Journal of Data Science 23 (2), 353–369
April 2025

DOI: 10.6339/25-JDS1181
Statistical Data Science

Exact Inference for Transformed Large-Scale Varying Coefficient Models with Applications

Tianyu Chen[1], Robert Habans[2], Thomas Douthat[3], Jenna Losh[4], Lida Chalangar Jalili Dehkharghani[5], and Li-Hsiang Lin[5,*]

[1]*School of Computer Science, Georgia Institute of Technology, Atlanta, GA, USA*
[2]*Kathleen Babineaux Blanco Public Policy Center, University of Louisiana at Lafayette, Lafayette, LA, USA*
[3]*Department of Environmental Sciences, Louisiana State University, Baton Rouge, LA, USA*
[4]*The Data Center, New Orleans, LA, USA*
[5]*Department of Mathematics and Statistics, Georgia State University, Atlanta, GA, USA*

Abstract

Studying migration patterns driven by extreme environmental events is crucial for building a sustainable society and stable economy. Motivated by a real dataset about human migrations, this paper develops a transformed varying coefficient model for origin and destination (OD) regression to elucidate the complex associations of migration patterns with spatio-temporal dependencies and socioeconomic factors. Existing studies often overlook the dynamic effects of these factors in OD regression. Furthermore, with the increasing ease of collecting OD data, the scale of current OD regression data is typically large, necessitating the development of methods for efficiently fitting large-scale migration data. We address the challenge by proposing a new Bayesian interpretation for the proposed OD models, leveraging sufficient statistics for efficient big data computation. Our method, inspired by migration studies, promises broad applicability across various fields, contributing to refined statistical analysis techniques. Extensive numerical studies are provided, and insights from real data analysis are shared.

Keywords *big data computation; dynamic dependencies; nonparametric regression*

1 Introduction

Understanding migration patterns driven by extreme environmental events is important for building a sustainable society and stable economy. As climate change intensifies, understanding these patterns helps policymakers anticipate and manage population shifts, mitigating the strain on urban infrastructure and resources. It also aids in planning to prevent future economic disruptions, ensuring that job markets and housing supplies align with incoming populations. This paper is motivated by the need to study migration patterns in coastal Louisiana, where recent hurricanes have increased flood risks. The major goal is to develop sophisticated regression models and fitting algorithms to understand the complex association among the migration patterns, spatial-temporal dependencies, and socioeconomic factors.

In migration studies, a primary objective is to develop effective methodologies for estimating Origin-Destination (OD) flows, which represent the number of migrations from one location to another at a specific time (Fields, 1979; Gurak and Caces, 1992). These patterns are inherently

[*] Corresponding Author Email: lhlin@gsu.edu.

Received August 16, 2024; Accepted March 20, 2025

influenced by socioeconomic and geographical factors, such as the occurrence of flood events, with complex dependencies on temporal and spatial information (LeSage and Fischer, 2009; Wood et al., 2010). To understand the dependencies, it becomes important to construct an OD regression modeling framework that incorporates appropriate features to quantify the dynamic impacts of these socioeconomic and geographical factors on migration patterns.

Existing OD regression methods have been widely applied to many applications, including passenger counts in public transportation systems (Pamuła and Żochowska, 2023), managing tourism (Flötteröd and Liu, 2014), and analyzing internet traffic (Tune et al., 2013), but are insufficient for addressing the migration problem comprehensively. Many methods focus on either temporal (Ashok, 1996) or spatial (LeSage and Fischer, 2009) information, but rarely both. Even when both temporal and spatial dimensions are considered (Noursalehi et al., 2021), interaction effects between temporal and spatial information and other input factors are often overlooked. These limitations highlight the need for an OD regression model that accounts for the dynamic spatial and temporal impacts on migration patterns and their interactions with socioeconomic and geographical variables. The broader applicability of OD matrices across various fields further underscores the importance of addressing these limitations. Thus, a more dedicated method is called for solving the problem of quantifying the dependencies for the OD regression on modeling migration patterns.

The problem of detecting the dynamic effects for migration patterns from these considered input factors can be solved by using varying coefficient (VC) regression models from statistical literature (Hastie and Tibshirani, 1993). Although many VC regression models and their fitting methods have been proposed, which are usually based on spline basis expansions (Hastie and Tibshirani, 1993), local polynomials (Fan and Zhang, 1999), and penalized basis expansions (Marx, 2009), two major challenges are still identified from the existing methods to better analyze the migration patterns. First, existing VC regression models are often limited to applications on time and two-dimensional spatial location. However, the migration OD pairs are associated with many inputs whose regression effects depend on time, location information from both origin and destination, and other social-economic factors. This requires higher-dimensional varying coefficient models. Another challenge of the analysis lies in how to efficiently fit large scale migration patterns once an adequate model is developed. As more and more migration data are shared online from existing literature, nowadays migration datasets are usually on a considerable scale; for example, the size of our motivation migration dataset is around 2TB. Implementing existing methods on such a large scale with higher varying coefficient dimensions is not straightforward because the data sets are too large to be loaded into memory. A recently developed method on varying coefficient models (Hung et al., 2022) may shed light on the development of our method, but the previous study focus on survival model with moderate dataset size, which cannot be used for the OD pair analysis. Thus, a more sophisticated statistical model and its fitting method are called for.

The proposed idea is based on a new Bayesian interpretation to connect with varying coefficient models. This connection further allows the identification of sufficient statistics (SS) of the likelihood of estimating the varying coefficient and shares an unexpected advantage in big data computing. The key benefit of the SS is their ability to be computed incrementally, processing one data point or a batch of data points at a time without the need to load the entire dataset into the computing environment. Such properties are highly beneficial for modern computing environments because they allow utilizing online or portable storage devices to analyze data through a personal laptop, which is usually associated with (relatively) small RAM and memory. This contrasts with common models used for OD pair data analyses, which lack such properties, making them less efficient for handling large datasets. For example, many OD pair data analysis are based on Poisson regression, but the sufficient statistics from the Poisson likelihood do not incorporate the SS properties the proposed method possesses (See Appendix A). Additionally, the utilization of SS paves the way for the development of advanced parallel computing techniques, enhancing the efficiency and scalability of big data processing further.

The study of sufficient statistics boasts a long and storied history in the field of statistics, serving as a foundational concept for summarizing data information effectively (Casella and Berger, 2024). While its practical application in the realm of big data is still evolving, especially concerning various models, significant strides have been made in areas such as Approximate Bayesian Computation (ABC) (Scott et al., 2022), and the Box-Cox transformation (Zhang and Yang, 2017), and data thinning (Dharamshi et al., 2024). However, to the best of our knowledge, these advancements and advantages have not yet been fully explored or implemented within the context of VC models. This gap in the literature motivates our studies for further research and development in leveraging sufficient statistics for enhanced VC model performance in big data analytics.

In addition to addressing large-scale fitting challenges, solving the migration model selection problem is crucial for improving the goodness of fit to migration data. Previous studies have employed normal-linear-based models (Beine et al., 2016) or log-linear-based models (Karemera et al., 2000). The log-linear model, inspired by gravity models derived from Newton's laws of gravity (Newton, 1687), explains migration flows as proportional to the product of area-specific factors and inversely proportional to distance. To determine whether normal-linear-based or log-linear-based models are more appropriate, we propose using a transformation on the response variable, enabling commonly used models to arise as special cases. Transformations, widely applied in statistical modeling to improve error normality (Box and Cox, 1964, 1982), ensure probability coherence (Dobson and Barnett, 2018), and achieve additivity (Lin and Roshan Joseph, 2020), can better align the model with its assumptions and enhance its goodness of fit. This motivates developing a method to select suitable transformations for practical migration modeling.

The organization of the proposed research is summarized as follows. Section 2 discusses the proposed model and its application to big data fitting. Section 3 presents several simulation studies to evaluate the method. In Section 4, we explore practical perspectives on the model, supported by additional simulation studies. Finally, Section 5 applies the proposed model to a motivating migration dataset with a conclusion and future direction discussed in Section 6.

2 Transformed Additive Varying Coefficient Model

Suppose there are I origins and J destinations. The migration number from origin $O_i \in I$ to destination $D_j \in J$ at time t is denoted by y_{ijt}. Consider the migration numbers are associated with p observed input variables $\{X_k\}_{k=1}^p$, which may change dynamically according to time t, origin O_i, destination D_j, and other socio-economic factors. The proposed VC regression model is

$$g(y_{ijt}) = \beta_0(\mathbf{h}_{ijt}) + \sum_{k=1}^{p} x_k \beta_k(\mathbf{h}_{ijt}) + \epsilon_{ijt}, \tag{1}$$

where $g(\cdot)$ is a given transformation function, such as logarithm transformation $\log(\cdot)$, $\{\epsilon_{ijt}\}$ represents the error terms following a distribution with mean zero and variance σ^2 independently for all i, j, and t, $\{\beta_k(\mathbf{h}_{ijt})\}_{k=0}^p$ are the unknown regression coefficients, which vary according to

$\mathbf{h}_{ijt}$, which is a vector including time, origin, destination, and other socio-economic or geographic factors. The core objective of this model is to develop a robust estimation method for these varying coefficient functions, particularly in the context of large-scale datasets that cannot be fully loaded into analysis software due to the memory limit of a single computing device. Other techniques about statistical computing on big data can be found in Wang et al. (2016).

We introduce this model as the Transformed Additive Varying Coefficient Model (TAVCM). TAVCM accommodates the dynamic dependency nature of migration patterns with a one-dimensional time variable, a two-dimensional spatial location, or even a higher-dimensional context. It also integrates various statistical models from the literature, including both standard VC models and log-varying coefficient models. In Section 4.3, we will discuss the selection of appropriate link functions and transformations, further enhancing the applicability and effectiveness of TAVCM in real-world scenarios. Before that, we will introduce the proposed method to fit TAVCM to a large dataset.

2.1 Fitting TAVCM in Big Data

To estimate the varying coefficients and the variance term σ^2 in model (1), we illustrate the proposed method by P-Spline method (Eilers and Marx, 1996, 2021), a nonparametric regression method through a basis expansion method with a difference penalty. An extension to other penalty methods will be discussed at the end of this section.

2.2 A Likelihood Function for TAVCM

The basis we choose for fitting is with tensor product structure (Szabó and Sriperumbudur, 2018). Specifically, if there are Q knots chosen for the basis expansions and M dimensions for the varying component $\mathbf{h}_{ijt}$, the coefficient functions can be represented as:

$$\beta_j(\mathbf{h}_{ijt}) = \sum_{q=1}^{Q} \alpha_{jq} \Phi_q(\mathbf{h}_{ijt}), \tag{2}$$

where $\Phi_q(\mathbf{h}_{ijt})$ is the kernel with tensor product structure $\prod_{m=1}^{M} \phi_q(h_{ijt}^{(m)})$, $h_{ijt}^{(m)}$ is the m-th element of vector $\mathbf{h}_{ijt}$ for $m = 1, \ldots, M$, and α_{jq} are the coefficients associated with q-th basis for j-th variable. The basis we used for each dimension of the tensor product structure is B-Spline, which can be implemented efficiently through the Cox–de Boor recursion formula (De Boor and De Boor, 1978).

To make the content clearer, we will introduce some vector and matrix notation when incorporating the proposed model into the data. Suppose all coefficients in (2) are collected and denoted by $\boldsymbol{\alpha} = (\alpha_{01}, \ldots, \alpha_{0Q}, \alpha_{11}, \ldots, \alpha_{1Q}, \ldots, \alpha_{p1}, \ldots, \alpha_{pQ})^T$. The input matrix incorporating the tensor product bases is represented by $\tilde{\mathbf{X}}$, i.e. $\tilde{\mathbf{X}} = [\mathbf{1} \otimes \phi_1(\mathbf{H}), \ldots, \mathbf{1} \otimes \phi_Q(\mathbf{H}), \mathbf{x}_1 \otimes \phi_1(\mathbf{H}), \ldots,$ $\mathbf{x}_1 \otimes \phi_Q(\mathbf{H}), \ldots, \ldots, \mathbf{x}_p \otimes \phi_1(\mathbf{H}), \ldots, \mathbf{x}_p \otimes \phi_Q(\mathbf{H})]$, where $\mathbf{H}$ is a matrix include the values of varying components for all data points. The response vector is ordered according to destination j, origin i, and time t, is denoted by $\mathbf{y} = (y_{111}, \ldots, y_{1J1}, \ldots, y_{I11}, \ldots, y_{IJ1}, \ldots, \ldots, y_{11T}, \ldots,$ $y_{1JT}, \ldots, y_{I1T}, \ldots, y_{IJT})^T$ and $\boldsymbol{\epsilon} = (\epsilon_{111}, \ldots, \epsilon_{1J1}, \ldots, \epsilon_{I11}, \ldots, \epsilon_{IJ1}, \ldots, \ldots, \epsilon_{11T}, \ldots, \epsilon_{1JT}, \ldots,$ $\epsilon_{I1T}, \ldots, \epsilon_{IJT})^T$. With these notation and follow previous results for applying P-Splines to varying coefficient models, the objective function can be expressed as $(\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\alpha})^T(\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\alpha}) + \sum_{j=0}^{p} \lambda_j (\mathbf{G}_j\boldsymbol{\alpha})^T(\mathbf{G}_j\boldsymbol{\alpha})$, where $\mathbf{G}_j$ constructs g-th order differences of $\boldsymbol{\alpha}$ for $j = 0, \ldots, p$. The estimated coefficients obtained from the objective function can be expressed as

$$\hat{\boldsymbol{\alpha}} = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} + \mathbf{P}_{\boldsymbol{\lambda}})^{-1}\tilde{\mathbf{X}}^T\mathbf{y}, \tag{3}$$

where $\mathbf{P}_{\boldsymbol{\lambda}}$ is a block diagonal matrix with diagonal elements $\lambda_0\mathbf{G}_0^T\mathbf{G}_0$, $\lambda_1\mathbf{G}_1^T\mathbf{G}_1, \ldots,$ and $\lambda_p\mathbf{G}_p^T\mathbf{G}_p$.

Because estimator (3) possesses a regression coefficient estimator from a linear model with an extra penalty term $\mathbf{P}_{\boldsymbol{\lambda}}$, this motivates us to connect a Bayesian linear model to interpret estimator (3). The Bayesian framework offers a finite-sample approach for quantifying uncertainty in varying coefficient (VC) models with multiple varying coefficient dimensions, such as model (1). This enhances the flexibility of VC modeling compared to the existing P-spline methods in the literature, which rely on asymptotic theory and are limited to models with a single varying coefficient dimension (Lu et al., 2008). The result is summarized in Theorem 1 with its proof given in Appendix B.

Theorem 1. *The posterior mean of the Bayesian linear model*

$$y = \tilde{\mathbf{X}}\boldsymbol{\beta} + \epsilon \quad \textit{with} \quad \boldsymbol{\beta} \sim N(0, \mathbf{P}_{\lambda^\star}^{-1}), \tag{4}$$

where $\epsilon \sim N(0, \sigma^2)$ *and* $\lambda^\star = (\lambda_1^\star, \ldots, \lambda_p^\star) \equiv (\lambda_1\sigma^{-2}, \ldots, \lambda_p\sigma^{-2})$, *are exactly the same as the estimator (3). Thus, the posterior variance covariance matrix of* $\hat{\boldsymbol{\alpha}}$ *can be used to quantify the uncertainty from the regression coefficient estimation.*

Note that $\mathbf{P}_{\lambda^\star}^{-1}$ in the inverse or the pseudo inverse of matrix $\mathbf{P}_{\lambda^\star}$. Theorem 1 provides a Bayesian perspective provides a basis for developing a likelihood-based method for estimating unknown parameters under big data as illustrated in the next subsection.

2.3 Large Scaled Algorithms for TAVCM Models

We consider the setting of the analysis dataset whose size is larger than the maximum memory size, making the conventional inference methods for VC models impractical as discussed in the fourth paragraph of the introduction section. To address this challenge, we divided the data into non-overlapping subdatasets, each small enough to be managed by the analysis software. To be more precise, suppose we partition the dataset into K subsets denoted by $\{\mathbf{D}_k\}_{k=1}^K = \{y_{k,\ell}, \mathbf{x}_{k,\ell}\}_{\ell=1}^{n_k}$, where n_k is the sample size of k-th sub-dataset. With the help of Theorem 1, we can record sufficient information to recover the estimators (3) and a consistent estimator of σ^2 in the error term of model (1). The results are summarized in Theorem 2, and its proof is given in Appendix C:

Theorem 2. *For each subdataset* $\mathbf{D}_k$ *for* $k = 1, \ldots, K$, *if the following information*

$$\boldsymbol{\Gamma}_k = \left(a_k \equiv \sum_{\ell=1}^{n_k} g(y_{k,\ell})^2, \mathbf{b}_k \equiv \sum_{\ell=1}^{n_k} g(y_{k,\ell})\tilde{\mathbf{x}}_{k,\ell}^T, \mathbf{C}_k \equiv \sum_{\ell=1}^{n_k} \tilde{\mathbf{x}}_{k,\ell}\tilde{\mathbf{x}}_{k,\ell}^T\right) \tag{5}$$

is recorded, then the estimator of p-spline coefficient $\boldsymbol{\alpha}$ *in (2) and a consistent estimator of* σ^2 *in (1) is recovered exactly by*

$$\hat{\boldsymbol{\alpha}} = (\sum_k \mathbf{C}_k + \mathbf{P}_\lambda)^{-1}(\sum_k \mathbf{b}_k), \tag{6}$$

and

$$\hat{\sigma}^2 = \frac{\sum_k a_k - (\sum_k \mathbf{b}_k)^T(\sum_k \mathbf{C}_k + \mathbf{P}_\lambda)^{-1}\sum_k(\mathbf{b}_k)}{n}, \tag{7}$$

respectively, where a_k, $\mathbf{b}_k$, *and* $\mathbf{C}_k$ *are as defined in Equation (5).*

Algorithm 1: Exact recover for large scale varying coefficient models.

```
Input : K partition non-overlapped datasets $\{\mathbf{D}_k\}_{k=1}^K$ and knots.
for k $i \in 1, \ldots, K$ do
    Record the sufficient information $\mathcal{P}_k$ from equation (3);
    Remove the k-th partition dataset in the memory space
Output: $\{\mathcal{P}_k\}_{k=1}^K$.
```

This theorem indicates that by recording and utilizing these sufficient statistics, one can estimate the model parameters without needing to load the entire dataset into memory, thus making the estimation feasible for large-scale datasets. Note that the calculation of $\mathbf{P}_{\lambda^\star}$ is usually not expensive, so can be easily plug-in into equations (6) and (7) directly. Additionally, as the estimator is recovered exactly, many theoretical results derived for the P-Spline method, such as the consistency properties of the estimators and the convergent rates from Claeskens et al. (2009), can be applied to the proposed estimator of $\boldsymbol{\alpha}$. The algorithm for the aforementioned procedure to fit the proposed model is given in Algorithm 1.

The proposed framework is illustrated by the P-Spline method, but it can be applied to other smoothing method. For example, if we use local linear expansion for expanding the varying coefficient all the varying coefficient functions $\{\beta_j(\mathbf{h}_{ijt})\}_{j=0}^p$ (Cai et al., 2000; Hung et al., 2022) at $\mathbf{h}_0$, then $\beta_j(\mathbf{h}_{ijt}) \approx \beta_j(\mathbf{h}_0) + \sum_{m=1}^M \frac{\partial \beta_j(\mathbf{h}_0)}{\partial h^{(m)}}(h_{ijt}^{(m)} - h_0^0)$ for $j = 0, 1, 2, \ldots, p$. For each dimension j, there are $M + 1$ coefficient functions $(\beta_j(\mathbf{h}_0), \frac{\partial \beta_j(\mathbf{h}_0)}{\partial h^{(1)}}(\mathbf{h}_0), \ldots, \frac{\partial \beta_j(\mathbf{h}_0)}{\partial h^{(M)}}(\mathbf{h}_0))$ corresponding to (M+1) columns $(\mathbf{1}, \mathbf{X}_j \otimes (h_{ijt}^{(1)} - h_0^{(1)}), \ldots, \mathbf{X}_j \otimes (h_{ijt}^{(M)} - h_0^{(M)}))$. With collecting all coefficients as a vector $\boldsymbol{\beta}$ and assigning their corresponding columns to $\tilde{X}$, we have the following results to fit the VC model based on the local polynomials. Its proof is given in Appendix D.

Theorem 3. *Suppose a local polynomial with bandwidth* $\boldsymbol{\delta}$ *are used. Then the posterior mean of the following Bayesian linear model*

$$y = \tilde{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{8}$$

where $\boldsymbol{\beta} \sim N(0, \mathbf{I})$ *and* $\epsilon \sim N(0, \mathbf{W})$ *and* $\mathbf{W}$ *is a diagonal matrix whose* i*-th diagonal elements is* $K_\delta(||\mathbf{h} - \mathbf{h}_0||)$ *for a given bandwidth* δ *and a known kernel function that can be evaluated at the distance between* $\mathbf{h}$ *and* $\mathbf{h}_0$*, denoted by* $||\mathbf{h} - \mathbf{h}_0||$*, is exactly the local polynomial estimator of* $\boldsymbol{\beta}$ *under model (1). Furthermore, suppose we partition the dataset into K subdatasets. For each subdataset* $\mathbf{D}_k$ *for* $k = 1, \ldots, K$*, if the following information*

$$\boldsymbol{\Gamma}_k = \left(\sum_{\ell=1}^{n_k} g(y_{k,\ell})^2, \sum_{\ell=1}^{n_k} g(y_{k,\ell}) \tilde{\mathbf{x}}_{k,\ell}^T, \sum_{\ell=1}^{n_k} \tilde{\mathbf{x}}_{k,\ell} \tilde{\mathbf{x}}_{k,\ell}^T \right) \tag{9}$$

is recorded, then its sufficient statistics for estimating $\boldsymbol{\alpha}$*, the local polynomial estimator and hence the estimator obtained from the average of the squares of residuals for* σ^2 *based on local polynomial method is also recovered exactly.*

3 Numerical Studies

In this section, we conduct numerical studies to evaluate the performance of our proposed method. In Section 3.1, the first simulation study focuses on whether our method can precisely reconstruct the underlying regression relationships from simulated data. The simulation setting is set to be tested conveniently in a common computing environment through R, so we can compare with related methods from other packages. Then, in Section 3.2, an advanced computing environment is designed, so we are able to test the proposed method under larger datasets on a scale close to our motivation real datasets. For both sections, we investigate the computational efficiency and root mean squared prediction error (MSE) of the proposed method with comparisons with other methods.

3.1 Comparisons with Existing Methods

In this simulation, data is generated for the following varying coefficient model:

$$\text{Model 1: } y = \exp(\beta_0 + \beta_1(t)X_1 + \beta_2 X_2 + \epsilon),$$

where there are two input variables with the first one $X_1 \sim N(0,1)$ associated with a varying coefficient $\beta_1(t) = \sin(2.5\pi t)$ and the other input variable $X_2 \sim \chi^2_4$ associated with a constant coefficient $\beta_2 = 2$, the time t is generated independently and identically following a discrete uniform distribution with values taking on $\{1/100, 2/100, 3/100, \ldots, 1\}$, and an error term is drawn from a normal distribution with mean 0 and variance 1 and served as noise in the regression. The simulation setting is repeated 1000 times under various sample sizes $n = 1000, 10000, 100000$.

The simulation results are summarized in Table 1, where the unit of time is second. The proposed method (column 1) implemented with number of basis 10 and difference order 1 is compared with the linear model from function lm in R (column 2), and varying coefficient regression methods from other R packages, including $tvReg$ (column 3, Casas and Fernandez-Casal (2023)), $tvem$ (column 4, Dziakm et al. (2023)), and $varycoef$ (column 5, Dambon et al. (2022)). The main differences from the varying coefficient methods are in the smoothing techniques, where $tvReg$ is based on local polynomials, $tvem$ uses truncated power basis with ridge penalty, and $varycoef$ is based on Gaussian processes. From Table 1, we observe that our method (column

Table 1: Summary of numerical comparison results in terms of root mean square prediction errors (RMSPE) and the fitting time of the methods for Section 3.1.

Case 1-I: Sample Size 1000 (Mean (sd))					
Methods	TAVCM	Linear Model	tvReg	tvem	varycoef
Time	1.29 (0.052)	0.001 (0.001)	5.445 (1.053)	8.096 (1.151)	4.271 (1.102)
RMSPE	0.103 (0.022)	0.697 (0.002)	0.131 (0.056)	0.131 (0.056)	0.127 (0.059)
Case 1-II: Sample Size 10000 (Mean (sd))					
Methods	TAVCM	Linear Model	tvReg	tvem	varycoef
Time	1.70 (0.067)	0.002 (0.005)	314.448 (724.141)	119.772 (16.227)	637.25 (15.375)
RMSPE	0.033 (0.007)	0.696 (0.001)	0.039 (0.008)	0.037 (0.009)	0.034 (0.008)
Case 1-III: Sample Size 100000 (Mean (sd))					
Methods	TAVCM	Linear Model	tvReg	tvem	varycoef
Time	2.94 (0.129)	0.002 (0.003)	X	X	X
RMSPE	0.002 (0.001)	0.683 (0.002)	X	X	X

1) has comparable prediction performance compared with other smoothing methods (columns 3 to 5), evaluated by a separated testing dataset with 1000 data points but the proposed method requires less time for fitting. Thus, in terms of both statistical efficiency and computational efficiency, our method is relatively better. Also, all the nonlinear smoothing methods are performed better than the linear method (Column 2), which matches our intuition because the true model is a nonlinear regression. Note that when the sample size is 100,000 (case III), the running time is more than three hours for each simulation, so we do not record the results, which is much less efficient compared with our method, so we do not present the results. Note that the simulation is compared under a laptop with a 2.2 GHz processor and 12 GB memory.

3.2 Larger Sample Size Testing

We further evaluated the proposed method on a larger sample size simulation using Databricks on Microsoft Azure. Note that running R on Databricks in Microsoft Azure offers seamless integration with Azure services and cluster computers. This integration provides optimized Spark clusters for distributed R computations, enabling R users to benefit from a scalable, secure, and collaborative analytics environment tailored to cloud workflows. However, based on our experience, not all R packages can be used on Databricks due to the distributed nature of the environment and its reliance on Spark for computation. For our simulation comparisons, there are no existing packages for direct comparison with the proposed method. Therefore, we implemented our own code based on Theorem 3 as the main comparison method. It is important to note that the simulation environment functions similarly to a personal laptop but with increased memory capacity to handle the larger sample sizes. The results are summarized below. From Table 2, we observe that the proposed method remains valid, as evidenced by the low MSE and reasonable computational time. Note that the used cluster computers are 3.0 GHz processors with 192 GB memory.

Table 2: Summary of numerical comparison results in terms of root mean square prediction errors (RMSPE) and the fitting time of the methods for Section 3.2.

Case 1-IV: Sample Size 10^6 (Mean (sd))			
1-4 Methods	TAVCM	Linear Model	Theorem 3
Time	1.292 (0.101)	0.001 (0.001)	1.371 (0.091)
RMSPE	0.002 (0.001)	0.697 (0.002)	0.002 (0.001)
Case 1-V: Sample Size 10^7 (Mean (sd))			
Methods	TAVCM	Linear Model	Theorem 3
Time	2.348 (0.156)	0.002 (0.005)	2.339 (0.127)
RMSPE	0.001 (0.001)	0.696 (0.001)	0.001 (0.001)
Case 1-VI: Sample Size 10^8 (Mean (sd))			
Methods	TAVCM	Linear Model	Theorem 3
Time	4.376 (0.215)	0.002 (0.003)	4.312 (0.207)
RMSPE	0.000 (0.001)	0.683 (0.002)	0.000 (0.001)

4 Practice Considerations

After testing the performance of the proposed method, we further discuss and extend the proposed method to be used in real data analysis. The topics include parallel computing, model selection, and parameter tuning.

4.1 Incorporation with Parallel Computing

The proposed method can be significantly enhanced by incorporating parallel computing techniques to accelerate the computational process. Notably, the collection of sufficient information from each k-th partition of the dataset, as outlined in Equation (3), is independent of the data in other partitions. This characteristic allows for the effective parallelization of the data processing workflow. Specifically, in Algorithm 1, the computation for each partition can be performed concurrently without the need for inter-process communication or dependency on results from other partitions. By leveraging parallel computing, we can distribute the task of collecting and processing data across multiple processors or cores. This approach not only reduces the overall computational time but also improves the efficiency of handling large-scale datasets. The implementation of parallel processing frameworks by using parallel libraries in R (Team, 2024; Daniel et al., 2022) or other high-performance computing environments can substantially accelerate the execution of the for-loop iterations in Algorithm 1. Consequently, this enhancement enables the proposed method to manage larger datasets more effectively, making it more suitable for practical applications where computational resources are a critical factor.

4.2 Variable, Basis, and Knot Place Selections

In extending the proposed method to model selection, we leverage the fact that the likelihood function, as specified in Equation (2), can be recovered exactly. This property allows us to apply important asymptotic results from previous studies to our framework. Specifically, the exact recovery of the likelihood function facilitates the use of likelihood ratio tests (Wilks, 1938) for model selection. The likelihood ratio test is a powerful statistical tool for comparing nested models by evaluating the ratio of their maximum likelihoods. The details are summarized in Algorithm 2. In our context, this approach enables us to systematically assess the adequacy of different models in terms of variable inclusion, basis functions, and knot placements (for some spline methods) or bandwidth for the local polynomial method. Note that for the demonstration method P-Spline, knot placement is usually not an issue due to the difference penalty (Eilers and Marx, 2021), so the method is usually equipped with equally-spaced knots.

By implementing Algorithm 2, we can rigorously determine the most appropriate model

Algorithm 2: Large scale likelihood ratio tests for model selection.

```
Input : K partition non-overlapped datasets $\{\mathbf{D}_k\}_{k=1}^{K}$ and null hypothesis $\mathcal{H}_0$ and an
  alternative $\mathcal{H}_1$ .
Step 1: Use Algorithm 1 to find estimates of regression coefficients under $\mathcal{H}_0$ and an
  alternative $\mathcal{H}_1$;
Step 2: Under the estimates from Step 1, use the SS information in Theorem 1 to
recover the likelihood values under $\mathcal{H}_0$ and an alternative $\mathcal{H}_1$.
Output: Returen the likelihood ratio values.
```

configuration that balances complexity and fit. This algorithm assesses various combinations of variables, basis functions, and knot locations by comparing their likelihood ratios, thereby ensuring that the selected model is both statistically robust and computationally efficient. This extension not only enhances the model's flexibility but also improves its performance in capturing the underlying structure of the data.

To demonstrate the idea, we modify the simulation section in Section 3.1 by adding three extra input variables denoted by X_1, X_2 and X_3. They are independently generated from a normal distribution with mean 0 and standard deviation 0.5. These extra input variables are not associated with the responses, i.e., the true model is the same one as we had in Section 3.1. Then to see the effectiveness of the method, we consider the following 4 hypothesis testing examples:

Example 1: $H_0 : \beta_3 = 0$ versus $H_a : \beta_3 \neq 0$
Example 2: $H_0 : \beta_0 = 0$ versus $H_a : \beta_0 \neq 0$
Example 3: $H_0 : \beta_2 = 0$ versus $H_a : \beta_2 \neq 0$
Example 4: $H_0 : \beta_1(t) = 0$ versus $H_a : \beta_1(t) \neq 0$

Intuitively, X_3 is not associated with the responses, so the frequency of rejection should be close to the significant level α we set. Such intuition is supported by the numerical results summarized in the first row of Table 3. We see the frequency of rejection of the test for example 1 (first row of Table 3) is close to $\alpha = 0.05$, and as sample size increases, the values are closer to α. For examples 2 to 4, the alternative hypothesis is true, so we report the absolute difference between the frequency of not rejection and the type II errors of the tests. The type II error is calculated under $\beta_0 = 2$ for example 2 and $\beta_2 = 2.5$ for example 3 for demonstration. The results are reported in the second and third rows of Table 3, and the results show the difference is approached to 0 when the sample size is increasing, matching our intuition. Additionally, we can also apply the ratio test to test whether there is a varying coefficient effect on the first input X_1. The power is calculated under the optimized values to approximate the true model from the basis expansion by using least square approximation. Note the degrees of freedom of the chi square test are different for the four tests, where for Example 4 the degrees of freedom (df) is 10 (the number of B-Spline basis) while for other examples the df is 1. The results are reported in the second last row of Table 3, and we also observe the absolute error converges to 0 as the same size increases.

4.3 Transformation Model Selection

While the previous method operates under a given transformation, we can extend our approach to include model selection involving parametric transformations. A notable example is the Box-Cox transformation, which can be integrated with the proposed model to encompass a broader range of statistical models applicable to migration datasets. The Box-Cox transformation (Box and Cox, 1964) is controlled by a single parameter, denoted by λ, which allows for a flexible adjustment of the response variable to better fit the data. By incorporating this transformation, our model can adapt to various types of non-linearity and heteroscedasticity present in the data. The parameter λ can be estimated using the proposed likelihood function with an appropriate Jacobian transformation, as this can be accurately recovered based on a corollary of Theorem 2. This extension necessitates an adaptation of Algorithm 2 to accommodate the estimation of the transformed varying coefficient models (TVCM) with parametric transformation. The revised algorithm will include steps to optimize λ along with the model parameters, ensuring that the transformation is appropriately applied to enhance the model's fit. This extension allows for

Table 3: Numerical simulation results for model selection by using Algorithm 2: Examples 1 and 5 are compared with a significant level $\alpha = 0.05$, and examples 2 to 4 demonstrate the power value of the associated test.

Methods	100	1000	10000	100000
Example 1	0.0479	0.0497	0.0503	0.0499
Example 2	0.98	0.99	0.99	1.00
Example 3	0.98	0.99	0.99	1.00
Example 4	0.97	0.99	0.99	1.00
Example 5	0.0471	0.0487	0.0492	0.0501

more comprehensive model selection, ensuring that the parametric transformation is effectively utilized to capture the underlying relationships in the data, thereby improving the robustness and accuracy of the statistical analysis.

The results of our analysis are demonstrated for specific settings of the Box-Cox transformation parameter, namely $\lambda = 0$ and $\lambda = 1$, which extend commonly utilized models for migration data. When $\lambda = 0$, the model extends log-linear models to a varying coefficient model with a logarithmic transformation on the responses, while $\lambda = 1$ aligns the model with a standard VC model, extending the normal linear models. This motivated the testing of the hypothesis.

Example 5: $H_0 : \lambda = 0$ versus $H_a : \lambda = 1$.

The results of this hypothesis test are reported in the last row of Table 3, where the true model is designed under $H0$ based on the simulation setting in Section 4.1. The results are close to the set type I error α value 0.05. This consistency underscores the robustness of our proposed model across different transformation settings and confirms its efficacy in analyzing migration data.

4.4 Cross Validation for Model Selection

Except for likelihood ratio tests, cross-validation is also a common method for model selection. Theorems 1 and 2 delineate the process for recovering the estimator $\boldsymbol{\alpha}$, which can subsequently be utilized to calculate the fitted values under partitioned data, as discussed at the beginning of this section. Given that our method functions as a linear smoother, we can apply the common cross-validation (CV) and generalized cross-validation (GCV) techniques within the partitioned dataset framework, as summarized by the theorem below. It is important to note that the theorem and its proof predominantly follow the standard CV and GCV implementations for linear smoothers (Simonoff, 2012). For the purpose of this study, we express the results within our subsampling (SS) framework (Theorem 2), and for the sake of completeness, the proof is provided in Appendix E. Note that in the formula $g(\hat{y}_v)$ and s_{vv} for all v are dependent on λ.

Theorem 4 (Cross Validation for Tuning). *Consider the linear smoother matrix* $\mathbf{S}$*, where* $\mathbf{S}\mathbf{y} = \tilde{\mathbf{X}}^T\hat{\boldsymbol{\alpha}} = \tilde{\mathbf{X}}^T(\sum_k \mathbf{C}_k + \mathbf{P}_\lambda)^{-1}(\sum_k \mathbf{b}_k)$*. Let* s_{vw} *be the* (v, w)*-th element of the smoother matrix* $\mathbf{S}$*. Then, the LOOCV criterion for tuning the penalty parameter* λ *is*

$$\mathrm{cv}_\lambda = \frac{1}{n}\sum_{v=1}^{n}\left(\frac{g(y_v) - g(\hat{y}_v)}{1 - s_{vv}}\right)^2.$$

Furthermore, the generalized cross-validation (GCV) criterion replaces the individual s_{vv} terms with their average $\mathrm{tr}(\mathbf{S})/n$*, resulting in* $\mathrm{gcv}_\lambda = \frac{1}{n}\sum_{v=1}^{n}\left(\frac{g(y_v)-g(\hat{y}_v)}{1-tr(\mathbf{S})/n}\right)^2$*, which is a simple function of the average squared residual.*

5 Analyzing a Large-Scale Migration Dataset

In this section, we apply the proposed method to a migration dataset from *Infutor*: As discussed in Diamond et al. (2019) and Phillips (2020), *Infutor* is a private data aggregator that provides address histories of individuals using a mix of private and public inputs. We used those address histories and their changes to create Census tract-to- Census tract migration flows by month, and to estimate monthly population change by aggregating individual micro data records. Sample data, and detailed methods are available in the OPEN ICPSR repository (Habans and Douthat, 2024).

The dataset we used mainly comprises origin-destination (OD) pairs recorded from 2000 to 2020. Each pair is identified using the Federal Information Processing Standard (FIPS) code, which uniquely distinguishes all county areas within the United States. The dataset also includes socioeconomic information for each county, encompassing approximately 19,872,391 OD pairs. Due to the substantial size of the dataset (1 TB), our big data methodology is essential for effective analysis. A summary of all input variables is provided in Table 2 of Appendix F in the supplemental material. Additionally, initial data exploration with visualizations, also in Appendix F, reveals that certain input factors are highly correlated with migration counts, though correlation strengths vary across states. This observation motivates further investigation with the following objectives: (i) Identify which input variables significantly influence migration patterns. (ii) Examine how the effects of these variables vary across spatial locations, specifically the origins and destinations of OD pairs. Insights from these findings may also offer a deeper understanding of localized migration dynamics. In this study, we further explore the third analysis goal (iii) exploring migration patterns in coastal Louisiana to know if the important factors are different from general USA patterns. Note that since the number of observations from Alaska state are relatively small compared to other states, they are excluded from this analysis.

To explore analysis goals (i) to (iii), we first discuss more insights about the connection and extension of the proposed model (1) and *Gravity models*, which serve as an important model in analyzing and interpreting migration patterns in the literature (Karemera et al., 2000). Gravity models initially derived from Isaac Newton's laws of gravity, describe how every particle in the universe attracts every other particle (Newton, 1687). These models have been extended beyond physics to various disciplines, including economics and social sciences. One notable application is in analyzing the volume of trade between countries. More recently, gravity models have been applied to study migration patterns, describing how the migration numbers between two areas are proportional to the product of factors from these areas and inversely proportional to the distance between them. This interpretation connects to our model (1) for analyzing migration patterns. By taking the exponential form of model (1) while considering the last input variable as the distance between two areas, denoted by $x_p = \delta_{ij}$, we can re-express the proposed model as:

$$y_{ijt} = \exp(\beta_0(\mathbf{h}_{ijt}))\frac{\prod_{k=1}^{p-1}\exp(x_k\beta_k(\mathbf{h}_{ijt}))}{\exp(\delta_{ij}\tilde{\beta}_p(\mathbf{h}_{ijt}))}\exp(\epsilon_{ijt}),$$

where $\tilde{\beta}_p(\mathbf{h}_{ijt}) = -\beta_p(\mathbf{h}_{ijt})$. This extension of the gravity model, incorporating varying coefficients from factors, motivated us to apply the proposed method to analyze migration patterns.

By using model (1) with applying Algorithm 1 to our dataset, we obtained varying coefficient estimates. To demonstrate the estimating results efficiently, we take the average of the effect values from the same state (according to the FIPS codes), and present the average effect values for each state in Figure 1. The figure includes the varying effects of three important variables. Figure 1A is for distance effects, demonstrating that larger distances of moving reduce the number of migration numbers, which matches our intuition. It also demonstrates that an increase in distance tends to result in a more pronounced decrease in migration events along the East Coast compared to the West Coast. Figure 1B is the varying effect from flood claim numbers in the origin of the migration patterns, and the claim numbers in the destination are shown in Figure 1C. From Figure 1B, we observe that if there are more flood claims, people living in the southeastern area are more likely to move than people in northern western area. Interestingly, Figure 1C demonstrates that if your destination is in northwestern area, the claim numbers possess less impact on the migration numbers. Additionally, we employed Algorithm 2 to evaluate

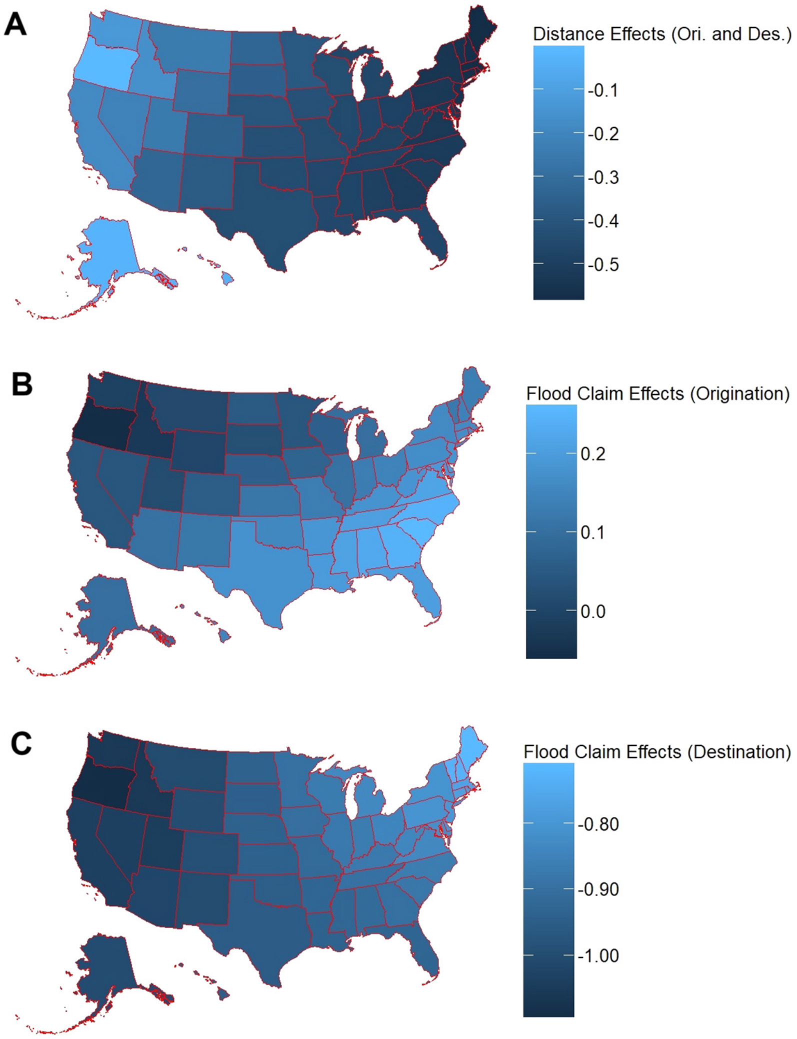
A
Distance Effects (Ori. and Des.)
-0.1
-0.2
-0.3
-0.4
-0.5
B
Flood Claim Effects (Origination)
0.2
0.1
0.0
C
Flood Claim Effects (Destination)
-0.80
-0.90
-1.00

Figure 1: Demonstration of some spatial varying coefficient effect figures.

and quantify the importance of input variables from the dataset. The results, summarized in Table 1 in Appendix F of the supplemental material, reveal the findings of significant important variables for migration.

One of the important variables is “the origin from the coastal Area”. Previous studies, such as Hauer et al. (2019), indicated no significant landward movement by examining 80 years of population migration and shoreline change in Louisiana (1932 to 2010). This finding motivated us to further examine these patterns with a focus on origins from Louisiana. We incorporate our methods with indicator functions as inputs to identify the movement direction and analyze the first nine years of data (2000 to 2008) with a focus on Louisiana origins, we found that the “origin from the coastal Area” was not an important variable during this period. Further analysis revealed that most origins exhibited landward population movement, perpendicular to the shoreline, exceeding 15 miles but not far inland. These findings suggest that coastal Louisiana’s historical population has moved in response to shoreline encroachment, demonstrating that our model can provide valuable insights into the dynamic patterns of migration.

6 Conclusion and Future Research Discussion

This study introduces a novel method for fitting large-scale transformed varying coefficient models, motivated by the needs of a human migration analysis. The method’s validity is underscored by the identification of a straightforward approach to recording sufficient statistics, which are then utilized to precisely recover all necessary estimators. This approach ensures the robustness and accuracy of the method, particularly in handling the complexity inherent in large-scale datasets. Theoretical extensions of the method are also provided, allowing for its application to other basis expansion techniques. Furthermore, practical considerations, such as testing for significant variables, assessing the goodness of model fit, and parameter tuning, have been addressed through extensive simulation studies. These simulations not only validate the method’s efficacy but also demonstrate its flexibility and adaptability to various modeling scenarios.

The proposed methods have been applied to a large-scale migration dataset observed in the United States from 2000 to 2020, yielding significant findings. Notably, one key observation from the real data analysis, detailed in Section 5, reveals an increased inward migration trend in Louisiana, a pattern not identified in previous studies. This finding highlights the importance of certain migration patterns and suggests the need for more advanced modeling techniques.

Although our model is inspired by extensive studies on migration data, its applicability extends far beyond this initial domain. It offers a robust framework for strengthening analytical results in various applications of varying coefficient models. Given the broad range of applications already identified in this research area, our developed method promises versatile utility. It can be applied across diverse fields to enhance the accuracy and depth of analytical insights, making it a valuable tool for researchers and practitioners alike. This broad applicability ensures that our method can contribute significantly to the ongoing efforts to refine and optimize statistical analysis techniques across different datasets and contexts. It would also be interesting to explore whether the concept of sufficient statistics can be extended to accommodate advanced deep neural network models (Hung et al., 2025).

We sincerely appreciate the valuable comments from the associate editor and reviewer. Our current approach incorporates temporal and spatial dependencies through the varying coefficients in model (1), but the error term assumes independence. This limitation can be addressed by extending the VC model to include random intercept and slope effects within a mixed-effects framework. These random effects serve as latent variables, capturing additional dependencies and unobserved heterogeneity among underserved groups, such as variations influenced by household income levels and educational attainment within regions. This research direction is both intriguing and practical. However, further discussions with domain experts are needed to define appropriate grouping criteria to accurately represent underserved populations. Additionally, advanced techniques may be required to efficiently extract latent random effects in large-scale datasets. Our framework could be extended to derive mixed-effects models for big data using an enhanced MCEM approach (Levine and Casella, 2001). Such advancements could yield valuable insights into clustering group behaviors and the migration patterns of underserved populations, providing a promising avenue for future research.

Supplementary Material

We provide more technical details, simulation results, and real data analysis as the pdf file in the supplemental material. Data files and simulation code used in the article can also be found in the supplemental material.

Funding

This study was supported by the U.S. Department of the Treasury through the Louisiana Coastal Protection and Restoration Authority's Center of Excellence Research Grants Program under the Resources and Ecosystems Sustainability, Tourist Opportunities, and Revived Economies of the Gulf Coast States Act of 2012 (RESTORE Act) (Award No. 1 RCEGR260007-01-00). The statements, findings, conclusions, and recommendations are those of the authors and do not necessarily reflect the views of the Department of the Treasury.

References

Ashok K (1996). Estimation and prediction of time-dependent origin-destination flows, Ph.D. thesis, Massachusetts Institute of Technology.

Beine M, Bertoli S, Fernández-Huertas Moraga J (2016). A practitioners' guide to gravity models of international migration. *World Economy*, 39: 496–512. https://doi.org/10.1111/twec.12265

Box GEP, Cox DR (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, 26: 211–243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

Box GEP, Cox DR (1982). An analysis of transformations revisited, rebutted. *Journal of the American Statistical Association*, 77: 209–210.

Cai Z, Fan J, Li R (2000). Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association*, 95: 888–902. https://doi.org/10.1080/01621459.2000.10474280

Casas I, Fernandez-Casal R (2023). tvReg: Time-varying coefficient linear regression for single and multi-equations in R. R package version 0.5.9.

Casella G, Berger R (2024). *Statistical Inference*. Chapman and Hall/CRC, Boca Raton, FL.

Claeskens G, Krivobokova T, Opsomer JD (2009). Asymptotic properties of penalized spline estimators. *Biometrika*, 96: 529–544. https://doi.org/10.1093/biomet/asp035

Dambon J, Sigrist F, Furrer R (2022). varycoef: Modeling Spatially Varying Coefficients. R package version 0.3.4.

Daniel F, Weston S, Tenenbaum D (2022). Parallel: Foreach Parallel Adaptor for the 'parallel'. Package. R package version 3.6.2.

De Boor C, De Boor C (1978). *A Practical Guide to Splines*, volume 27. Springer, New York, NY.

Dharamshi A, Neufeld A, Motwani K, Gao LL, Witten D, Bien J (2025). Generalized data thinning using sufficient statistics. *Journal of the American Statistical Association*, 120: 511–523. https://doi.org/10.1080/01621459.2024.2353948

Diamond R, McQuade T, Qian F (2019). The effects of rent control expansion on tenants, landlords, and inequality: Evidence from San Francisco. *American Economic Review*, 109(9): 3365–3394.

Dobson AJ, Barnett AG (2018). *An Introduction to Generalized Linear Models*. Chapman and Hall/CRC, Boca Raton, FL.

Dziakm J, Coffman DL, Li R, Litson K, Yajnaseni C (2023). varycoef: Modeling Spatially Varying Coefficients. R package version 1.4.1.

Eilers PH, Marx BD (1996). Flexible smoothing with b-splines and penalties. *Statistical Science*, 11: 89–121. https://doi.org/10.1214/ss/1038425655

Eilers PH, Marx BD (2021). *Practical Smoothing: The Joys of P-Splines*. Cambridge University Press, Cambridge, United Kingdom.

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

Fields GS (1979). Place-to-place migration: Some new evidence. *Review of Economics and Statistics*, 61: 21–32. https://doi.org/10.2307/1924827

Flötteröd G, Liu R (2014). Disaggregate path flow estimation in an iterated dynamic traffic assignment microsimulation. *Journal of Intelligent Transportation Systems*, 18(2): 204–214. https://doi.org/10.1080/15472450.2013.806854

Gurak DT, Caces F (1992). Migration networks and the shaping of migration systems. In: *International Migration Systems: A Global Approach* (Mary M. Kritz, Lin Lean Lim, Hania Zlotnik, eds.), Chapter 9: 150–176.

Habans R, Douthat T (2024). *Past and Future Migration in Coastal Louisiana*. Inter-university Consortium for Political and Social Research [distributor], Ann Arbor, MI.

Hastie T, Tibshirani R (1993). Varying coefficient models. *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, 55: 757–779. https://doi.org/10.1111/j.2517-6161.1993.tb01939.x

Hung NYT, Lin LH, Calhoun VD (2025). Deep p-spline: Theory, fast tuning, and application. arXiv preprint: https://arxiv.org/abs/2501.01376.

Hung Y, Lin LH, Wu CJ (2022). Varying coefficient frailty models with applications in single molecular experiments. *Biometrics*, 78: 474–486.

Karemera D, Oguledo VI, Davis B (2000). A gravity model analysis of international migration to North America. *Applied Economics*, 32: 1745–1755.

LeSage JP, Fischer MM (2009). Spatial econometric methods for modeling origin-destination flows. In: *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications* (Manfred M Fischer, Arthur Getis, eds.), 409–433. Springer, New York, NY.

Levine RA, Casella G (2001). Implementations of the Monte Carlo em algorithm. *Journal of Computational and Graphical Statistics*, 10(3): 422–439.

Lin LH, Roshan Joseph V (2020). Transformation and additivity in Gaussian processes. *Technometrics*, 62: 525–535.

Lu Y, Zhang R, Zhu L (2008). Penalized spline estimation for varying-coefficient models. *Communications in Statistics - Theory and Methods*, 37(14): 2249–2261.

Marx BD (2009). P-spline varying coefficient models for complex data. In: *Statistical Modelling and Regression Structures*, (Thomas Kneib, Gerhard Tutz, eds.), 19–43. Springer, New York, NY.

Noursalehi P, Koutsopoulos HN, Zhao J (2021). Dynamic origin-destination prediction in urban rail systems: A multi-resolution spatio-temporal deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 23(6): 5106–5115.

Pamuła T, Żochowska R (2023). Estimation and prediction of the od matrix in uncongested urban road network based on traffic flows using deep learning. *Engineering Applications of Artificial Intelligence*, 117: 105550.

Phillips DC (2020). Measuring housing stability with consumer reference data. *Demography*, 57(4): 1323–1344.

Scott SL, Blocker AW, Bonassi FV, Chipman HA, George EI, McCulloch RE (2016). Bayes and big data: The consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management*, 11: 78–88. Routledge: Oxfordshire, United Kingdom.

Simonoff JS (2012). *Smoothing Methods in Statistics*. Springer, New York, NY.

Szabó Z, Sriperumbudur BK (2018). Characteristic and universal tensor product kernels. *Journal of Machine Learning Research*, 18: 1–29.

Team RC (2024). Parallel Package. R package version 1.0.17.

Tune P, Roughan M, Haddadi H, Bonaventure O (2013). Internet traffic matrices: A primer. *Recent Advances in Networking*, 1: 1–56.

Wang C, Chen MH, Schifano E, Wu J, Yan J (2016). Statistical methods and computing for big data. *Statistics and its Interface*, 9: 399. https://doi.org/10.4310/SII.2016.v9.n4.a1

Wilks SS (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9: 60–62.

Wood J, Dykes J, Slingsby A (2010). Visualisation of origins, destinations and flows with od maps. *The Cartographic Journal*, 47: 117–129.

Zhang T, Yang B (2017). Box-Cox transformation in big data. *Technometrics*, 59: 189–201. https://doi.org/10.1080/00401706.2016.1156025