anisotropy which can also be corrected by applying a sum of linear transformations. One other less restrictive class of models for anisotropic kernels is based on the deformation approach; some relevant papers are by Sampson and Guttorp, in particular Sampson and Guttorp 1992.

### 3 Computational Approach

In the functional mixed model (5), the parameters to be estimated are the fixed effects  $\beta$ , the random effects  $\gamma$ , the space- and time-varying variances of the random effects,  $\sigma_{i,s}^2$  and  $\sigma_{j,t}^2$  and the variance of the measurement error  $\sigma_\epsilon^2$ . The fitting algorithm of our model is an EM-type iterative procedure and it is described in the diagram below.

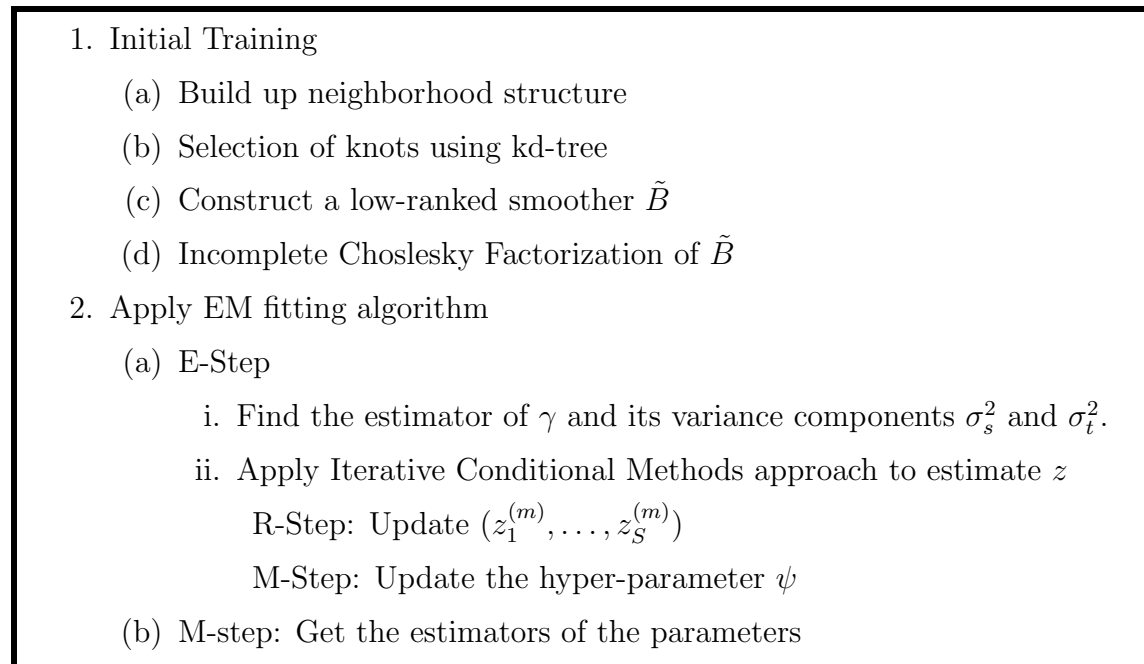


Table 1. A fast FSC algorithm of two-stage approximation.

### 3.1 Fitting Algorithm

A modified EM algorithm is applied to estimate the model parameters by maximizing the expectation of the complete likelihood,

$$f(Y, \gamma, Z; \beta_0, \beta_1, \dots, \beta_C, \sigma_\varepsilon^2, \sigma_s^2, \sigma_t^2, \psi) = f(Y|\gamma, Z; \beta_0, \beta_1, \dots, \beta_C, \sigma_\varepsilon^2) \cdot f(\gamma; \sigma_s^2, \sigma_t^2) \cdot f(z_1, \dots, z_S; \psi) \quad (6)$$

- In the *E-step*, the cluster membership  $(Z_1, \dots, Z_S)$  and the random effects  $\gamma$  are estimated. To identify the cluster membership, we adopted the Iterative Conditional Models(ICM) approach commonly implemented for estimating the hidden variables in HMRF models (Besag, 1986). ICM is a two-step (Restoration-step and Maximization-step) iterative procedures. At the  $m$ th iteration step,

1. *R-step*: Create a restoration  $z_{s_i}^{(m)}$ , i.e. find an approximate MAP estimator

$$p(z_{s_i}^m | z_{\partial s_i}^{m-1}, y_{s_i}; \psi^{m-1}, \theta^{m-1}) \propto f(y_{s_i} | z_{s_i}^m; \theta^{m-1}) p(z_{s_i}^m | z_{\partial s_i}^{m-1}; \psi^{m-1})$$

where  $\theta$  are all the parameters associated with  $Y$ .

2. *M-Step*: Find an estimator of  $\psi$  by maximizing the pseudo-likelihood

$$f(z_1, \dots, z_S; \psi) \approx \prod_{s_i=1}^S f(z_{s_i} | z_{\partial s_i}; \psi).$$

Since  $Y_s|Z$  are assumed conditional dependent, we derive a modified R-step using the facts that  $y_s|z_s, \gamma \sim N(X_s\beta_s + B_s\gamma, \sigma_\varepsilon^2 I_T)$  for  $s = 1, \dots, S$  are conditionally independent as well as  $Z$  and  $\gamma$  are independent (see the proof in the Appendix).

*Modified R-step*: Create a restoration  $z_{s_i}^{(m)}$ , i.e. find an approximate MAP estimator

$$p(z_{s_i}^m | z_{\partial s_i}^{m-1}, y_{s_i}; \psi^{m-1}, \theta^{m-1}) \propto f(y_{s_i} | z_{s_i}^m, \gamma; \theta^{m-1}) p(z_{s_i}^m | z_{\partial s_i}^{m-1}; \psi^{m-1})$$

The random effect  $\tilde{\gamma}$  is provided by  $\tilde{\gamma}|Y, Z$  distributed as

$$N\left(\left(\sigma_\varepsilon^2 \Gamma^{-1} + \sum_{s=1}^S \tilde{B}'_s \tilde{B}_s\right)^{-1} \sum_{s=1}^S \tilde{B}'_s [Y_s - X_s(\beta_0 + \beta_z)], \left(\Gamma^{-1} + \sum_{s=1}^S \tilde{B}'_s \tilde{B}_s / \sigma_\varepsilon^2\right)^{-1}\right).$$

The estimation of  $\tilde{\gamma}$  is equivalent to solving a so-called mixed-model equation. The resulting estimator is the Best Linear Unbiased Predictor (BLUP) and is derived from

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}B \\ B'R^{-1}X & B'R^{-1}B + \Gamma^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} X'R^{-1}Y \\ B'R^{-1}Y \end{bmatrix}.$$

• In the *M-step*, the parameters  $\beta_0, \beta_1, \dots, \beta_C, \sigma_\varepsilon^2$  are estimated by maximizing  $f(Y, \gamma|Z; \beta_0, \beta_1, \dots, \beta_C, \sigma_\varepsilon^2)$ . The constraint on the cluster fixed effects ( $\beta_1 + \dots + \beta_C = 0$ ) ensures identifiability of the model parameters. The estimates are

$$\hat{\beta}_0 = \left( \sum_{s=1}^S X'_s X_s \right)^{-1} \sum_{s=1}^S X'_s (Y_s - X_s \beta_{z_s} - B_s \hat{\gamma}),$$

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{ST} \sum_{s=1}^S [ \|Y_s - X_s(\beta_0 + \beta_{z_s}) - B_s \hat{\gamma}\|^2 ]$$

$\hat{\beta}_1, \dots, \hat{\beta}_C$  are estimated by solving a linear equation system,

$$\begin{cases} \sum_{s=1}^S z_{s2} X_s X'_s \beta_2 - \sum_{s=1}^S z_{s1} X_s X'_s \beta_1 = \sum_{s=1}^S (z_{s2} - z_{s1}) X'_s (Y_s - X_s \beta_0 - B_s \hat{\gamma}) \\ \dots \\ \sum_{s=1}^S z_{sC} X_s X'_s \beta_C - \sum_{s=1}^S z_{s1} X_s X'_s \beta_1 = \sum_{s=1}^S (z_{sC} - z_{s1}) X'_s (Y_s - X_s \beta_0 - B_s \hat{\gamma}) \\ \beta_1 + \dots + \beta_C = 0 \end{cases}$$

The variance components of the random effect  $\gamma$  is estimated by

$$\hat{\sigma}_{s_k}^2 = E\left[\frac{1}{K} \hat{\gamma}' (J_K(k) \otimes C^{-1}) \hat{\gamma}\right] \text{ and } \hat{\sigma}_{t_{k'}}^2 = E\left[\frac{1}{K'} \hat{\gamma}' (V^{-1} \otimes J_{K'}(k')) \hat{\gamma}\right]$$

where  $J_n(i)$  denotes a  $n \times n$  matrix with only  $i$ th diagonal element equal to 1 and 0 otherwise. We can calculate them by applying simple matrix algebra  $E(\hat{\gamma}A\hat{\gamma}) = \text{tr}(A\hat{\gamma}) + \text{Acov}(\hat{\gamma})$ , where  $A$  is  $(J_K(k) \otimes C^{-1})$  or  $(V^{-1} \otimes J_{K'}(k'))$  for  $\hat{\sigma}_{s_k}^2$  and  $\hat{\sigma}_{t_{k'}}^2$  respectively.

### 3.2 Kernel Matrix Approximations

The computational burden in the estimation procedure arises from operations with a large kernel matrix  $B$  used to model the spatial-temporal dependence. Particularly, each iterative step requires the inversion involving  $B'B$  which takes a large amount of CPU time and memory storage. Therefore, we propose a fast algorithm of two-stage approximation for  $B$ . The first-stage approximation reduces the size of  $B$  from  $ST$  to  $\kappa_K \kappa_{K'}$  with  $\kappa_K \ll S$  and  $\kappa_{K'} \leq T$ . The second-stage approximation aims at reducing the computational cost of the inversion of  $\tilde{B}'\tilde{B}$  from  $O(\kappa_K^3 \kappa_{K'}^3)$  to  $O(\kappa_K \kappa_{K'} m^2)$  with  $m \ll \kappa_K \kappa_{K'}$ .

**First Approximation: Reduced-rank Smoothers.** We use a low-ranked approximation as defined by Hastie, 1996. Let  $\kappa_1, \dots, \kappa_K$  be a set of spatial knots and  $\kappa_1, \dots, \kappa_{K'}$  be a set of temporal knots; the low-ranked smoother is

$$\tilde{B} = (K_{sp} \Omega_V^{-1/2}) \otimes (K_{temp} \Omega_C^{-1/2})$$

with  $K_{sp} = \{K_{sp}(\|s_i - \kappa_{k_j}\|)\}_{i=1, \dots, S, j=1, \dots, K}$ ,  $\Omega_V = \{K_{sp}(\|\kappa_{k_i} - \kappa_{k_j}\|)\}_{i=1, \dots, K, j=1, \dots, K}$

and  $K_{temp} = \{K_{temp}(\|t_i - \kappa_{k'_j}\|)\}_{i=1, \dots, T, j=1, \dots, K'}$ ,  $\Omega_C = \{K_{temp}(\|\kappa_{k'_i} - \kappa_{k'_j}\|)\}_{i=1, \dots, K', j=1, \dots, K'}$ .

After transforming  $\tilde{B} = B\Omega$  and  $\tilde{\gamma} = \Omega^{-1}\gamma$  with  $\Omega = \Omega_V \otimes \Omega_C$ , the model can be written as

$$Y \sim X(\beta_0 + \beta_Z) + \tilde{B}\tilde{\gamma} + \varepsilon, \quad \text{cov} \begin{bmatrix} u \\ \varepsilon \end{bmatrix} = \begin{bmatrix} \Gamma & 0 \\ 0 & R \end{bmatrix}.$$

One common approach to spatial knots selection is using the space-filling algorithm (Nychka et al. 1998), which is based on minimax design (Johnson et al. 1990).

However, space-filling has several drawbacks: it is sensitive to the initial configuration of points; the quality of the solution depends on the density of the points on the region; for dense spatial space ( $S \gg 5000$ ), optimization can be computationally prohibitive. Consequently, we investigate a new knots selection technique by constructing a kd-tree (see Moore, 1991) which is a space-partitioning data structure for storing a finite set of points from a k-dimensional space and is useful to speed up the nearest neighbors searches. It has two advantages: it is computationally efficient as compared to minimax design.

A kd-tree is constructed by starting with the full space - the parent node of all subspaces; a first step is to split the space into two subspaces according to the splitting hyperplane of the nodes. At each tree level consisting of a series of subspaces or nodes, each node is further split into two other sub-spaces if the number of spatial units in the node is larger than a pre-specified threshold. The splitting is terminated when the count of points in each node is smaller than this pre-specified threshold, and these lowest tree-level nodes are called leave nodes. The centers of the leave nodes become the selected knots. The algorithm is deterministic once we decide the position of the splitting hyperplane.

Figures 1 in the supplemental materials shows the location of knots selected by space-filling algorithm and splitting kd-tree applied to the irregular grid of 3141 counties throughout the US. They result in similar sets of knots; kd-tree is much more stable and computationally efficient. Space-filling takes several hours while splitting a kd-tree only needs a few seconds. Moreover, the performance of space-filling depends on the starting points, which may be selected using multiple optimization but at the price of a higher computational cost.

**Second Approximation: Incomplete Cholesky Factorization.** In E-step, the calculation of  $\hat{\gamma}$  and its variance components requires the inversion of  $\sigma_\epsilon^2 \Gamma^{-1} + \sum_{s=1}^S \tilde{B}'_s \tilde{B}_s$  where  $\tilde{B}'_s \tilde{B}_s$  has very large dimensionality assuming densely observed space domain. Therefore, we use matrix factorization on  $\tilde{B}$  to reduce the computational cost.

The use of matrix factorization has been recently discussed to speed up algorithms

involving operations on large kernel matrices (also called kernel machines), such as eigen-decomposition in Kernel PCA or the covariance matrix inversion in Gaussian processes. All the above operations are of order  $O(n^3)$ , where  $n$  is the dimensionality of the matrix to be factorized. Wu et al., 2005 proposed a Kronecker factorization which is able to reduce the computation cost to  $O(2n^2)$ . This factorization was originally developed by Loan and Pitsianis in 1993. They proposed two algorithms; one algorithm makes use of SVD and the second is based on a separable least squares algorithm. In more recent work, Wu et al., 2006 suggested using Incomplete Cholesky Factorization (ICF) for which the computational cost reduces to  $O(m^2n + \frac{2}{3}m^3)$ , where  $m \ll n$ . In our case,  $Q = \sum_{s=1}^S \tilde{B}'_s \tilde{B}_s$  has size  $n = K \times K'$ . Using ICF, the matrix  $Q$  is approximated as  $Q \approx H_m H_m^T$  where  $H_m \in R^{n \times m}$  with  $m \ll n$  and  $H_m$  is ‘close’ to  $Q$ ’s exact Cholesky factor. The algorithm follows Bash and Jordan 2002. The approximation accuracy is measured by  $\frac{\|Q - H_m H_m^T\|}{\|Q - H_1 H_1^T\|} \leq \eta$  where  $\eta$  is a precision parameter and  $H_1$  is the approximate ICF factor with  $m = 1$ . The total computational cost of ICF is  $O(m^2n)$ .

Once  $H$  has been calculated, Sherman-Morrison-Woodbury formula can be applied to calculate the inverse of  $Q + \Gamma$ . That is

$$(Q + \Gamma)^{-1} \approx (HH^T + \Gamma)^{-1} = \Gamma^{-1} - \Gamma^{-1}H(I + H^T\Gamma^{-1}H)^{-1}H^T\Gamma^{-1}$$

where  $E = I + H^T\Gamma^{-1}H$  is a  $m$  by  $m$  matrix. Consequently, the cost of inverse operation is  $O(\frac{2}{3}m^3)$ . Moreover, the calculation of  $H$  is outside of EM loop and only needed once, thus in each iteration, the cost of inversion step is reduced from  $O(n^3)$  to  $O(nm^2)$ .

### 3.3 Number of Clusters vs. Resolution Levels

One primary issue in cluster analysis is the determination of the number of clusters. However, the choice of the number of cluster is resolution dependent, which in our study, depends on both the resolution level of the observed data (county vs census tract) and the number of knots used to evaluate the spatial structure. Fixing the

observed resolution level (census tract in our application), for a large number of knots, we will discover more clusters but there will be higher uncertainty in the clusters; for a small number of knots, we will discover less clusters but there will be more homogeneity within a cluster. Therefore, the number of knots coarsely adjust the resolution level while the number of cluster does fine adjustment. A rigorous approach of the estimation of the number of clusters is through iteratively updating  $C$  (number of clusters) and  $K$  (number of spatial knots) and  $K'$  (number of temporal knots). In this paper, we simply fix  $K$  and  $K'$  to find the number of cluster  $C$ .

If we recognize that different number of clusters corresponds to different models, the problem can be considered as a model selection problem. AIC is a common likelihood-based model selection criterion. When the model under consideration contains random effects, the definition of the likelihood in AIC is not straightforward. Vaida and Blanchard (2005) discussed this issue by defining two variations of AIC, i.e., marginal AIC (mAIC) and conditional AIC (cAIC). Applying their arguments to our model framework, if only the fixed effect ( $\beta_Z$ ) contains information about clustering,  $mAIC = -2\log f(y|\hat{\theta}) + 2df$  should be used; within our model, the marginal likelihood  $f(y|\hat{\theta})$  is derived from integrating out  $\gamma$  and therefore,

$$Y = X\beta + \epsilon, \epsilon = \tilde{B}\gamma + \varepsilon, \epsilon \sim N(0, \tilde{B}\Gamma\tilde{B}' + \sigma_\varepsilon^2 I_{ST})$$

where  $df$  is the number of parameters in both the fixed effects and random effects.

On the other hand, our model formulation is different from Vaida and Blanchard in that there are two multivariate random variables to condition on  $Y$ , the latent variable  $Z = (Z_1, \dots, Z_S)$  in the fixed effect and the random effect  $\gamma$ . If we were to use the marginal likelihood as defined above, we would integrate out  $\gamma$  which incorporates information about the resolution level. Since the number of clusters depends on the resolution level, we consider the joint likelihood in (6) rather the marginal likelihood to select the number of clusters. Following Vaida and Blanchard's notation, we define the AIC variant with joint likelihood as  $jAIC = -2\log f(y, \gamma, Z) + 2df$  where  $df$  is a function of  $C$  and  $K$ .

## 4 Simulation Study

In the simulation study, our primary objective is to assess the estimation accuracy of the cluster membership and dependence structure under a series of spatial correlation structures between curves and varying the level of signal-to-noise ratio within each curve. We generate a synthetic data with six clusters of curves from the model

$$Y_s(t_{is}) = f_s(t_{is}) + \epsilon_{is} \text{ with } t_{is} = (i - 1)/(T - 1), \quad i = 1, \dots, T \quad (7)$$

$$f_s(t) = \sum_{k=1}^5 (\theta_{z_s, k} + \gamma_{s, k}) b_k(t) \quad s = 1, \dots, S. \quad (8)$$

where  $T = 11$  and  $S = 8484$ . The spatial coordinates are based on the location of census tracts of five south-east state - Florida, Georgia, South Carolina, North Carolina and Tennessee. The set of coefficients  $\theta_{z_s} = (\theta_{z_s, 1}, \dots, \theta_{z_s, 5})$  are the first five coefficients obtained from the Fourier decomposition of a functional pattern  $F_{z_s}$  where  $z_s \in \{1, \dots, C\}$  with  $C = 6$  (number of clusters). In this simulation study, we consider the following six patterns after re-scaling:

$$\begin{aligned} F_1(t) &= \exp(5t), \quad F_2(t) = \exp(t) \cos(t), \quad F_3(t) = \cos(2\pi t) \\ F_4(t) &= -t, \quad F_5(t) = \sin(2\pi t), \quad F_6(t) = \left(\frac{2 - 5t}{2}\right) \wedge \left[\frac{(5t - 2)^2}{3} + \sin \frac{5\pi t}{2}\right]. \end{aligned}$$

The functions  $b_k(t), k = 1, 2, \dots$  are the cosine basis functions. The error term is simulated as  $\epsilon_{is} \sim N(0, \sigma_\epsilon^2)$  with a signal-to-noise ratio approximately 1.

**Spatial dependence.** We simulate conditional dependence  $Y_s|Z$  through  $\gamma_k \sim N(0, K_s)$  where  $K_s$  is the Matérn covariance matrix of order  $\nu = \frac{2}{3}$  and varying range parameter  $\phi$ . The cluster membership  $(z_1, \dots, z_s)$  is generated from Gibbs distribution. Following HMRF methodology, the probability that a site  $s$  belongs to cluster  $c = 1, \dots, C$  is  $p(z_{s_i} = c) \propto \exp(\psi \sum_{s_j \in \partial s_i} I(z_{s_j} = c)), c \in 1, \dots, C$ .

**Simulation scenarios.** We investigate the estimation accuracy of the cluster membership and dependence structure by varying three model factors:

1. Spatial Dependence on  $Z$  controlled by the hyperparameter in Gibbs distribu-

tion  $\psi$ . The larger  $\psi$  is, the more extensive the spatial dependence of cluster membership  $Z$  is.

2. Conditional Spatial Dependence on  $Y|Z$  controlled by the hyperparameter in the Matern Kernel matrix  $\phi$ . With small  $\phi$ , each spatial unit is dependent on those that closely surround it, and consequently, the spatial pattern is more heterogenous.
3. Signal-to-Noise Ratio (SNR) of the functional pattern controlled by the ratio  $\theta/\gamma$  where  $\theta$  and  $\gamma$  are defined according to the simulation model in (7). The larger (relative to  $\theta$ )  $\gamma$  is, the more distorted the functional is from its cluster base pattern, and therefore, the functional will be less accurately identified in its corresponding cluster.

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

Table 1. *Model Settings: Spatial Dependence(Left), Conditional Dependence(Middle), SNR of the functional pattern(Right)*

	$\psi = 0.5$		$\psi = 0.9$	
	$\phi = R$	$\phi = 0.1R$	$\phi = R$	$\phi = 0.1R$
$\gamma = \gamma_0$	Weak; Strong; Low	Weak; Weak; High	Strong; Strong; High	Strong; Strong; High
$\gamma = 2\gamma_0$	Weak; Strong; Low	Weak; Weak; Low	Strong; Strong; Low	Strong; Strong; Low

where  $R = \max(d_{i,j}) = 14.67$  is the parameter used in defining  $\phi$ , the range parameter for the correlation matrix of  $\gamma$  ( $\gamma_0 \sim N(0, K_s)$ ).

**Number of clusters.** We applied jAIC to all the above eight setting for  $C$  ranging up to 10 clusters; for all cases, the minimum value for jAIC is attained at  $C = 6$  clusters as initially simulated validating jAIC as a possible criterion for selecting the number of clusters.

## 4.1 Accuracy of the Clustering Membership

In our synthetic example, as we have the true clustering membership, we can assess the accuracy of the clustering estimation for the method introduced in this paper and

other existing methods using a clustering/classification error. We use Rand index (Rand 1971), which provides the fraction of all misclustered pairs of curves representing a dissimilarity measure between the estimated clustering and true clustering. Let  $\mathcal{C} = \{f_1, \dots, f_S\}$  denote the true curves,  $\hat{\mathcal{C}} = \{\hat{f}_1, \dots, \hat{f}_S\}$  denote the estimated curves,  $T$  and  $\hat{T}$  denote the true and estimated clustering maps respectively. Rand index is defined by

$$\mathcal{R}(\mathcal{C}, \mathcal{C}') = 1 - \frac{\sum_{r < s} I(T_k(f_r, f_s) \neq \hat{T}_k(f_r, f_s))}{\binom{N}{2}}.$$

To assess the clustering accuracy, we compare:

1. *Mclust*: the model-based regularization clustering method introduced in Fraley and Raftery, 2002;
2. *Fclust*: the model-based filtering clustering technique (Fclust) introduced by James and Sugar, 2003;
3. *FMRF*: which is derived from our modeling procedure with the restriction of conditional independence on  $Y|Z$ ; and
4. *FSCM*: FSCM model using 2-step approximation (low rank kernel and ICF).

The Rand index values for the five different clustering algorithms are provided in Table 2. We compare the relative improvement in cluster estimation accuracy using

$$\text{Relative Improvement} = \frac{\text{QM}(\text{Method}_i) - \text{QM}(\text{FSCM})}{\text{QM}(\text{Method}_i)}, \quad i = 1, \dots, 3$$

where  $QM$  is a quality measure - Rand index in our comparisons. The relative improvement results are presented in Table 2. We summarize our findings as follows:

- FSCM outperforms both Mclust and Fclust under all the experimental settings.
- ➔ Under strong spatial correlation in the cluster membership  $Z$  controlled by  $\psi$ , our method improves the estimation clustering accuracy significantly.
- ➔ When SNR (controlled by  $\theta$  and  $\gamma$  levels) is low, we observe a less accurate clustering estimation overall. However, under low SNR, we find that FSCM improves the estimation significantly as compared to Mclust and Fclust.

- Comparing FCSM and FMRF, the estimation clustering accuracy does not improve under the extra conditional spatial dependence assumption when the parameter that controls it ( $\psi$ ) is small.

**Other simulation settings.** We also compared the four methods with fewer time points ( $T = 6$ ) and/or more clusters ( $C = 11$ ). Our method improves the clustering accuracy under all situations. The simulation results are not included in this manuscript due to page limitation but they may be available in a supplemental material from the authors of this article.

## 4.2 Comparisons: Accuracy of the Dependence Structure

In addition to accurate clustering estimation, our method provides a means for recovering the dependence structure between curves.

By accounting for dependence structure, we estimate the variance  $\sigma_\varepsilon^2$  accurately, which is important in inference analysis. We compare the estimated variance for different filtering methods (Fclust, FMRF and FSCM). We did not include Mclust because the variance estimation is not comparable between filtering methods and regularization methods. Table 3 shows that the variance  $\sigma_\varepsilon^2$  estimated by Fclust is larger than our method which account for spatial dependence.

The spatial dependence of the cluster membership  $Z_1, \dots, Z_S$  is modelled from a Markov Random Field where  $Z_1, \dots, Z_S$  follow the Gibbs distribution. The hyperparameter  $\psi$  in the Gibbs distribution determines the extension of the spatial correlation and it is estimated in the E-step of ICM. 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 4 lists the estimated value of  $\psi$  for all eight settings. The results show that all the estimated value are close to the true values.

Using the vector notation of the simulated functionals  $Y(t_i) = (Y_1(t_i), \dots, Y_S(t_i))'$  with  $\theta_k = (\theta_{z_1,k}, \dots, \theta_{z_S,k})'$  and  $\gamma_k = (\gamma_{1,k}, \dots, \gamma_{S,k})'$ ,  $k = 1, \dots, 5$ , the simulated

covariance structure is provided by

$$\text{cov}(Y(t_i)|Z) = \sum_{k=1}^5 b_k^2(t_i) \text{cov}(\gamma_k) + \sigma_\epsilon^2 I_S = \sum_{k=1}^5 b_k^2(t_i) K_s + \sigma_\epsilon^2 I_S.$$

On the other hand, in the model formulation of FSCM, the dependence structure on the conditional observations  $Y|Z$  is estimated by the covariance matrix:  $\text{cov}(Y(t_i)|Z) = b(t_i) \tilde{B}_s V \tilde{B}_s + \sigma_\epsilon^2 I_S$  where  $b(t_i) = \sum_{k'=1}^{K'} \sigma_{k'}^2 \tilde{B}_t^2(t_i, t_{k'})$ . To compare the estimation of the true covariance with its estimate from FSCM, we compute their largest eigenvalues. Figure 2 in the supplemental materials presents the eigenvalues of the true covariance and the ones estimated by Mclust, Fclust, FMRF and our method under setting three on the log-scale. The covariance matrix estimated by our method follows closely the true structure according to their largest eigenvalues.

## 5 Demographics Trends in Southeast U.S.

### 5.1 Description of Dataset

In this section, we apply our method to the demographics data from Sourcebook America - ESRI, a leading GIS industrial company. The data are released each year starting with 1996 to 2006 except 2002. The Census Tracts Database consists of data collected at  $> 60,000$  spatial units over the U.S. each year.

Sourcebook provides around 50 variables including basic demographics, business and spending potential attributes. In this paper, we only focus on five variables: the Per Capita Income, Percentage of Asian, Black, Hispanic and White population in five southeast states: Florida, Georgia, North Carolina, South Carolina and Tennessee. A more comprehensive study of multiple variables covering the U.S. will be provided in a subsequent application paper.

Since the boundaries of census tracts are updated by the Census Bureau every ten years (1980, 1990, 2000, 2010), our dataset includes the boundary updates from 2000. Census Bureau provides the so called 'relationship files' to document the revisions of the 1990 to 2000 census tracts. We therefore, map the data collected based on 1990

census tracts to 2000 boundaries using the information in the relationship files.

## 5.2 Results and discussion

When we apply jAIC criterion to select the number of clusters, jAIC reaches a minimum point at  $k = 17$  for Per Capita Income but does not reach a minimum for the four population percentage variables. However, if we take the difference of jAIC, i.e.,  $dAIC(k) = jAIC(k - 1) - jAIC(k)$ , the decrease of jAIC stabilizes around a  $k$  value. A more parsimonious penalty may identify an optimal value for the number of clusters; however, a small number of clusters may not capture the spatial structure complexity. To simplify the clustering interpretation, we choose the estimated number of clusters  $\hat{K} = 17$  for Per Capita Income and  $\hat{K} = 19$  for the four population variables.

For Per Capita Income, each data point is the average income over the population within a census tract, and since each census tract consists of a few thousands of people, the average income will follow an approximate normal distribution by Central Limit Theorem. We may also apply the normal limiting distribution to the percentages of Asian, Black, Hispanic and White population; however, since for many census tracts, the probability of a specific ethnic group may be very small, we used a Box-Cox transformation of these variables and cluster the transformed percentages.

In the supplemental material, for each demographic variable analyzed in this study, we provide three plots: a cluster map of the five southeast states, where clusters are coded by different colors; a cluster map of the Atlanta area with its suburban surroundings and a plot with a few selected time-dependent patterns where each pattern corresponds to the average trend within a cluster. For all the maps included in this paper, we roughly colored the high-value clusters with shades of red (hot spots) and the low-value clusters with shades of blue (cold spots). Yellow and green are between the two levels. We used purple to color the areas/clusters that experience dramatic changes over the 11 year period. Below are some interpretations for each variables:

Per Capita Income Firstly, wealth is in the hands of few census tracts of the five

states. Only 6.24% tracts fall into the top four clusters (high-income) compared with 37.5% in the bottom four clusters (low-income). High-income census tracts are located in cities such as Atlanta (Georgia), Research Triangle Area (North Carolina), Miami and beach area (Florida), and Memphis and Nashville (Tennessee). Except for the main cities, there is a lower income in states such as Georgia, Florida and Tennessee, but rather mixed for North Carolina. To visualize the spatial clustering at a finer resolution, we zoomed in the Atlanta area (Figure 3 (b), supplemental material); high income census tracts are highly concentrated in the central region (e.g. Midtown and Buckhead areas) and it decreases as the census tracts are further away from the Atlanta's central regions. This may be a common spatial pattern for metropolitan cities. Secondly, the time-dependent trend of per capita income is diverging. The increase in the trends of the low-income clusters are flatter than the ones for high-income. The low-income cluster trends increase very slowly where this slight increase could be due to economic inflation rather than a real income enhancement. On the other hand, the time-dependent cluster trends of high-income census tracts have a sharper increasing trend indicating that the income inequality worsens over a long time span.

Asian Population. Overall, Asian population has low to median percentages in Southeast US except Carolinas. Figure 4 (b) in the supplemental material shows that Metropolitan Atlanta is more dense in Asian population; this may be supported by the fact that Atlanta has the largest Korean town and highest Korean population in the US. The spatial clusters are rather small as compared to the clustering provided by other population groups. The large percentage Asian population census tracts are isolated (hot-spots) whereas there are many large regions very low in Asian population. Importantly, Figure 4 (c), supplemental material shows that most of the Asian population cluster trends increase over time as well as that the high percentage clusters tend to increase in Asian population faster than low percentage clusters.

Hispanic Population. The level of Hispanic population is low to medium in the five states except Florida, which has very high concentration of Hispanic population. The overall spatial pattern is very mixed. Figure 5 (b), supplemental material

indicates that Atlanta also has low to medium Hispanic population. However, the time-dependent pattern of most clusters show that Hispanic population is increasing (see Figure 5 (c), supplemental material).

Black and White Population. We discuss the spatial and temporal patterns of Black and White population together since we found that they contrast each other with a very clear division in both. The spatial distribution of Black and White population is roughly divided into three regions: Tennessee is very low in Black population and high in White population, Georgia, North Carolina and South Carolina are mixed with higher Black population and lower White population, and Florida has low Black population whereas upper Florida has high White population. The division is even more striking in Atlanta area: South Atlanta is high in Black population but very low in White population while North Atlanta shows the opposite pattern; moreover, the Southern Atlanta is more diverse than Northern Atlanta. The spatial clusters are larger for Black and White population than for Asian and Hispanic population; one reason is that the five states are predominantly populated by Black and White population. Moreover, the very high-value and very-low spatial clusters are the most extensive.

The time-dependent trends of Black population are mixed with both increasing and decreasing patterns. Interestingly, the White population trends decrease in most of the clusters. The highest value cluster trend is close to 100% for white population and close to 90% for Black population.

In Atlanta Metropolitan area, Cobb, DeKalb and part of Gwinnett counties (the areas colored purple (see Figure 7 (a) and 7 (c) in the supplemental material) have the most dramatic decrease (40%) of White population and the sharpest increase (30%) of Black population. Although these areas used to have high White population, they are surrounded by areas with high black population (red colored in Figure 7 (a)) and low white population (blue colored in Figure 7 (c)).

## 6 Conclusions and Further Considerations

In our model formulation, we allow for spatial dependence in both the clustering membership  $Z$  and the conditional distribution of  $Y|Z$  to define the dependence in the complete data  $(Y, Z)$ . By allowing for spatial dependence, the functionals are estimated by borrowing information from the spatially correlated functions within and between clusters. Accounting for spatial dependence results in enhanced estimation accuracy of the cluster membership under sparse temporal grid and under low signal-to-noise ratio. As provided by our simulation study, when the functions are disturbed by large errors ( $\sigma_\epsilon^2$  large) and/or highly disturbed from the base cluster pattern by increasing the ratios of  $\gamma/\theta$ , our method outperforms the existing methods which ignore the dependence structure; the cluster membership is more accurately estimated under a series of settings.

Another aspect of our estimation method is that we allow for different smoothness levels across space and time by varying the scaling variance parameters at each spatial and temporal knot ( $\sigma_{s,k}^2$  and  $\sigma_{t,k'}^2$ ). Assuming space-varying local smoothness levels is particularly important for densely observed spaces providing heterogeneous spatial information across space. Importantly, for examples with the change-of-support problem, the temporal variation may change as the support of the spatial units changes.

Generally, inference analysis relies on the estimation of the variance of each individual curve (e.g. confidence bands, temporal prediction) and of the covariance/correlation relationship between curves (e.g. spatial prediction). By correctly assuming spatial dependence in the estimation model, one may use the estimated correlation structure to make inference about each curve as well as about the relationship between curves within the same cluster or different clusters.

The clustering algorithm presented in this paper is one of the first endeavors in handling densely sampled space domains using rigorous statistical modeling. In our further studies, we will focus on even larger-scale spatial spaces - the whole US. There are several computational difficulties in applying our estimation algorithm to 60,000 spatial units. One possible solution to overcoming the large dimensionality of the correlation matrix is using compactly supported smoothing kernels instead

of globally supported smoothing kernel; advanced computational algorithms which exploit the sparsity in the kernel matrix may be therefore considered. We are also currently investigating extensions of our approach to the multivariate case with a focus on co-location of bivariate functional-spatial varying attributes.

## Acknowledgement

I would like to thank Professor Alex Gray and his PhD student Nikolaos Vasiloglou II for their C++ library - Fastlib and their valuable input along the way. Huijing Jiang has been supported by a Tennenbaum Institute fellowship while performing this research.

## Appendix: Modified R-step in Hidden Markov Random Field

We show here the derivation of *Modified R-step* in HMRF.

$$\begin{aligned}
f(z_1^m, \dots, z_S^m | Y_1, \dots, Y_S; \theta^{m-1}, \phi^{m-1}) &= f(z_1^m, \dots, z_S^m | Y_1, \dots, Y_S, \tilde{\gamma}; \theta^{m-1}, \phi^{m-1}) \\
&= \frac{f(Y_1, \dots, Y_S, z_1^m, \dots, z_S^m | \tilde{\gamma}; \theta^{m-1}, \phi^{m-1})}{f(Y_1, \dots, Y_S | \tilde{\gamma})} \\
&\propto f(Y_1, \dots, Y_S, z_1^m, \dots, z_S^m | \tilde{\gamma}; \theta^{m-1}, \phi^{m-1}) \\
&\propto f(Y_1, \dots, Y_S | z_1^m, \dots, z_S^m, \tilde{\gamma}; \theta^{m-1}) f(z_1, \dots, z_S; \phi^{m-1}) \\
&= \prod_{i=1}^S f(y_{s_i} | z_{s_i}^m, \tilde{\gamma}; \theta^{m-1}) p(z_{s_i}^m | z_{\partial s_i}^{m-1}; \psi^{m-1})
\end{aligned}$$

The first equation is valid under the assumption that  $Z$  and  $\gamma$  are independent. The last one is because  $y_s | z_s, \tilde{\gamma}$  are conditionally independent and the pseudo-likelihood function  $f(z_1, \dots, z_S) \approx \prod_{s_i=1}^S f(z_{s_i} | z_{\partial s_i}; \psi)$

## References

- [1] Archer, G.E.B., Titterton, D.M. (2002), “Parameter estimation for hidden Markov chains”, *Journal of Statistical Planning and Inference*, 108, p365.
- [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*, p39-48.
- [3] Booth, J.G., Casella, G., Hobert, J.P. (2007), “Clustering Using Objective Functions and Stochastic Search”.
- [4] Chen, L., Fuentes, M., and Davis, J. (2006). “Spatial-temporal Statistics for Ecological data”, Chapter in *Modern Statistical Computation*, to appear. Editor Jim Clark and Alan Gelfand. Wiley, New York.
- [5] Cressie, N.A.C. (1993), *Statistics For Spatial Data*, Wiley, NY.
- [6] Cressie, N., Huang, H.-C. (1999), “Classes of nonseparable, spatio-temporal stationary covariance functions.”, *Journal of the American Statistical Association*, 94, 1330-1340.
- [7] Fraley, C., Raftery, A. E. (2002), “Model-Based Clustering, Discriminant Analysis, and Density Estimation”, *Journal of the American Statistical Association*, 97, 611-631.
- [8] Gneiting, T.(2002) “Nonseparable, stationary covariance functions for space-time data“, *Journal of the American Statistical Association*, 97, 590-600.
- [9] Hastie, T. (1996), “Pseudosplines“, *Journal of the Royal Statistical Society, B*, 58, 379-396
- [10] 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):research0003.1-0003.21.

- [11] James, G.M., Sugar, C.A. (2003), “Clustering for sparsely sampled functional data”, *Journal of the American Statistical Association*, 98, p397.
- [12] Johnson M.E., Moore L.M., Ylvisaker D. (1990), “Minimax and Maximin Distance Designs“, *Journal of Statistical Planning and Inference*, 26, 131-148.
- [13] Kamman, E.E., Wand, M.P.(2003), “Geoadditive Models”, *Applied Statistics*, 52(1), p1-18.
- [14] Loan, C.V., Pitsianis, N.(1993), “Approximation with Krockner Products“, in *Linear Algebra for Large Scale and Real-Time Applications*, 293314.
- [15] Luo, Z., Wahba G., Johnson, D.(1998), “Spatial-Temporal Analysis of Temperature Using Smoothing Spline ANOVA”, *Journal of Climate*, 11, 18-28.
- [16] Matern, B. (1986), “Spatial Variation”, 2nd ed. *Lecture Notes in Statistics*, Springer Verlag, New York.
- [17] Moore, A.W. (1991), “Efficient Memory-based Learning for Robot control”, PhD thesis, Computer Laboratory, University of Cambridge.
- [18] Nychka, D.W.(2000), “Spatial-Process Estimates as Smoothers”, in *Smoothing and Regression*
- [19] Nychka, D., Saltzman, N.(1998), “Design of air quality monitoring networks. “ *Lecture Notes in Statistics*, 132, 51-76.
- [20] Ramsay, J.O., Silverman, B.W. (1997,2005), *Functional Data Analysis*, Springer, New York.
- [21] Rand, W.M. (1971), “Objective Criteria for the Evaluation of Clusterings Methods“, *J. of American Statistical Association*, 66, 846-850
- [22] Ruppert, D., Wand, M.P., Carroll, R.J. (2003), *Semiparametric Regression*, Cambridge University Press.

- [23] Sampson, P.D. and Guttorp, P. (1992), “Nonparametric estimation of nonstationary spatial covariance structure”, *Journal of the American Statistical Association*, 87, 108-119.
- [24] Serban, N., Wasserman, L. (2005), “CATS: Cluster Analysis by Transformation and Smoothing”, *J. of the American Statistical Association*, 100 ,471.
- [25] Serban, N. (2007), “Clustering in the Presence of Heteroscedastic Errors”, *Journal of Nonparametric Statistics*, to appear.
- [26] Stein, M.L.(1999), *Interpolation of Spatial Data: Some Theory for Kriging*, New York: Springer-Verlag.
- [27] Vaida, F., Blanchard, S.(2005), “Conditional Akaike information for mixed-effects models”, *Biometrika*, 92(2), 351-370
- [28] 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*.
- [29] Wu, G., Zhang, Z., Chang, E.Y. (2005), “Kronecker factorization for speeding up kernel machines” *SIAM International Conference on Data Mining* .
- [30] Wu, G., Chang, E., Chen, Y.K., Hughes, C. (2006), “Incremental Approximate Matrix Factorization for Speeding up Support Vector Machines”, *Knowledge Discovery and Data Mining*.
- [31] Zhang, H. (2004), “Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics”, *Journal of the American Statistical Association*, 99, 250-261.

Table 2. *Rand index for the clustering membership*

*Mclust(Relative Improvement), Fclust(Relative Improvement), FMRF, FSCM*

	$\psi = 0.5$	
	$\phi = R$	$\phi = 0.1R$
$\gamma = \gamma_0$	0.117(3.41%), 0.155(27.1%), 0.113, 0.113	0.122(3.27%), 0.157(24.8%), 0.118, 0.118
$\gamma = 2\gamma_0$	0.117(5.98%), 0.155(29.0%), 0.111, 0.110	0.162(11.1%), 0.176(18.2%), 0.146, 0.144
	$\psi = 0.9$	
	$\phi = R$	$\phi = 0.1R$
$\gamma = \gamma_0$	0.137(45.3%), 0.152(50.7%), 0.074, 0.075	0.126(31.0%), 0.160(45.6%), 0.089, 0.087
$\gamma = 2\gamma_0$	0.117(35.9%), 0.155(51.6%), 0.071, 0.075	0.168(28.0%), 0.194(37.6%), 0.122, 0.121

Table 3. *Variance Estimation*

*Fclust(Relative Improvement), FMRF, FSCM*

	$\psi = 0.5$	
	$\phi = R$	$\phi = 0.1R$
$\gamma = \gamma_0$	107.81(10.96%), 95.54, 95.99	110.55(10.78%), 98.20, 98.63
$\gamma = 2\gamma_0$	109.28(10.94%), 96.87, 97.32	119.97(11.05%), 106.19, 106.71
	$\psi = 0.9$	
	$\phi = R$	$\phi = 0.1R$
$\gamma = \gamma_0$	108.09(8.53%), 98.38, 98.87	110.17(7.62%), 101.27, 101.77
$\gamma = 2\gamma_0$	106.46(6.44%), 99.13, 99.60	118.34(6.45%), 110.21, 110.71

Table 4. *Estimation of the parameter  $\psi$  in the Gibbs distribution (FMRF, FSCM)*

	$\psi = 0.5$		$\psi = 0.9$	
	$\phi = R$	$\phi = 0.1R$	$\phi = R$	$\phi = 0.1R$
$\gamma_0$	0.45, 0.46	0.47, 0.47	0.89, 0.89	0.90, 0.91
$2\gamma_0$	0.47, 0.47	0.53, 0.54	0.89, 0.89	0.89, 0.90