Journal of Sciences, Islamic Republic of Iran 35(3): 205 - 219 (2024) http://jsciences.ut.ac.ir
University of Tehran, ISSN 1016-1104

Bayesian Clustering of Spatially Varying Coefficients Zero-Inflated Survival Regression Models

S. Asadi, M. Mohammadzadeh*

Department of Statistics, Tarbiat Modares University, Tehran, Islamic Republic of Iran

Received: 15 October 2024 / Revised: 1 January 2025 / Accepted: 13 January 2025

Abstract

The study addresses the challenges of analyzing time-to-event data, particularly emphasizing the discrete nature of durations, such as the number of years until divorce. This frequently results in zero-inflated survival data characterized by a notable frequency of zero observations. To address this, the study employs the zero-inflated discrete Weibull regression (ZIDWR) model, which serves as a suitable framework for evaluating the impact of explanatory variables in survival analysis. However, challenges such as nonstationarity in the relationship between variables and responses and spatial heterogeneity across geographical regions can result in a model with too many parameters To mitigate this, we propose a spatial clustering approach to summarize the parameter space. This Paper leverages nonparametric Bayesian methods to explore the spatial heterogeneity of regression coefficients, focusing on the geographically weighted Chinese restaurant process (gwCRP) for clustering the parameters of the ZIDWR model. Through simulation studies, the gwCRP method outperforms unsupervised clustering algorithms clustering K-means and the standard Chinese restaurant process (CRP), exhibiting superior accuracy and computational efficiency, particularly in scenarios with imbalanced cluster sizes. This improved performance is quantitatively demonstrated through higher Rand indices, lower average mean squared error (AMSE) in parameter estimation and superior log pseudo-marginal likelihood (LPML) values. Applying this methodology to Iranian divorce data reveals distinct spatial clusters characterized by varying covariate effects on the probability of divorce within the first five years of marriage and the subsequent time to divorce.

Keywords: Survival Analysis; Varying Coefficient; Spatial Clustering.

Introduction

Survival analysis is a statistical technique used to evaluate time-to-event data. While survival time is generally treated as a continuous random variable, it is often recorded at discrete intervals (e.g., 0, 1, 2, 3…). This discretization may result in zero observations, indicating events that occurred before the first time recording unit (e.g., daily, monthly, or yearly). These zero values, sometimes referred to as "sampling zeros"(1), arise from events that take place right at the)commencement of the study. Such occurrences are

* Corresponding Author: Tel:+989122066712; Fax:+982182883483: Email:mohsen_m@modares.ac.ir

prevalent across various domains. For example, in the healthcare sector, pregnant women might spend less than a day in the hospital before delivery. Likewise, in studies related to job placement, a zero survival time may signify an immediate job placement. Traditional survival models often struggle to accommodate these instances. As a result, researchers have turned to "zero-inflated survival models" to tackle these challenges more effectively. Applications of these models include zero-inflated Cox models for analyzing rat sleep time following ethanol exposure (2), Weibull models for investigating time until banking fraud occurs (3-4), and zero-inflated cure models employed in studies of labor duration and cervical cancer (5-6). The choice of the baseline distribution is crucial in zero-inflated discrete models. While the Poisson distribution is frequently used for its intuitive interpretation of count data, its inherent assumption of equality between mean and variance often fails in practice. This limitation leads to over- or under-dispersion, resulting in inaccurate inferences and underestimated standard errors. Although the negative binomial distribution effectively addresses over-dispersion, it is unsuitable for under-dispersed data. Furthermore, by modeling the probability of a specific number of events within a defined period and assuming independence, the Poisson distribution is not directly analogous to time-to-event distributions. Consequently, generalizing discrete distributions in survival analysis is necessary to accommodate all types of dispersion and relax the independence assumption, mainly when dealing with correlated data. The Type I Discrete Weibull distribution proposed (7) is designed to mirror its continuous counterpart, is well-suited for discrete survival data and effectively handles both over- and under-dispersion. The Zero-Inflated Discrete Weibull (ZIDW) regression model is ideal for zero-inflated discrete survival data as it captures dispersion in zero and non-zero modes (8). This model includes two regression relationships:one for the effect of explanatory variables on the rate of non-zero responses and another for the probability of zero, allowing each explanatory variable to have two regression coefficients. Considering the spatial references of survival data, known as survival spatial analysis, enables the estimation and comparison of survival across different geographical areas, revealing spatial patterns. This helps identify areas with the highest and lowest survival rates.

One notable aspect of spatial variability is the difference in the influence of explanatory variables on survival time across different locations, a phenomenon known as spatial heterogeneity. Spatial heterogeneity refers to how the relationship between explanatory and response variables alters with geographical displacement. This variation arises because different locations exhibit different properties or values. Consequently, the values of regression coefficients can differ significantly from region to region. As a result, traditional regression models may fail to accurately capture the nature of these relationships in the context of spatial data analysis. Two main methods exist for estimating regression coefficients in models with spatially variable coefficients. The first is geographically weighted regression, a local method that estimates model parameters by weighting them at any point in the examined space. Unlike conventional regression, which describes general relationships between variables, geographically weighted regression provides spatial information on the variations in these relationships. The second method treats regression coefficients as random variables following spatial distributions. The spatial distribution can be assessed by selecting appropriate prior probability functions for the parameters. (9) examined the application of a geographically weighted regression model for accelerated failure time in spatial survival data. (10) investigated the influence of explanatory variables through a geographically weighted regression model on the Cox survival models, explicitly applying the Weibull distribution to handle the data.

The second method, Spatial Variable Coefficient (SVC), addresses this spatial heterogeneity by treating regression coefficients as spatial random variables. This method considers regression coefficients as spatial random variables that follow spatial distributions. By selecting suitable prior probability functions for the parameters, it is possible to assess the spatial variability of the parameters at different locations in the Bayesian spatially-varying coefficient (BSVC) model and to estimate the regression coefficients from their posterior probability (11). Simulation studies indicate that SVC processes outperform GWR by accurately estimating regression coefficients, so GWR must be considered a purely exploratory tool (12). In survival data analysis, the spatial variable coefficients (SVCs) method in the Cox model with a frequency-oriented perspective has been suggested by (13). (14) have suggested an AFT model with prior spatial distributions for spatially variable regression coefficients. (15) have suggested a geographically weighted Cox regression model for sparse spatial survival data.

Although these methods enhance survival prediction accuracy by considering the spatial variability of regression coefficients, they also raise model fitting complexity for ZIDWR models, assuming spatial variability of parameters for both regression coefficient vectors. This is because not only can the effect of explanatory variables on non-zero response have spatial

variability, but their effect on the probability of zeroing can also be different in different places. Thus, all regression coefficients have spatial variability.

Moreover, there are often censored observations in these data due to limited follow-up time that cannot be overlooked. Clustering the model parameters with similar spatial features is a proper method for efficiently reducing model dimensions and summarizing data. Spatial clustering methods, such as the K-means method, can be used to summarize data efficiently. Hence, Bayesian nonparametric processes, such as the Dirichlet mixture process, are used to investigate the spatial heterogeneity of regression coefficients (16). This process simultaneously considers intra-cluster correlation and heterogeneity between clusters in the spatial clustering structure. These processes turn the model into a simple parametric form by clustering with complex data on the model parameter space. Given its computational ease, the Dirichlet Process is one of the best random processes among nonparametric processes (for instance, the Gaussian process, Pólya tree process (17), and so on) for clustering parameters in models with SVCs. In this method, we can cluster coefficients into homogeneous groups by choosing prior probabilities on the distribution of discrete partitions, where several parameters get the same value simultaneously.

As a representation of the infinite mixture Dirichlet process, the Chinese Restaurant Process (CRP) introduced by (18,19) allows for dividing model parameters into homogeneous clusters without predetermined assumptions about the cluster count. As our purpose of spatial clustering of parameters is to reduce the spatial heterogeneity in the data, it is necessary to consider the geographical location of the units in the allocation of clusters because the presence of common factors in close areas causes the parameters in them to be similar. In other words, each member's allocation within each cluster will be such that if that member is closer to the other cluster members regarding geographical distance, it has a better chance of being in that cluster. Hence, the distance between regions has a significant role in this clustering. (20) offers a compelling alternative:the distance-dependent Chinese restaurant process (ddCRP). This model directly incorporates the probability of assigning data points to existing clusters, making the assignment dependent on the distance between data points. An excellent way to do this is to make this function one of the weighting functions in geographically weighted regression models. Recently (21) introduced a Geographically Weighted Chinese Restaurant Process (gwCRP) to analyze the spatial heterogeneity of regression coefficients. This method simultaneously considers intra-cluster correlation and spatial clustering structure heterogeneity and estimates the number of clusters using a nonparametric Bayesian approach. While recent studies have explored the spatial heterogeneity of regression coefficients in count data models (22) and zero-inflated models (23), the spatial clustering of coefficients within survival models incorporating both zero-inflated and right-censoring remains an uncharted area of research.

Here, we demonstrate the adaptability of the geographically weighted Chinese restaurant process (gwCRP) clustering method for zero-inflated and right-censored survival models and show that compared to traditional CRP and k-means methods, gwCRP consistently estimates the number of clusters regarding distances while maintaining precise parameter estimation of each component of our two-part generalized linear regression model. To our knowledge, we are the first to introduce the spatial varying coefficients in the ZIDW regression model. Finally, we demonstrate how a Zero-Inflated Discrete Weibull (ZIDW) model, incorporating covariates such as the husband's employment status, wife's financial autonomy, age gap, and spousal similarity, could best fit the data. By spatial analysis, we will reveal significant regional variations in the effects of these covariates on both the probability of early divorce and the duration of marriage when divorce occurs later. Our novel approach, combining survival analysis with spatial clustering, provides a more nuanced understanding of divorce than traditional CRP and k-mean methods and offers valuable insights for targeted policy interventions.

The remainder of the paper is organized as follows. Section 2 summarizes ZIDW regression models. Section 3 defines the variability of SVC regression coefficients in ZIDW survival data and provides an overview of CRP and gwCRP methods. Section 4 presents the Bayesian analysis with a Gibbs sampling algorithm for clustering parameters of the ZIDW model with spatial variability of regression coefficients. Section 5 compares the existing methods in a simulation study. Then, numerical results on divorce data are presented in Section 6.

Materials and Methods

Let the random variable T has a discrete Weibull distribution $T \sim DW(q, \beta)$ with probability mass function $f(t) = P(T = t) = q^{t^\beta} - q^{(t+1)^\beta}, t = 0,1,2, \ldots$ One uses the discrete Weibull regression model with some link functions of the parameters q or β to consider the effects of some covariates on T. To define a ZIDW regression model, let the survival time T be a non-negative random count variable with the probability mass function

$$P(T=t \mid X,Z) = \begin{cases} p(Z)+(1-p(Z))(1-q(X)), & t=0 \\ (1-p(Z))\left(q(X)^{t^{\beta}}-q(X)^{(t+1)^{\beta}}\right), & t=1,2,\dots \end{cases}$$

denoting by $T \mid X,Z \sim \text{ZIDW}(p(Z),q(X),\beta)$, where the parameters $q \equiv q(X)$ and $p \equiv p(Z)$ depend on the covariates $X_{n\times(m_1+1)} = \left(1,X_1,\dots,X_{m_1}\right)$ and $Z_{n\times(m_2+1)} = \left(1,Z_1,\dots,Z_{m_2}\right)$, respectively, through the link functions (24):

$$\log\left(-\log(q(X))\right) = X'\boldsymbol{\alpha},\ \Rightarrow\ q \equiv q(X) = e^{-e^{X'\alpha}} \quad (1)$$

$$\text{logit}(p(Z)) = Z'\boldsymbol{\gamma},\ \Rightarrow\ p \equiv p(Z) = \frac{e^{Z'\gamma}}{1+e^{Z'\gamma}} = \left(1+e^{-Z'\gamma}\right)^{-1} \quad (2)$$

where $\alpha = \left(\alpha_0,\dots,\alpha_{m_1}\right)$ and $\boldsymbol{\gamma} = \left(\gamma_0,\dots,\gamma_{m_2}\right)$ are the vectors of regression coefficients. The ZIDW regression models assume that the effect of explanatory variables on the response variable is the same in different places. However, other conditions in each region may cause spatial heterogeneity. Here, we consider the spatial variability for all regression coefficients. Let $T_{\ell i}$, for $i=1,\dots,n$, and $\ell=1,\dots,n_i$ denote the survival time for the case ℓ at site $s_i = (u_i,v_i)$, n_i denotes the number of subjects at site s_i, and $X_\ell(s_i), Z_\ell(s_i)$ are the vectors of covariates. Let $T_{\ell i} \mid X,Z \sim ZIDW(p_{\ell i}(Z),q_{\ell i}(X),\beta)$, then the equations (1) and (2) considering the spatial variability of regression coefficients $\alpha_{\ell i}$ and $\gamma_{\ell i}$ will be as follows:

$$p_{\ell i}(Z) = \frac{e^{Z_\ell(s_i)\gamma(s_i)}}{1+e^{Z_\ell(s_i)\gamma(s_i)}},\ q_{\ell i}(X) = e^{-e^{X_\ell(s_i)\alpha(s_i)}}$$

Where $\gamma(s_i) = \left(\gamma_0(s_i),\dots,\gamma_p(s_i)\right)$ and $\boldsymbol{\alpha}(s_i) = \left(\alpha_0(s_i),\dots,\alpha_p(s_i)\right)$ are the model components that can be estimated by fitting two separate models. So

$$P\left(T_{i\ell}=t \middle| X,Z,\beta\right) = \begin{cases} \frac{1}{1+e^{Z_\ell(s_i)\gamma(s_i)}}\left[1+e^{Z_\ell(s_i)\gamma(s_i)}-e^{-e^{X_\ell(s_i)\alpha(s_i)}}\right] & t=0 \\ \frac{1}{1+e^{Z_\ell(s_i)\gamma(s_i)}}\left[\left(e^{-e^{X_\ell(s_i)}}\right)^{t^{\beta}}-\left(e^{-e^{X_\ell(s_i)}}\right)^{(t+1)^{\beta}}\right] & t=1,2\dots \end{cases}$$

Zero-Inflated Discrete Weibull (CZIDW) model, if $T_{\ell i}$ is the survival time of the ℓ-th unit and $C_{\ell i}$ the censored from the right that is independent of $T_{\ell i}$, then for a censored unit, the only available information is $C_{\ell i} < T_{\ell i}$. By defining $Y_{\ell i} = \min(T_{\ell i}, C_{\ell i})$, $\delta_{i\ell} = 1$ if $T_{\ell i} \geq C_{\ell i}$ and $J_{\ell i} = 1$, if $Y_{\ell i} = 0$, we can divide all the data, $\boldsymbol{D} = \left\{\left(T_{\ell i},\delta_{\ell i},X_\ell(s_i)\right), i=1,\dots,n,\ell=1,\dots,n_i\right\}$, as follows

$$\begin{cases} J_{\ell i}=1,\delta_{\ell i}=0 & Y_{\ell i} \text{ is zero and not right-censored} \\ J_{\ell i}=0,\delta_{\ell i}=0 & Y_{\ell i} \text{ is non-zero and not right-censored} \\ J_{\ell i}=0,\delta_{\ell i}=1 & Y_{\ell i} \text{ is non-zero and right-censored} \end{cases}$$

In this case, the likelihood of the CZIDW model can be defined as follows

$$L(\beta,\boldsymbol{\alpha},\gamma \mid n,Y,X,Z) = \Pi_{i=1}^{n}\Pi_{\ell=1}^{n_i}[F_{\ell i}+(1-F_{\ell i})(1-G_{\ell i})]^{J_{\ell i}(1-\delta_{\ell i})} \times\left[(1-F_{\ell i})\left(G_{\ell i}^{Y_{\ell i}^{\beta}}-G_{\ell i}^{(Y_{\ell i}+1)^{\beta}}\right)\right]^{(1-J_{\ell i})(1-\delta_{\ell i})}\left[1-F_{\ell i}-(1-F_{\ell i})\left(1-G_{\ell i}^{C_{\ell i}^{\beta}}\right)\right]^{\delta_{\ell i}} \quad (3)$$

where $F_{\ell i} = \left(1+e^{-Z_\ell(s_i)\gamma(s_i)}\right)^{-1}$ and $G_{\ell i} = e^{-e^{X_\ell(s_i)\alpha(s_i)}}$.

1. Clustering of Model Coefficients

For each particular location $s_i, i=1,\dots,n$, we define $\boldsymbol{\theta}(s_i) = (\boldsymbol{\alpha}(s_i)^{\top},\boldsymbol{\gamma}(s_i)^{\top})^{\top}$ the collection of parameters. CRP assumes n customers enter a Chinese restaurant with unlimited tables (5). In our setting, we assume that the n parameter vectors can be clustered into k groups, i.e., $\boldsymbol{\theta}(s_i) = \boldsymbol{\theta}_{\lambda_i} \in \{\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_k\}$, where $\lambda_i \in \{1,\dots,k\}$, with k being the total number of clusters. One popular way to model the joint distribution of $\boldsymbol{\lambda} = (\lambda_1,\dots,\lambda_k)$ is the CRP, which is an essential representation of the Dirichlet process and defines a series of conditional distributions as

$$P(\lambda_i = c \mid \boldsymbol{\lambda}_{-i}) \propto \begin{cases} \frac{n_{i,c}}{\alpha^*+i-1} & \text{existing cluster} \\ \frac{\alpha^*}{\alpha^*+i-1} & \text{new cluster} \end{cases} \quad (4)$$

where $\boldsymbol{\lambda}_{-i} = (\lambda_1,\dots,\lambda_{i-1})$ and $n_{i,c}$ is the number of elements in cluster c, and α^* is the concentration parameter of the underlying Dirichlet process. Equation (4) expresses the conditional probability of placing the i^{th} unit in the c^{th} cluster, given that the $i-1$ of the previous unit is clustered. (15,20) introduced the "geographically weighted Chinese Restaurant Process" (gwCRP) clustering method based on the weight functions of distances. So in equation (3), we have $n_{i,c} = \sum_{j=1}^{i-1} w_{ij} I(\lambda_j = c)$, where w_{ij} s are elements of the weight matrix W. Spatial weights are accommodated using a Stochastic Neighborhood Conditional Autoregressive (SNCAR) model (25), extending the conventional Conditional Autoregressive (CAR) model (26) to account for areal data. (27) defined a weight matrix based on graph distance. Assume that the whole area we are considering is a graph A with a set of vertices $V(A) = \{v_1,\dots,v_n\}$ and a set of edges $E(A) = \{e_1,\dots,e_m\}$ then the matrix elements are $w_{ij} = 1$ if $d_{v_iv_j} \leq 1$, otherwise $w_{ij} = \exp\left(-\frac{d_{v_iv_j}}{h}\right)$ where

$$d_{v_iv_j} = \begin{cases} |V(e)|, & \text{if } e \text{ is the shortest path connecting } v_i \text{ and } v_j \\ \infty, & \text{if } v_i \text{ and } v_j \text{ are not connected} \end{cases}$$

is the distance graph between A_i and A_j, and h is the bandwidth (28). Moreover, $|V(e)|$ is the cardinality of the $V(e)$ set, where e is the shortest path to the two vertices. It is evident that when $h=0$, the suggested gwCRP technique is identical to the traditional CRP technique. In this particular situation, the CRP technique tends to cluster excessively. Another significant pattern is that as h rises, the estimated number of clusters decreases before rising again. Simultaneously, the Rand

index demonstrates an initial increase followed by a decrease as h becomes excessively large. This pattern emerges because, starting from $h = 0$, the gwCRP technique effectively begins to capture the inherent spatial relationships in the data. Nevertheless, as $h \to \infty$, the geographic weights w_{ij} for spatial-discontinuous areas decrease to zero. As a result, only neighboring areas are categorized within the identical cluster, bringing back the problem of excessive clustering.

2. Bayesian Analysis

Suppose for the CZIDW model for the set of parameters $\Theta = (\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\pi}, k)$, we have separated the model parameters by to $k \leq n$. In that case, we expect that each member of the parameter space $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$ where $\boldsymbol{\theta}(s_i) = \boldsymbol{\theta}_{\lambda_i}$ is equal to one of the k separate values of the separation set $\boldsymbol{\theta}_1^*, \dots, \boldsymbol{\theta}_{K^*}^*$ If K^* denotes the number of clusters excluding the i-th observation $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{i-1}$. Thus, if G_0 is a continuous distribution Polya Urn scheme, the conditional distribution of θ_i given $\boldsymbol{\theta}_{-i} = \{\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_n\}$ will be as follows:

$$P(\theta_i \mid \boldsymbol{\theta}_{(-i)}, \alpha^*, G_0) \propto \begin{cases} \frac{1}{\alpha^* + i - 1} \sum_{k^*=1}^{K^*} \sum_{j=1}^{i-1} w_{ij}^* I(\boldsymbol{\theta}(s_j) = \boldsymbol{\theta}_{k^*}) \delta_{\theta_k^*}(\boldsymbol{\theta}(s_i)) & \text{existing cluster} \\ \frac{\alpha^* G_0(\boldsymbol{\theta}(s_i))}{\alpha^* + i - 1} & \text{new cluster.} \end{cases}$$

Where $\delta(\cdot)$ is the indicator function. Then by defining Prior hierarchically as follows:

$$T \mid X, Z, U \sim ZIDW\big(p_{\lambda_i}(Z), q_{\lambda_i}(X), \beta\big),\ i = 1, \dots, n,$$
$$\boldsymbol{\alpha}_h \sim \mathrm{N}(\mathbf{0}, \Sigma_\alpha),\ \gamma_h \sim \mathrm{N}\big(\mathbf{0}, \Sigma_\gamma\big),\ h = 1, \dots, k,$$
$$G_0(\alpha, \gamma) \propto P(\alpha)P(\gamma) = MVN(0, \Sigma_0),$$
$$\lambda_i \mid \boldsymbol{\pi}, k \sim \mathrm{Multinomial}(\pi_1, \cdots, \pi_k),$$
$$\boldsymbol{\pi} \sim \mathrm{gwCRP}(\alpha^*, h),\ k \sim P(\cdot).$$

For data $D = (Y, X, Z, J, \delta)$, with $L(\Theta \mid D)$, our goal is to sample from the posterior distribution of the parameters $k, \boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_n) \in \{1, \dots, k\}, \boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_k)$, and $\gamma = (\gamma_1, \dots, \gamma_k)$. In nonparametric Bayesian models with the prior Dirichlet Processes (8), due to the unavailability of the analytical form for the posterior distribution of Θ, we employ the Gibbs sampling (27) to repeatedly draw values for each θ_i from its conditional distribution given both the data and the θ_j for $j \neq i$. Then, we combine this result with the likelihood and derive the full conditional distribution for θ_i for use in Gibbs sampling:

$$\theta_i \mid \theta_{-i}, Y \sim Q\Big[\sum_{i \neq r} L(\theta_i \mid n, Y, X_1, X_2)\delta_{\theta_r}(\theta_i) + \alpha^*\Big(\int L(\theta_i \mid n, Y, X_1, X_2) dG_0(\theta)\Big) H_i(\theta_i)\Big],$$

Where Q is the normalizer constant, $H_i(\theta)$ is the posterior distribution of θ obtained by combining information from the prior distribution G_0 and observed data D_i.

3. Cluster Configurations

Using Dahl's method introduced by (28) allows for obtaining posterior estimates of cluster memberships $\lambda_1, \dots, \lambda_n$ and other model parameters γ and α. This method selects an "average" clustering using all posterior clusterings in the three below steps:

Step 1. Define membership matrices $\mathcal{A}^{(b)} = \big(\mathcal{A}^{(b)}(i,j)\big)_{i,j \in \{1,\dots,n\}} = \Big(I\big(\lambda_i^{(b)} = \lambda_j^{(b)}\big)\Big)_{n \times n}$, where $b = 1, \dots, B$ is the index for the retained MCMC draws after burn-in, and $I(\cdot)$ is the indicator function.

Step 2. Calculate the element-wise mean of the membership matrices over MCMC draws $\overline{\mathcal{A}} = \frac{1}{B}\sum_{b=1}^{B} \mathcal{A}^{(b)}$.

Step 3. Identify the most representative posterior $\overline{\mathcal{A}}$ draw based on minimizing the element-wise Euclidean distance $\sum_{i=1}^{n}\sum_{j=1}^{n}\big(\mathcal{A}^{(b)}(i,j) - \overline{\mathcal{A}}(i,j)\big)^2$ among the retained $b = 1, \dots, B$ posterior draws.

The algorithm accuracy can be evaluated using the Rand index (29) for comparing cluster configurations obtained with different methods to the actual clusters. The Rand index computes a similarity measure between two clusterings by considering all sample pairs and counting pairs assigned in the same or different clusters in the predicted and true clusterings. This index allows us to measure the similarity between different clustering results, providing valuable insights into the match ability of these configurations. To measure the agreement between $\boldsymbol{\lambda}^{(C_{LS})}$ and the true clustering configuration. The Rand index of two partitions, $\mathcal{S}_1 = \{U_1, \dots, U_r\}$ and $\mathcal{S}_2 = \{V_1, \dots, V_s\}$, of a set of n objects $S = \{o_1, \dots, o_n\}$, is defined as

$$RI = \frac{a + b}{a + b + c + d} = \frac{a + b}{\binom{n}{2}}$$

where a represents the number of pairs of objects in set S that are in the same cluster in $\mathcal{S}_1$ and the same cluster in $\mathcal{S}_2$, b represents the number of pairs of objects in set S that are in different clusters in $\mathcal{S}_1$ and different clusters in $\mathcal{S}_2$, c represents the number of pairs of objects in set S that are in the same cluster in $\mathcal{S}_1$ and different clusters in $\mathcal{S}_2$ and d represents the number of pairs of objects in set S that are in different clusters in $\mathcal{S}_1$ and the same cluster in $\mathcal{S}_2$. The Rand index varies from 0 to 1, where a higher value signifies more excellent agreement between the two partitions. When the partitions are in complete agreement, the Rand index equals 1.

For model selection, the decaying effect parameter h for geographical weights needs to be tuned, and we use the logarithm of the Pseudo-Marginal Likelihood (30) based on conditional predictive ordinate to select h. The LPML is defined as LPML $\sum_{i=1}^{N} \log(\mathrm{CPO}_i)$, where CPO_i is the $i - the$ conditional predictive ordinate. The

Table 1. Comparison of LPML for different h values in both scenarios

	h-values				
Scenario	1	1.6	2	2.6	3
Balanced	-22402	-155942	-12658	-8360	-9653
Imbalanced	-20188	-21070	-13181	-5133	-6671

Monte Carlo estimate of the CPO, within the Bayesian framework, can be obtained as $\widehat{\mathrm{CPO}_i}^{-1} = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{f\left(D_i|\boldsymbol{\theta}_{\lambda_i}^b\right)}$, where B is the total number of Monte Carlo iterations, $\boldsymbol{\theta}_{\lambda_i}^b$ is the $b-the$ posterior sample, and $f(\cdot)$ is the likelihood function defined in (3). An estimate of the LPML can subsequently be calculated as $\widehat{\mathrm{LPML}} = \sum_{i=1}^{N}\log\left(\widehat{\mathrm{CPO}_i}\right)$. A model with a more considerable LPML value is preferred.

4. Simulation Study

A simulation study compares the K-means and the CRP clustering methods with the proposed gwCRP clustering for zero-inflated discrete time-to-event data with spatially varying covariates. The study will examine two balanced and imbalanced scenarios for data geographical clustering patterns. Under the balanced scenario, each group contains an equal number of units. Under the imbalanced scenario, the group sizes differ, and we have three to four clusters over two scenarios. The number of sites is set to the number of provinces in Iran, i.e., 31. We then generated a sample of size $n_i = 5$ for each province with center s_i, so the total number of observations is $n = 155$. We assumed X is equivalent to Z and a similar set of covariates affect q and p parameters. Then, we generated spatial covariates $X_\ell(s_i)$ from Normal distribution $N(0,1)$. The temporal component pdf for the $\ell^{th}, \ell = 1, \dots, n_i$ observation in province s_i, follows the distribution $T_{\ell i} \mid X_{\ell i} \sim \mathrm{ZIDW}(p(X_{\ell i}), q(X_{\ell i}), \beta)$, with a fixed value of $\beta = 1.2$. So, two related responses were controlled under two generalized linear models, $\mathrm{logit}(p)$ and $\log(-\log(q))$. We set initial values for model coefficient parameter α_{real}, $(-2,0.5)$, $(1.5,0.6)$, $(2.1,-0.4)$, $(1.1,0.3)$, and for γ_{real}, $(0.95,1.1)$, $(-0.4,0.6)$, $(0.5,0.8)$, $(1,1.5)$ corresponding to each of the 4 clusters, respectively. Then, to investigate the right censoring, we considered the quantile 93% of data as the censored point $C_{i\ell}$ and as a threshold to cut the simulated sample, such that all values $y_{\ell i} \geq C_{\ell i}$ were re-valued to be equal to $C_{\ell i}$. Also, if $T_{\ell i}$ is not greater than the generated censored time $C_{i\ell}$, we set $\delta_{\ell i} = 1$, otherwise, it is considered zero. To add a zero-inflated feature for each response, first, a random vector from a uniform distribution $U = (u_1, \dots, u_n) \sim U(0,1)$ is generated if $u_{\ell i} \leq p_{\ell i}$, set $J_{\ell i} = 0$ and $Y_{\ell i} = 0$ otherwise, we considered $J_{\ell i} = 1$ and generated $Y_{\ell i}$ from DW distribution. We generated the outcome data under the following two generalized linear models

$$\mathrm{logit}(p_{i\ell}) = \gamma_{0\ell}(s_i) + \gamma_{1\ell}x_\ell, \quad (5)$$

$$\log(-\log(q_{i\ell})) = \alpha_{0\ell}(s_i) + \alpha_{1\ell}(s_i)x_\ell \quad (6)$$

We used Normal prior distributions $N(0,\sigma_\alpha^2)$ for regression coefficients α_0 and α_1, with precision parameters, $\sigma_\alpha^{-2} \sim T(10^{-5}, 10^{-5})$. Similarly, for γ_0 and γ_1, the Normal priors $N\left(0,\sigma_\gamma^2\right)$, are considered, respectively, with $\sigma_\gamma^{-2} \sim T(10^{-5}, 10^{-5})$. To assess gwCRP's clustering performance across a range of h values, we will evaluate it from 1 to 3 in a grid of 0.2. The optimal value of h will be determined using LPML (Table 1). We fixed the concentration parameter $\alpha^* = 1$. We provide information on estimating the number of clusters and the compatibility of clustering configurations. The maximum distance in the spatial structure of the 31 regions is 10k. m, so yielding an optimal bandwidth $\left(h_{\text{opt}} = 2.6\right)$ induces a weighting scheme that ensures relative weights are assigned appropriately. Each replicate involves running an MCMC chain of length 10,000 with a thin of one and burn-in of 2,000 samples.

Results

After meticulously examining the MCMC chain length, we run our proposed algorithm in 100 separate data replicates. A vital part of this process is obtaining 100 RI values, which we then compare with the real values to validate our results. We calculated the mean in the 100 replicates and the posterior means of the parameters. Each replicate runs a total of MCMC iterations. We calculated the cover rate for each scenario, which equals the percentage of replicates in which our proposed algorithm accurately recovers the number of clusters. In our gwCRP model for scenario 2, we observe that the correct number of clusters is inferred in at least 25 out of 100 instances. Specifically, for model 1 under scenario 2, the final estimate of the number of clusters consistently reaches five across 90 replicates. However, in scenario 1, 75 cases underestimate the number of clusters by 10. We also provide a detailed comparison of our method with the K-means Algorithm. As the K-means algorithm cannot infer the number of clusters,

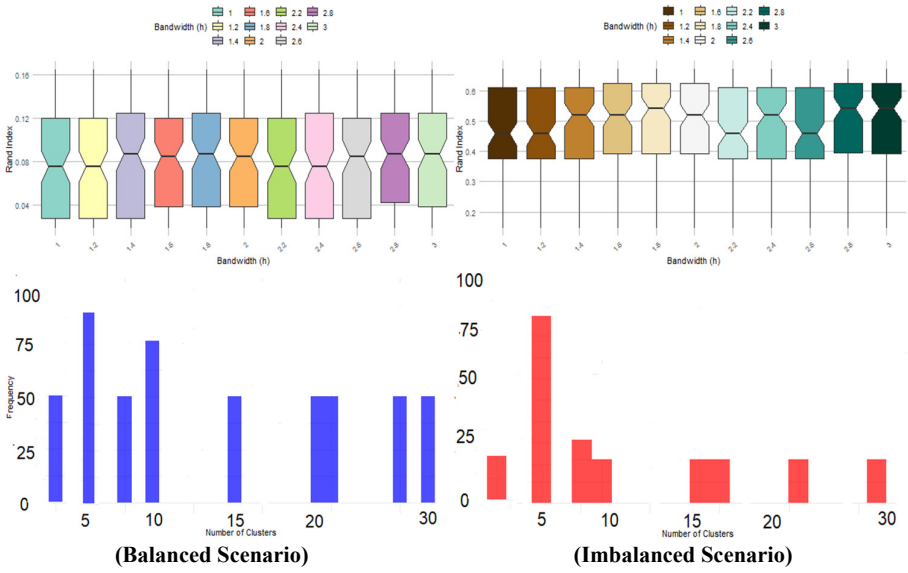


Figure 1. Histogram of estimates of k under h-optimal and box plot of Rand index under different h and LPML selection.

Table 2. RI indexes for different clustering methods ($h_opt = 2.6$)

Type	CRP	K-means	gwCRP
Balanced	0.8931	0.6865	0.9344
Imbalanced	0.8122	0.7624	0.9581

such values must be pre-specified. We supplied them with the number of clusters inferred by our method in each replicate, providing a comprehensive understanding of their differences. We also present the histogram of the final number of clusters inferred for each cluster scenario and data generation model combination in (Figure 1).

As the Gibbs sampler does not directly yield the posterior distribution of k, we employed Dahl's method to estimate it. The RIs for each scenario and data generation model are reported in Table 2. We also thoroughly compared our method to the K means algorithm. As the K-means algorithm cannot infer the number of clusters, such values need to be pre-specified, and we supplied them with the number of clusters inferred by our method in each replicate. As the Gibbs sampler does not directly yield the posterior distribution of k, we employed Dahl's method to estimate it. Table 2 demonstrates the significant improvement in computational efficiency offered by our proposed gwCRP model with vectorization for both scenarios. This enhancement, coupled with the model's highest RI, underscores its innovative approach and high accuracy in clustering. The K-means model, while having an RI greater than 0.6, does not match the performance of our proposed model. Furthermore, using the optimal value of h determined by LPML has resulted in excellent clustering performance. In addition to assessing clustering performance, we also evaluate the estimation performance of covariate coefficients.

Let $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_n)$ be the actual clustering label vector, $\boldsymbol{\theta}_r(s_i)$ be the true parameter value of cluster $j, \kappa_r = \sum_{i=1}^{31} I(\lambda_i = r)$ be the number of provinces in cluster r (where $r = 1, \dots, k$ and $\sum_{r=1}^{k} \kappa_r = n$). For the simulated dataset t, let $\widehat{\boldsymbol{\theta}}_{(t)}(s_i)$ be the estimate of Dahl's method at location s_i. Then, the average of mean squared error (AMSE) is calculated as

$$\text{AMSE} = \frac{1}{k}\sum_{r=1}^{k}\frac{1}{\kappa_r}\sum_{i|\lambda_i=r}\frac{1}{100}\sum_{b=1}^{100}\left(\widehat{\boldsymbol{\theta}}_{(b)}(s_i) - \boldsymbol{\theta}_r(s_i)\right)^2$$

Which calculates mean squared errors for each cluster first and then averages across clusters. Table 3 presents the AMSE results for parameter estimation of gwCRP using optimal values of h in two different scenarios. Table 3 presents the AMSE results for parameter estimation of gwCRP using optimal values of h in two different scenarios. Generally, the K-mean method has a higher AMSE than other methods. Our research Identifies a pattern in clustering performance, showing that gwCRP exhibits a lower AMSE than traditional

Table 3. Performance of parameter estimates under the two true cluster scenarios with AMSE (h = 2.6)

	Balanced				Imbalanced			
Method	α_0	α_1	γ_0	γ_1	α_0	α_1	γ_0	γ_1
gwCRP	0.0115	0.0266	0.0029	0.0350	0.0023	0.0023	0.0143	0.0009
CRP	0.0268	0.0214	0.1059	0.0901	0.0137	0.0319	0.0269	0.0401
K-mean	0.1810	0.0815	0.0297	0.3112	0.0997	0.1069	0.1018	0.1704

CRP. This result indicates the importance of selecting the optimal h based on LPML for accurate estimation. The AMSE fluctuates more in the balanced scenario than the imbalanced scenario;in this scenario, AMSE values are lower overall due to being mis-clustered.

In conclusion, our simulation studies clearly show that the gwCRP models outperform the standard CRP models in terms of clustering accuracy and parameter estimation. Our proposed model selection criterion, the LPML, effectively identifies the optimal h value, yielding superior results for clustering and parameter estimation tasks. These conclusions should convince the audience of the strength of our research findings.

The computational costs of our different clustering methods vary significantly. K-means has a time complexity of 37,200 units and is faster when the number of clusters is pre-defined, but it cannot automatically determine the optimal number of clusters. The Chinese Restaurant Process (CRP) has a more complex time complexity of around 24,025 units due to its iterative evaluation of potential cluster assignments, making it less efficient for larger datasets. In contrast, the proposed method, gwCRP, utilizes vectorization and optimized techniques, achieving a time complexity of approximately 930 units per iteration, leading to faster convergence. Additionally, the use of the LPML criterion helps identify the optimal value for parameter (h), further enhancing efficiency. In conclusion, while K-means is computationally efficient for fixed clusters but lacks flexibility, CRP is more adaptable but computationally intensive. The gwCRP method offers a balance of robust clustering performance and improved efficiency. Simulation studies confirm that gwCRP outperforms standard CRP in clustering accuracy and parameter estimation, with carefully designed parameters reflecting realistic scenarios in geographical data, highlighting the strengths of the proposed research.

1. Analysis of Divorce Data

Understanding the dissolution of marriages is crucial in addressing the social issue of divorce through Survival analysis. Recent studies show a worrying inflation of divorce in the first five years of marriage. To further investigate, we have partitioned the time axis into six 5-year periods, $[0,5), [5,10), \ldots, [30,35)$. The starting points of these intervals, namely $0,1, \ldots,6$, define the discrete survival times. Fifty couples who had experienced one or more marriages between 1989 and 2019 were selected from each of Iran's 31 provinces. The final dataset comprised 1,550 couples, of which 874 had experienced divorce. Other couples who did not experience divorce by the end of 2019 were considered

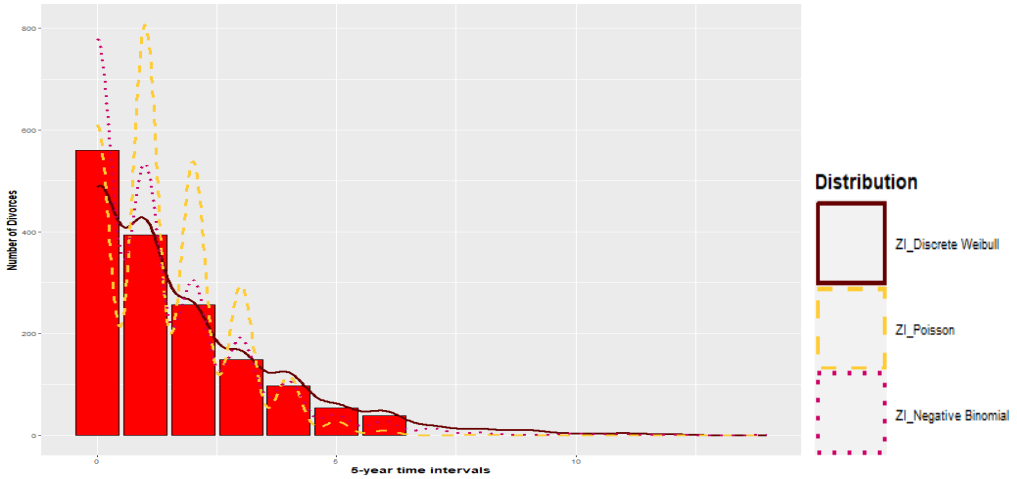


Figure 2. Histogram and Zero-inflated distributions of marriage duration among couples.

Table 4. Demographic characteristics.

Variable	Group	Number	Percent
Husband's employment status	Fixed	514	8.3
	Temporary	907	58.5
	Unemployed	129	33.2
Wife's Financial Autonomy	Independent	1128	72.8
	Dependent	422	27.2
Age gap	Less than fifteen years	1128	72.8
	More than fifteen years	422	27.2
Similarity	No	1283	82.8
	Yes	267	17.2

right-censored data. Approximately %36 of divorces occurred within the initial five years of marriage. Consequently, it is imperative to employ censored zero-inflated discrete distributions to model the data. The dispersion index represents the ratio of the observed variance from the data to the observed mean. In this case, the dispersion index equals 1.69, indicating over-dispersion in the data. Three distributions, namely Zero-Inflated Discrete Weibull (ZIDW), zero-inflated negative Binomial, and zero-inflated Poisson, have been fitted to the time data to reach the divorce event. Based on the observation in (Figure 2), it is evident that the ZIDW is better suited for these data than the other two distributions.

Due to the multidimensional nature of the divorce issue and the existence of various economic, social, cultural, demographic, etc. factors influencing the risk of divorce during the marriage period and also the probability of divorce less than five years, the demographic information of people such as the age difference of spouses, employment status are included in the CZIDW model as auxiliary variables according to (Table 4). In this study, we also examine the effect of spousal similarity.

To fit the distribution $T_{\ell i} \mid X, Z(s_i) \sim ZIDW\big(p_{\ell i}\big(X(s_i)\big), q_{\ell i}\big(X(s_i)\big), \beta\big)$ to the data, first, it is necessary to build the $5 \times n$ scenario matrix $X = (1, X_1, \dots, X_4)$, including the covariates "Husband's employment status" X_1, "Similarity" X_2, "Age gap" X_3, and "Wife's Financial Autonomy" X_4. Then we have:

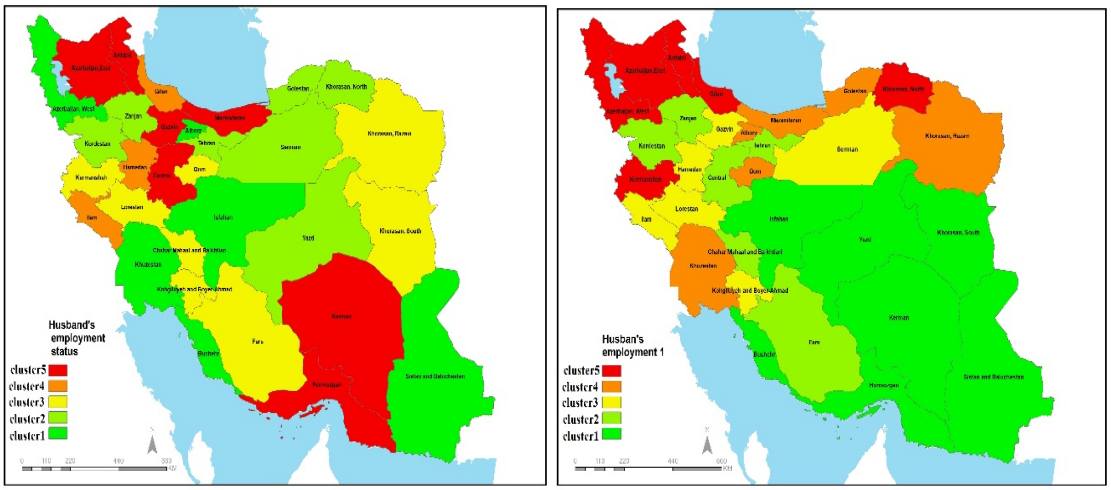
$$\log\left(-\log\left(q_{i\ell}\right) = \alpha_{0\ell}(s_i) + \sum_{m=1}^{4} \alpha_{m\ell}(s_i) x_{m\ell}(s_i)\right. \quad (7)$$

$$\operatorname{logit}(p_{i\ell}) = \gamma_{0\ell}(s_i) + \sum_{m=1}^{4} \gamma_{m\ell}(s_i) x_{m\ell}(s_i) \quad . \quad (8)$$

We first fit the two-part ZIDW model for each area using the covariates selected. Before being visualized, the covariates are adjusted to have a mean of 0 and a standard deviation of 1. According to the geographical patterns specified in (Figure 2-5) for each of the four covariates in both models, the probability of divorce in less than five years (zeroing the marriage survival time) and the duration of cohabitation provided that the couple has lived together for at least five years (non-zero count values), emphasizes the necessity of using SVC model. Also, it is seen that some provinces have similar characteristics, not limited to only adjacent counties, indicating possibilities of globally discontiguous clusters. In more detail, (Figure 3) shows significant spatial variation in divorce rates across Iranian provinces, strongly influenced by the husband's employment status (temporary, permanent, or unemployed). This variation reflects substantial socioeconomic disparities, including unemployment rates, job security, access to social services, and cultural and religious factors. These factors affect the relationship between a husband's employment and divorce probability, leading to stronger associations in some provinces than others. This is demonstrated by the varying regression coefficients for the husband's employment status across the country, as mapped in (Figure 3) for both models (6 and 7).

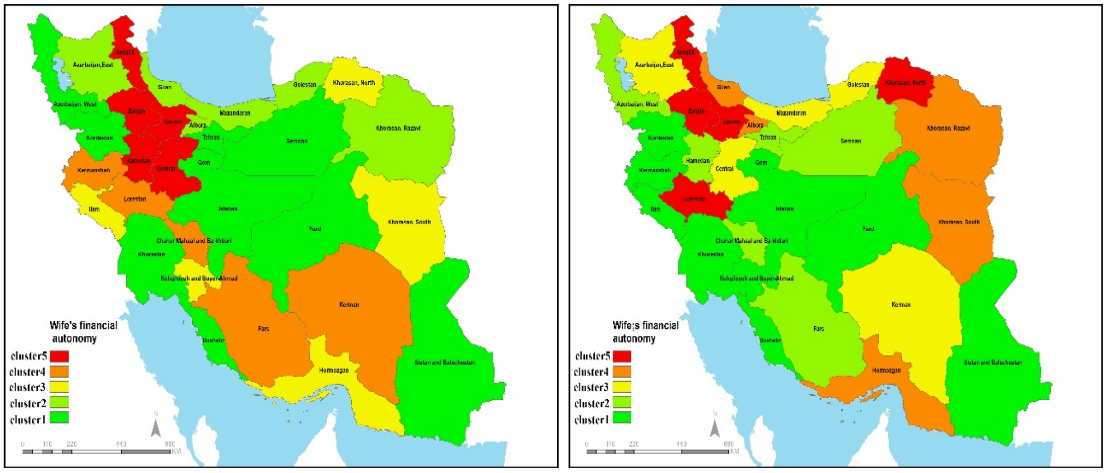
Additionally, according to spatial disparities shown in (Figure 4), the regression coefficient for both models in (7) and (8) for the wife's financial autonomy covariate in Iran is expected to vary spatially due to significant regional differences in socioeconomic development and cultural norms. More developed provinces with higher female education and employment may show weaker links between financial autonomy and divorce, while less developed, more conservative regions might exhibit stronger negative correlations, reflecting societal pressures and differing views on gender roles and responsibilities.

Moreover, As shown in (Figure 5) the impact of spousal similarity (education, socioeconomic status,



(a) (b)

Figure 3. The spatial varying covariate effects of Husband's employment on the two-part ZIDW model of provinces in Iran a:the probability of marriage survival becoming zero, b:the duration of time to divorce.



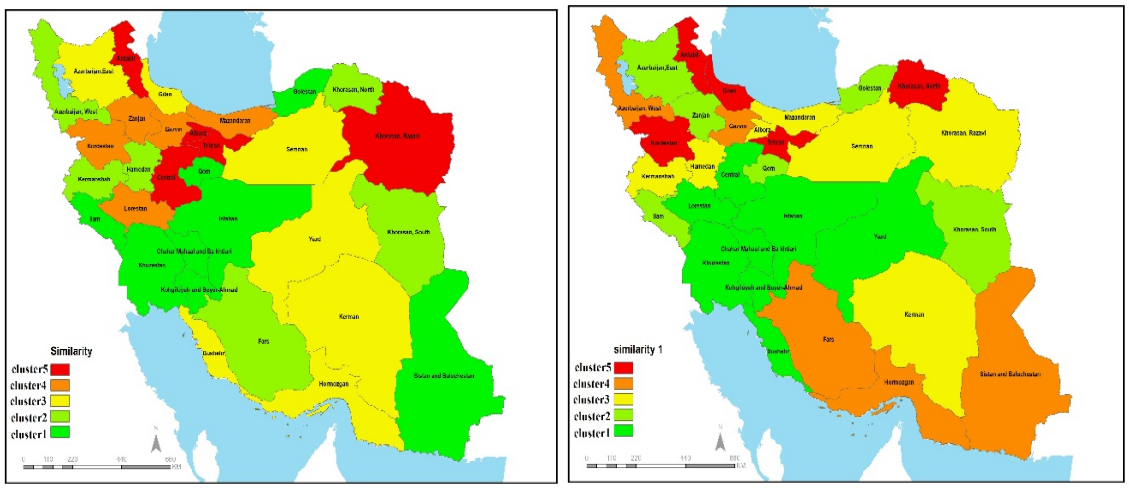
(a) (b)

Figure 4. The spatial varying covariate effects of Wife's Financial Autonomy on the two-part ZIDW model of provinces in Iran a:the probability of marriage survival becoming zero, b:the duration of time to divorce.

religious observance, ethnicity, and attitudes/personality) on time until the divorce event occurs in Iran varies significantly across provinces. For example, educational similarity is greater in provinces with higher literacy rates, while socioeconomic disparity's negative impact is stronger in provinces with high-income inequality. Similarly, religious similarity matters more in religiously conservative provinces, and ethnic similarity is more impactful in ethnically diverse regions.

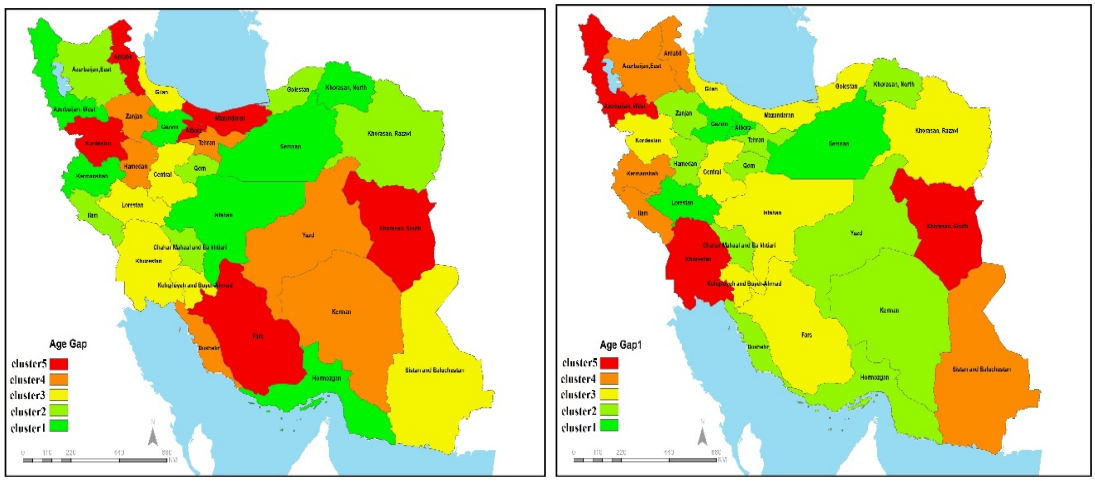
Finally, we visualize how the impact of age differences in couples varies significantly across provinces of Iran (Figure 6). Societies with traditional values or limited opportunities may show less adverse effects from larger age gaps than those with more liberal views or better opportunities.

We run 10,000 MCMC iterations, dropping the first 2000 as burn-in. We retained every fifth observation to reduce autocorrelation. We adopted a non-informative prior for the bandwidth and estimated the optimal bandwidth by the LPML method, choosing an optimal value of h at 4.2. The maximum distance between any two points is 10. The result from Dahl's method for the gwCRP model suggests that all couples are to be classified into five groups. However, our proposed



(a) (b)

Figure 5. The spatial varying covariate effects of Similarity on the two-part ZIDW model of provinces in Iran a:the probability of marriage survival becoming zero, b:the duration of time to divorce.



(a) (b)

Figure 6. The spatial varying covariate effects of the Age gap on the two-part ZIDW model of provinces in Iran a:the probability of marriage survival becoming zero, b:the duration of time to divorce.

gwCRP model, with its unique features, presents a different perspective. The sizes of the five groups in our model are7,9,4,6 and 5, respectively. The arrangement of these cluster assignments is based on Dahl's method, and the mode is depicted in (Figure 7), which illustrates their spatial distribution.

From (Figure 7), our gwCRP approach effectively identifies spatially connected and disconnected clusters. Provinces in the "light green" cluster exhibit spatial contiguity, and provinces in the "dark green" cluster display spatial discontinuity. Several interesting observations can be made from (Figure 6 and Table 5):

1. WestAzarbaijan, Kermanshah, Ilam, Khuzestan, Isfahan, Qom, Semnan, Khorasan North, Sistan, and Baluchestan all four covariates have moderate hazard effects compared with other counties.

2. East Azarbaijan, Golestan, Bushehr, Hormozgan, Kohgiluyeh, Buyer Ahmad, Chahar Mahall, and Bakhtiari starkly contrast in risk effects. Husband's employment status demonstrates significantly higher risk effects than Wife's Financial Autonomy status.

3. North Khorasan, Razavi Khorasan, Yazd, Gilan,

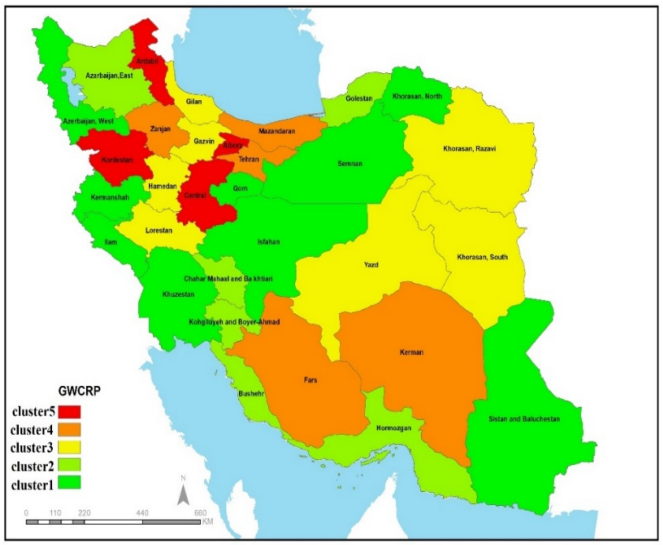


Figure 7. Clustering by gwCRP for the mean of estimated coefficients of the model

Table 5. Dahl's method estimates regression coefficients by gwCR

Cluster	$\widehat{\alpha_0}$	$\widehat{\alpha_1}$	$\widehat{\alpha_2}$	$\widehat{\alpha_3}$	$\widehat{\alpha_4}$	$\hat{\gamma}_0$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$	$\hat{\gamma}_4$
1	0.216	0.135	0.535	1.419	1.063	1.052	0.975	0.988	0.999	1.1103
2	2.492	0.364	0.785	1.668	1.110	0.965	0.981	0.065	1.000	0.3042
3	0.304	0.066	0.428	1.056	1.066	1.013	0.936	0.072	0.152	0.2587
4	1.975	-0.280	0.578	2.011	1.197	-0.259	0.753	-0.588,	-1.633	1.089
5	-0.801	0.135	0.047	-0.441	0.583	-0.417	1.539	2.444	-0.821	0.876

Ghazvin, Hamedan, and Lorestan are similar and have the highest risk effects in both model parts.

4. Mazandaran, Tehran, Kerman, and Fars:The spouses' Age differences have a negative risk effect in the model of non-zero count values and a positive effect in the probability of survival time becoming zero. The husband's employment has the most impact compared to other provinces.

5. Ardebil, Kordestan, Markazi, Alborz:The Wife's Financial Autonomy has the least hazard on the average duration of cohabitation, provided that the couple has lived together for at least 5 years.

Table 5 shows that the Bayes estimates of our spatial varying regression covariates coefficients through the gwCRP approach are quite different across different clusters.

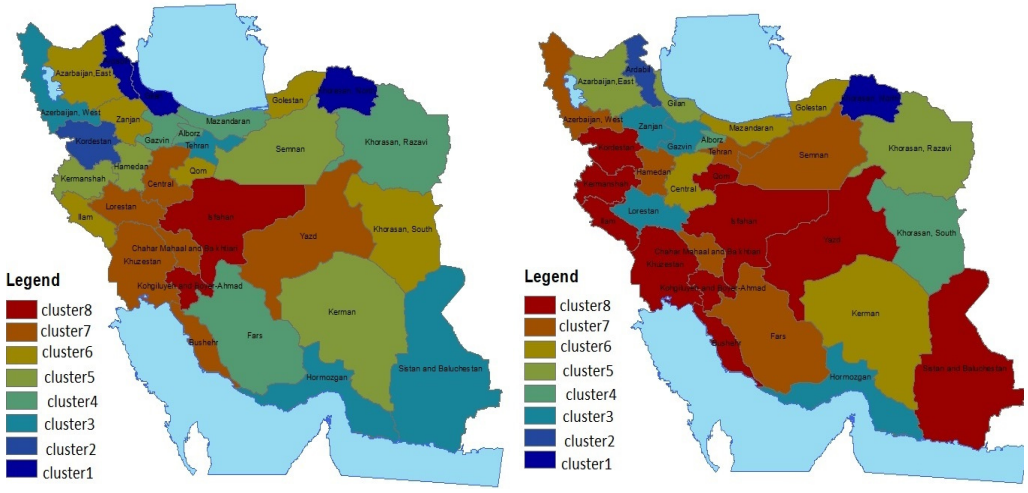
$\hat{\alpha}$, and $\hat{\gamma}$ are respectively the estimated regression coefficients for the models (7) and (8) in each cluster. They represent the effect of predictor variables (Husband's employment status, Wife's Financial Autonomy, Age gap, Similarity) on the duration of cohabitation provided that the couple has lived together for at least five years, within each cluster. For example, in Cluster 1, the estimated coefficient for Husband's employment status is 0.135. This means, that within Cluster 1, a one-unit increase in this covariate is associated with a 0.135 unit increase in c-log-log(q) in (1), holding all other variables constant. The intercept $\widehat{\alpha_0}$ represents the value of c-log-log(q) when all predictor variables are zero within that cluster. Also, in this cluster, the coefficient $\hat{\gamma}_1$ is 0.975 for the same covariate Husband's employment status in the logit model (8), suggesting a positive relationship between the Husband's employment status and the probability of divorce in less than five years represented by p. The effect is more substantial here than in the c-log-log model.

Finally, to show that our proposed clustering method, gwCRP, performs better than the two methods, traditional CRP and K-mean, in clustering regression coefficients in models 7 and 8 and determine which method yields estimation that best suits the data, the LPML values are calculated. As a more considerable LPML value indicates a better fit, we base our conclusion on the gwCRP results (Table 6).

It can be seen in (Figure 8) that the traditional K-mean method, high dimensional supervised

Table 6. LPML values for different methods in modeling divorce data

Method	gwCRP	CRP	K-mean
LPML	-367700.24	-406588.20	-416597.30



(a) (b)

Figure 8. Clustering by the k-mean with 8 clusters for the mean of a:$\hat{\alpha}$ and b:$\hat{\gamma}$

classification, and clustering, categorizes the provinces into 8 clusters, which leads to over-clustering.

To determine the optimal number of clusters (k), the Elbow method is used in k-means clustering, an empirical approach. This method is based on examining a graph that shows the value of the Within-Cluster Sum of Squares (WCSS) in terms of the number of clusters. WCSS is the sum of the squares of the distances of each data point to the center of its corresponding cluster. In (Figure 9), the horizontal axis represents the number of clusters (k), and the vertical axis represents WCSS. In this graph, the value of WCSS usually decreases as k increases. This decrease is significantly rapid at first, but after reaching a certain point (number 8), this decrease loses its speed and the downward trend becomes slower. This point, which resembles an elbow, indicates the optimal number of clusters.

Also, Comparing the clustering results using the traditional CRP method shown in (Figure 10) with the proposed method, we see that our proposed method successfully detects both spatially continuous clusters and discontinuous clusters simultaneously, however in the traditional CRP clustering method, neighboring provinces are more likely to be in the same cluster.

Discussion

In the present study, we propose an innovative Bayesian clustered coefficients regression model that employs a gwCRP to capture the spatial homogeneity of the regression coefficient proficiently. Our gwCRP models effectively address the intricate challenges associated with spatially varying coefficients in datasets characterized by right censoring and zero inflation. Through a combination of theoretical foundations and empirical evaluations, we provide compelling evidence that our methodologies yield precise parameter estimates within the ZIDW model while adeptly identifying the number of clusters and their configurations, even amidst varying proportions of zero counts. Furthermore, a comparative analysis with established clustering methodologies, such as K-means and traditional Chinese restaurant processes, illustrates that our approach achieves superior clustering concordance without additional tuning parameters, as indicated by higher Rand indexes, lower average mean squared error (AMSE), and improved log pseudo-marginal likelihood (LPML). Extensive simulation results are carried out using R version 4.3.3., to show that our proposed method has better clustering performance than the others. No issues with likelihood calculation were encountered in the simulations or the application to Iranian divorce data, however, the existence of two indicator functions often leads to extremely small likelihood functions, complicating the modeling process and requiring careful

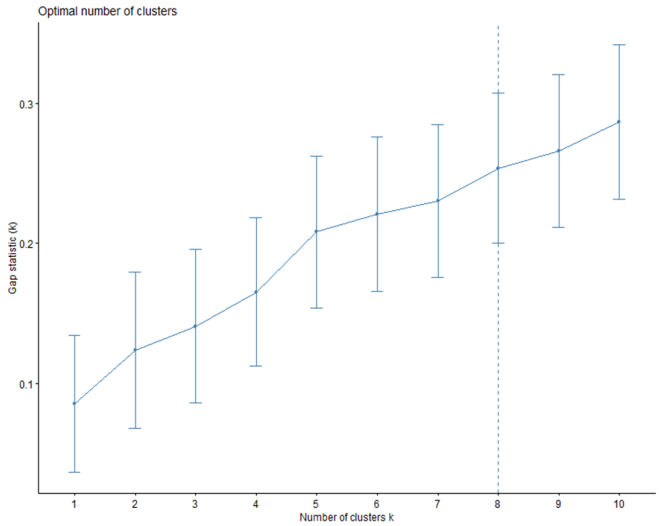
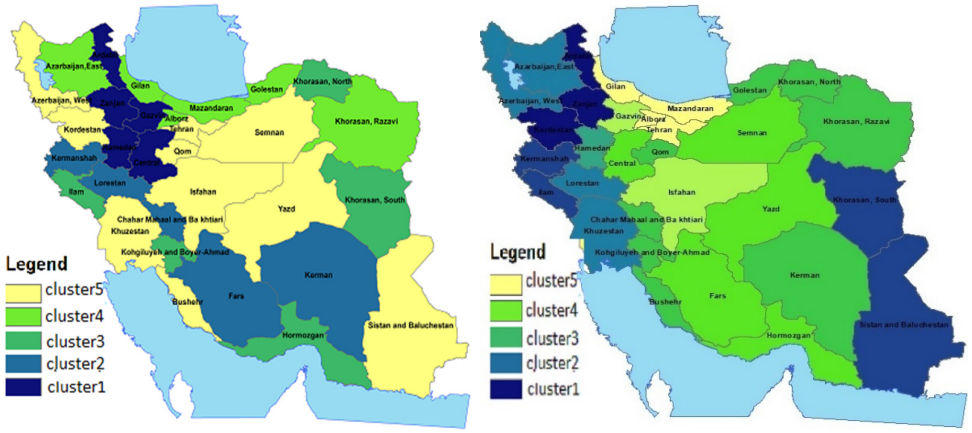


Figure 9. a: Elbow curve to determine the number of optimal clusters (k = 8), b:visualize the clustering results in the K-mean method



(a) (b)

Figure 10. Clustering by the CRP with 5 clusters for the mean of a:$\hat{\alpha}$ and b:$\hat{\gamma}$

consideration. While the discrete Weibull distribution proved beneficial for simulation data generation, using two-part regression models increased computational demands due to high-dimensional parameter spaces, resulting in extended convergence times. Furthermore, spatial heterogeneity and the inherent complexity of Bayesian hierarchical models contributed to substantial computational costs, particularly when analyzing the Iranian divorce dataset. Despite these computational challenges, our gwCRP model provides a robust and superior approach for analyzing spatially varying coefficients in complex datasets.

There are several possible directions for further investigation. The current model needs to be adapted to handle other related data (e.g., number of events) and longitudinal data (repeated measurements over time). Additionally, in this paper, our posterior sampling is based on the Chinese restaurant process, allowing for the inference of the number of clusters based on the unique latent cluster labels. To enhance the model, we suggest using a Mixed Finite Mixture (MFM) prior, allowing for the joint estimation of both regression coefficients and the probabilities of zero inflation (23) along with their associated clustering information. Finally, research is needed to improve computational efficiency, particularly for handling high-dimensional and sparse datasets, which

can be challenging to analyze.

Acknowledgments

The authors thank the editorial team and the anonymous reviewers for their comments, helpful suggestions, and encouragement, which helped improve the final version of this paper. Receiving support from the Center of Excellence in Analysis of Spatio-Temporal Correlated Data at Tarbiat Modares University is acknowledged.

References

1. Eissa S. Diagnostic biosensors for coronaviruses and recent developments. Advanced Biosensors for Virus Detection. 2022;8:261-278.
2. Moattari G, Izadi Z, Shakhsi-Niaei M. Development of an electrochemical genosensor for detection of viral hemorrhagic septicemia virus (VHSV) using glycoprotein (G) gene probe. Aquaculture. 2021; 536:736451.
3. Li J, Jin X, Feng M, Huang S, Feng J. Ulteransensitive and highly selective electrochemical biosensor for HIV gene detection based on Amino-reduced graphene oxide and b-cyclodextrin modified glassy carbon electrode. International Journal of Electrochemistry Science. 2020;15:2722-2738.
4. Zhao F, Bai Y, Cao L, Han G, Fang C, Wei S, Chen Z. New electrochemical DNA sensor based on nanoflowers of Cu3 (PO4) 2-BSA-GO for hepatitis B virus DNA detection. Journal of Electroanalytical Chemistry. 2020;867:114184.
5. Chowdhury AD, Takemura K, Li T-C, Suzuki TM, Park EY. Electrical pulse-induced electrochemical biosensor for hepatitis E virus detection. Nat, Commun. 2019;10:3737.
6. Lee T, Park SY, Jang H, Kim GH, Lee Y, Park C, Mohammadniaei M, Lee MH, Min J. Fabrication of electrochemical biosensor consisted of multi-functional DNA structure/porous au nanoparticle for avian influenza virus (H5N1) in chicken serum. Material Science Engineering. 2019;99:511-519.
7. Faria HAM, Zucolotto V. Label-free electrochemical DNA biosensor for zika virus identification. Biosensors and Bioelectronics. 2019;131:149-155.
8. Sabzi RE, Sehatnia B, Pournaghi-Azar MH, Hejazi MS. Electrochemical detection of human papilloma virus (HPV) target DNA using MB on pencil graphite electrode. Journal of the Iranian Chemical Society. 2008;5:476–483.
9. Campos-Ferreira DS, Souza EVM, Nascimento GA, Zanforlin DML, Arruda MS, Beltrão MFS, et al. Electrochemical DNA biosensor for the detection of human papillomavirus E6 gene inserted in recombinant plasmid. Arabian Journal of Chemistry. 2016;9:443–450.
10. Balvedi RPA, Castro ACH, Madurro JM, Brito-Madurro AG. Detection of a specific biomarker for epstein-barr virus using a polymer-based genosensor. International Journal of Molecular Sciences. 2014;15:9051–9066.
11. Dong S, Zhao R, Zhu J, Lu X, Li Y Qiu S, et al. Electrochemical DNA Biosensor Based on a Tetrahedral Nanostructure Probe for the Detection of Avian Influenza A (H7N9) Virus. ACS Applied Material Interfaces. 2015;7:8834–8842.
12. Shakoori Z, Salimian S, Kharrazi S, Adabi M, Saber R. Electrochemical DNA biosensor based on gold nanorods for detecting hepatitis B virus. Analytical and Bioanalytical Chemistry. 2015;407:455–461.
13. Manzano M, Viezzi S, Mazerat S, Marks RS, Vidic J. Rapid and label-free electrochemical DNA biosensor for detecting Hepatitis A virus. Biosensors and Bioelectronics. 2018;100:89-95.
14. Shawky SM, Awad AM, Allam W, Alkordi MH, EL-Khamisy SF. Gold aggregating gold: A novel nanoparticle biosensor approach for the direct quantification of hepatitis C virus RNA in clinical samples. Biosensors and Bioelectronics. 2017;92:349–356.
15. Mohammadi J, Moattari A, Sattarahmady N, Pirbonyeh N, Yadegari H, Heli H. Electrochemical biosensing of influenza A subtype genome based on meso/macroporous cobalt (II) oxide nanoflakes-applied to human samples. Analytica Chimica Acta. 2017;979:51–57.
16. Ilkhani H, Farhad S. A novel electrochemical DNA biosensor for Ebola virus detection. Analytical Biochemistry. 2018;557:151–155.
17. Marrazza G, Ramalingam M, Jaisankar A, Cheng L, Selvolini G, Vitale IA. Advancements and emerging technologies in biosensors for rapid and accurate virus detection. TrAC Trends in Analytical Chemistry. 2024;172:117609.
18. Rezaei B, Jamei HR, and Ensafi AA. An ultrasensitive and selective electrochemical aptasensor based on RGO-MWCNTs/Chitosan/carbon quantum dot for the detection of lysozyme. Biosensors and Bioelectronics. 2018;115:37-44.
19. Izadi Z, Sheikh-Zeinoddin M, Ensafi AA, Soleimanian-Zad S. Fabrication of an electrochemical DNA-based biosensor for Bacillus cereus detection in milk and infant formula. Biosensors and Bioelectronics. 2016;80:582–589.
20. Mansor NA, Zain ZM, Hamzah HH, Noorden MSA, Jaapar SS, Beni V, Ibupoto ZH. Detection of Breast Cancer 1 (BRCA1) Gene Using an Electrochemical DNA Biosensor Based on Immobilized ZnO Nanowires. Open Journal of Applied Biosensors. 2014;3:9–17.
21. Izuan J, Rashid A, Azah N. The strategies of DNA immobilization and hybridization detection mechanism in the construction of electrochemical DNA sensor : A review. Sensing and Bio-Sensing Research. 2017;16:19–31.
22. Chun L, Kim SE, Cho M, Cheo WS, Nam J, Lee DW, Lee Y. Electrochemical detection of HER2 using single stranded DNA aptamer modified gold nanoparticles electrode. Sensors and Actuators B. 2013;186:446–450.
23. Kavita V J. DNA Biosensors-A Review. Journal of Bioengineering Biomedical Science. 2017;7:1-5.
24. Gunasekaran B M, Srinivasan S, Ezhilan M, Nesakumar N. Nucleic acid-based electrochemical biosensors. Clinica Chimica Acta. 2024;559:119715.
25. Eksin E, Yildirim A, Bozoglu A, Zor E, Erdem, A. Paper-based nucleic acid biosensors. TrAC Trends in Analytical Chemistry. 2024;171:117511.