IRANIAN STATISTICAL SOCIETY 1990

Tarbiat Modares University

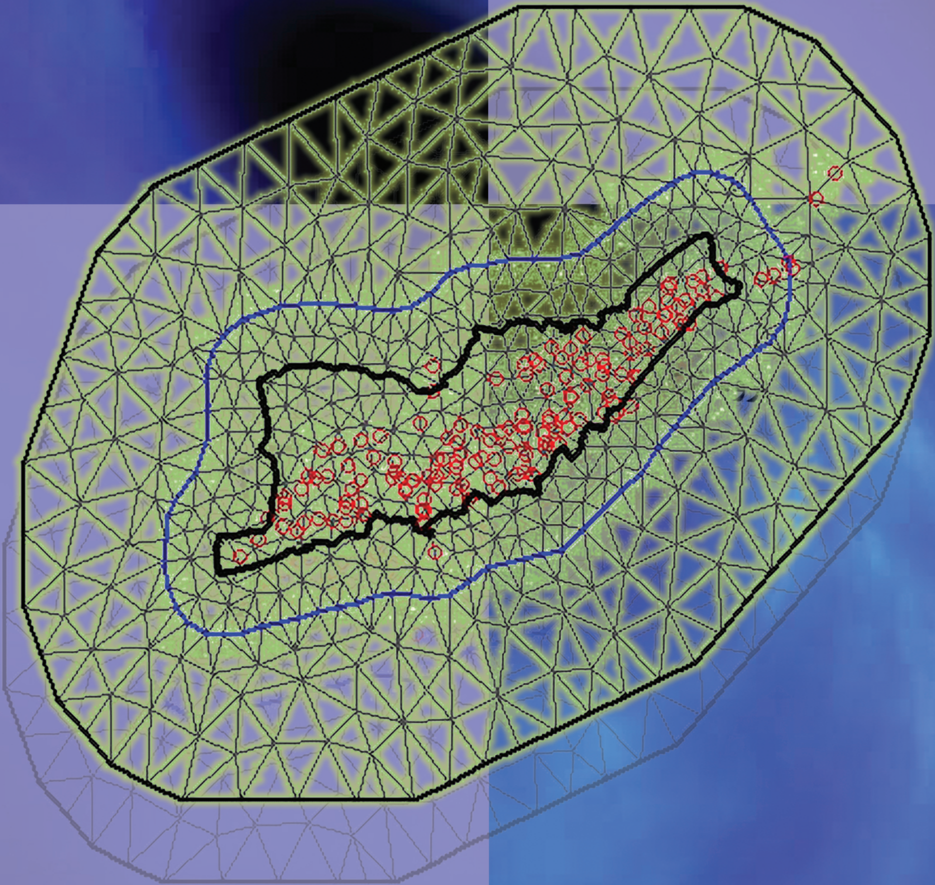
IMAM KHOMEINI INTERNATIONAL UNIVERSITY

# پنجمین سمینار
# آمار فضایی و کاربردهای آن

۳-۴ آبان ماه ۱۴۰۲
دانشگاه بین‌المللی امام‌خمینی، قزوین

# 5th Seminar on Spatial Statistics and Its Applications

Oct. 2023 25-26
Imam Khomeini International University



PROCEEDING

In the Name of God

# Proceedings of the 5th Seminar on Spatial Statistics and Its Applications

**25-26 October 2023**

International University of Imam Khomeini (IUIK)

Qazvin, Iran.

Proceedings of the 5th Seminar on Spatial Statistics and Its Applications

| | |
|---|---|
| Producer: | Mohsen Mohammadzadeh |
| Printing: | October 2023 |
| Address: | International University of Imam Khomeini, Qazvin, Iran. |
| Phone: | 028 – 33901379 |
| E-mail: | spatial5@ikiu.ac.ir |
| Website: | conf.ikiu.ac.ir/spatial5 |

# Preface

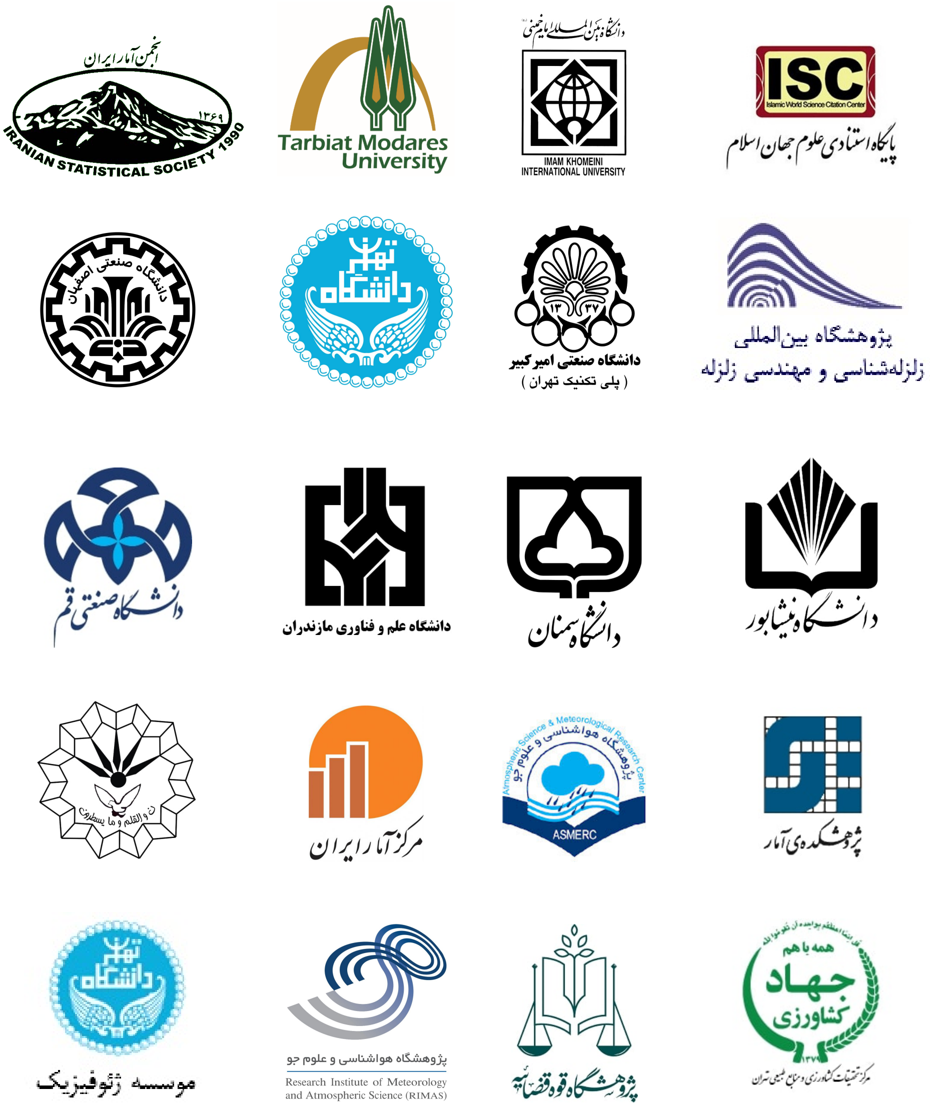
In the Name of GOD

The increasing development of human societies in various fields is gaining momentum every day. To overcome and control this increasing growth, the need for advanced methods of modeling different phenomena becomes doubly important. Most experimental phenomena have a series of dependent and independent variables. Discovering and modeling the dependence of these variables has a vital role in a better and fact-based understanding of those phenomena. Statistical sciences and new methods of data science play a key role in this regard and promote interdisciplinary collaboration. Spatial statistics is a powerful tool, which examines their correlations by analyzing spatial and temporal data. With this in mind, spatial statistics methods can be used in a wide range of areas. Including Earthquake Science and Engineering, Risk Engineering, Crisis Management, Atmospheric and Meteorological Sciences, Water Resources, Environment, Geology, Mining, Urban and Regional Planning, Traffic, Transportation, Remote Sensing, Health and Treatment, epidemiology, forensics, social sciences, oil and gas, economics, and insurance have a wide range of applications. To provide opportunities for the exchange of views of experts in various related fields to spatial statistics, the Fifth Seminar on Spatial Statistics and Its Applications, to be held from 25 to 26 October 2023, is hosted by the International University of Imam Khomeini in collaboration the Centre of Excellence in Analysis of Spatio-Temporal Correlated Data Tarbiat Modares University and the Iranian Statistical Society will be held. This seminar provides a unique opportunity for academics, professionals, government agencies, the private sector and other institutions active in related fields to exchange views and present the results of their research by presenting the latest scientific achievements. Thanks to the esteemed experts from inside and outside the country in various fields who contribute to the scientific fruitfulness of this seminar by presenting their valuable articles and the respected referees, scientific committee, and executive committee who took great efforts to hold this seminar. We hope that with your active presence and participation in this seminar, it will be possible to achieve its predicted goals like the previous successful seminars.

**Secretary of the Scientific Committee**
**Professor Mohsen Mohammadzadeh**
**October 2023**

## Seminar Partners and Sponsors:

This seminar is hosted by the International University of Imam Khomeini in cooperation with the Centre of Excellence in Analysis of Spatio-Temporal Correlated Data at Tarbiat Modares University and the Iranian Statistics Society, as well as the support of organizations and institutions listed below. We would like to express our gratitude to all the individuals and organizations that supported the seminar.


انجمن آمار ایران
IRANIAN STATISTICAL SOCIETY 1990
Tarbiat Modares University
IMAM KHOMEINI INTERNATIONAL UNIVERSITY
ISC
Islamic World Science Citation Center
پایگاه استنادی علوم جهان اسلام
دانشگاه صنعتی اصفهان
دانشگاه تهران
دانشگاه صنعتی امیرکبیر
( پلی تکنیک تهران )
پژوهشگاه بین المللی
زلزله‌شناسی و مهندسی زلزله
دانشگاه صنعتی قم
دانشگاه علم و فناوری مازندران
دانشگاه سمنان
دانشگاه نیشابور
مرکز آمار ایران
ASMERC
پژوهشکده‌ی آمار
موسسه ژئوفیزیک
پژوهشگاه هواشناسی و علوم جو
Research Institute of Meteorology and Atmospheric Science (RIMAS)
پژوهشگاه قوه قضاییه
جهاد کشاورزی

# Scientific Committee

| | |
|---|---|
| 1- Mohsen Mohammadzadeh (Secretary) | Tarbiat Modares University |
| 2- Afshin Fallah | Imam Khomeini International University |
| 3- Ali Aghamohammadi | Zanjan University |
| 4- Hossein Baghishani | Shahrood University of Technology |
| 5- Morteza Bastami | International Institute of Earthquake Engineering and Seismology |
| 6- Fatemeh Hosseini | Semnan University |
| 7- Ramin Kazemi | Imam Khomeini International University |
| 8- Omid Karimi | Semnan University |
| 9- Mousa Golalizadeh | Tarbiat Modares University |
| 10- Kiomars Motarjem | Tarbiat Modarres University |

# Excutive Committee

| | |
|---|---|
| 1- Afshin Fallah (Seminar Secretary) | Imam Khomeini International University |
| 2- Sedigheh Zamani Mehreyan | Imam Khomeini International University |
| 3- Elias Shivanian | Imam Khomeini International University |
| 4- Maryam Dargahi | Imam Khomeini International University |
| 5- Ramin Kazemi | Imam Khomeini International University |
| 6- Arezo Hajrajabi | Imam Khomeini International University |
| 7- Mahsa Nadifar | University of Pretoria |

# Contents

# Determining the Anisotropic Spatial Correlation of $V_{s30}$ Values in Tehran

Morteza Abbasnejad Fard*, Morteza Bastami
International Institute of Earthquake Engineering and Seismology, Tehran, Iran.

**Abstract:**

Spatial correlation and cross-correlation of earthquake intensity measures (IMs) are essential for seismic hazard and risk assessment of spatially distributed assets, such as portfolios of buildings or infrastructure networks. Recent studies have shown that the spatial correlation characteristics of local soil conditions, represented by the average shear-wave velocity in the upper 30 m of soil (Vs30), significantly impact the spatial correlations of earthquake IMs. This study aims to analyze the spatial correlation characteristics of the soil profile in the Tehran region by collecting accurate Vs30 measurements and obtaining the parameters of a multivariate anisotropic spatial correlation model of earthquake IMs for seismic hazard and risk assessment applications in Tehran.

**Keywords:** Spatial correlations, Earthquake, Latent dimensions, Anisotropy.
**Mathematics Subject Classification (2010):** 62P30, 62N01, 62H11.

## 1 Introduction

Vs30 is the time-averaged shear-wave velocity in the upper 30 m of soil, which is a key measure to characterize the site response and classify the site conditions. Vs30 values are not constant over a large area but vary spatially due to the heterogeneity of the soil layers and the topography. It is necessary to account for the spatial correlation of Vs30 values when estimating the ground motion IMs at different locations, especially for large-scale urban areas prone to earthquakes. Ignoring the spatial correlation of earthquake intensity measures (IM) and Vs30 values can lead to unrealistic loss estimation and inaccurate resilience assessment, as well as increased uncertainty and variability in the ground motion

*Speaker: m.abbasnejad@iiees.ac.ir

IMs Abbasnejadfard et al. (2021a,b). Several studies have proposed different methods to model the spatial correlation of Vs30 values, such as geostatistical methods, proxy-based methods, and latent dimensions methods. Recently developed methods (e.g., Abbasnejadfard et al. (2020)) can also capture the anisotropy of the spatial correlation, which means that the correlation depends on the direction and the distance between two locations. Anisotropy is another critical factor that should be considered in spatial correlation modelling of earthquake IMs, as it can affect the results significantly. Abbasnejadfard et al. (2019) developed a latent dimensions method to model the anisotropic spatial correlation of earthquake IMs, which relies on the anisotropic spatial correlation characteristics of Vs30 values in the target region. To apply this method in Tehran, the capital city of Iran, this study collected measured Vs30 values within and around the boundaries of Tehran and analyzed the spatial correlation properties of the Vs30 random field in the area. The parameters of the latent dimensions model for the anisotropic spatial correlation of earthquake IMs in the region were obtained, which can be used for further spatially correlated seismic hazard and risk assessment of the buildings and infrastructure networks in Tehran metropolitan area.

## 2 Application of the Latent Dimensions Method

Apanasovich and Genton (2010) proposed an innovative approach based on the latent dimensions method based on existing covariance models of univariate random fields to define closed-form cross-covariance function. The key idea is to represent a vectors components as points in $k$ dimensional space and convert a multivariate problem to a multidimensional univariate one. In this regard, each component $\alpha$ of a multivariate random field $\varepsilon'(\mathbf{s})$ considered as a point of univariate random field in $k$ dimensional space. Based on these latent dimensions, $C_{\alpha\beta}(\mathbf{s}_1, \mathbf{s}_2) : \ \mathbf{s}_1, \mathbf{s}_2 \in \mathbb{R}^n$ becomes as $C\left((\mathbf{s}_1, \xi_\alpha), (\mathbf{s}_2, \xi_\beta)\right)$, a covariance of a univariate random field which its arguments are from $\mathbb{R}^{n+k}$ space instead of $\mathbb{R}^n$. In the case of the current study, $n$ equals 2 (because data have been recorded in 2-dimensional space), and $k$ is considered as 1, so a 2-dimensional three-variate random field of normalized intra-event residuals of PGA, PGV and PGD is represented as a 3-dimensional univariate random field. Consequently, using valid covariance models of univariate random fields, the covariance matrix is guaranteed to be non-negative definite. The current study implements the cross-covariance function form of equation (2.1) proposed by Apanasovich and Genton (2010) based on the latent dimension method.

$$C_{\alpha\beta}(\mathbf{h}) = C(\mathbf{h}, \upsilon_{\alpha\beta} - \mathbf{\Gamma}_\xi \mathbf{h}) = \frac{\sigma_{\alpha\beta}}{|\upsilon_{\alpha\beta} - \gamma\omega^{\mathrm{T}}\mathbf{h}| + 1} \exp\left\{ -\frac{a\,\|\mathbf{h}\|}{\left(|\upsilon_{\alpha\beta} - \gamma\omega^{\mathrm{T}}\mathbf{h}| + 1\right)^{1/2}} \right\}, \quad (2.1)$$

where $\alpha, \beta = 1, 2, 3$ represent $\varepsilon'_{PGA}$, $\varepsilon'_{PGV}$ and $\varepsilon'_{PGD}$, normalized intra-event residuals of PGA, PGV and PGD respectively which are the components of multivariate random field in original space, $\underset{\sim}{h} = \mathbf{s}_i - \mathbf{s}_j$ is relative position vector of points i and j, $\sigma_{\alpha\beta}$ is variance parameter and $\omega^{\mathrm{T}} = \{\omega_1, \omega_2\}^{\mathrm{T}}$ is a 2-dimensional vector such that $\omega^{\mathrm{T}}\omega = 1$. In this equation, $\arctan(\omega_2/\omega_1)$ shows the anisotropy direction and $\gamma \geq 0$ defines anisotropy ratio. Normalized intra-events residuals are defined as (2.2) assuming that the standard deviation of intra-event residuals are independent of location (Jayaram and Baker, 2009; Du and Wang, 2013; Garakaninezhad and Bastami, 2017) and consequently they have unite standard deviation.

$$\varepsilon'_{ij} = \frac{\varepsilon_{ij}}{\sigma} = \frac{\ln(Y_{ij}) - \ln(\overline{Y_{ij}})}{\sigma} \tag{2.2}$$

Instead of using latent dimension,$\xi_1 = \{\xi_{11}, \xi_{12}, \xi_{13}\}^{\mathrm{T}}$ it is possible to treat with latent distances $\upsilon_{\alpha\beta} = \xi_{1\alpha} - \xi_{1\beta}$, $\alpha, \beta = 1, 2, 3$. The larger latent distance $\upsilon_{\alpha\beta}$ indicates the smaller cross-correlation between components $\alpha$ and $\beta$. The vector $\omega$ and the parameter $\gamma$ determine the anisotropy direction and anisotropy extend (ratio), respectively. Moreover, the latent distance parameter $\upsilon$ determines the correlation between different components (variable), so a larger value of $\upsilon$ shows a more negligible correlation. In this regard, $\upsilon$ equals to 0 for marginal-covariance models.

According to the Abbasnejadfard et al. (2020), the anisotropy ratio and anisotropy direction in the model mentioned above are directly affected by the anisotropy ratio and anisotropy direction of the random field of the Vs30 values in the region. For this reason, the following sections are dedicated to investigating these parameters.

## 3 Description of Collected Data

VS30 values of the 158 sample points within the study area and an area with a distance of about 17 km from the study area's borders are selected for the statistical investigations. Figure 1 demonstrates the location of the sample points in the study area and its vicinity. The box plot and histogram of the collected data are also presented in Figure 1. Moreover, the statistical characteristics of the observations are provided in Table 1.

Hainings method is used to identify the outlier observations. In this regard, the observation that satisfies one of the inequalities of equation (3.1) is considered an outlier and excluded from the collected observations.

$$Z(\mathbf{s}) < Q_L - 1.5\,(Q_U - Q_L)\,, \quad Z(\mathbf{s}) > Q_L + 1.5\,(Q_U - Q_L) \tag{3.1}$$

In (3.1), $Z(\mathbf{s})$ is the observed value at location s, $Q_L$ is the lower quartile, and $Q_U$ is

Table 1: Geometric parameters of the seismic source

| Statistical Characteristic | Value |
|---|---|
| Number | 158 |
| Mean (m/s) | 517.2 |
| Standard Deviation (m/s) | 166.4 |
| Minimum (m/s) | 187.0 |
| Lower Quartile (m/s) | 388.0 |
| Median (m/s) | 512.0 |
| Upper Quartile (m/s) | 634.0 |
| Maximum (m/s) | 1236.0 |
| Interquartile Range (m/s) | 246.0 |

the upper quartile of observations. Using equation (3.1), one of the observations with a VS30 value equal to 1236 m/s is considered as outlier data and excluded from the observations. To examine the stationarity of the data in terms of mean values, scatter plots of the observed data are drawn versus the relative distances between locations in the east-west (X) and north-south (Y) directions. Figure 2 demonstrates that the data have considerable trends in both directions, and the north-south trend is more significant than the trend in the east-west direction. In order to reduce the adverse effects of the existing trend on the estimation of the variables and predictors, it is necessary to remove trends from the data. The trend model is first determined using the linear regression approach in this context. Subtracting the modeled trend value from the observed value at each location provides the residual, also known as the detrended data. Figure 2 depicts the scatter plots of the detrended data versus the relative distances between locations in the east-west and north-south directions. Moreover, the histogram of the detrended data is presented in Figure 3. According to this figure, detrended VS30 values follow a normal distribution. In order to further investigate the normality of the detrended data, the Q-Q plot of Figure 3 is presented. According to this figure, the detrended VS30 values are symmetric and follow the normal distribution with a good approximation. Considering the results presented in Figures 2 through 3, it can be concluded that the random field of the detrended VS30 values in the study area is Gaussian and stationary.

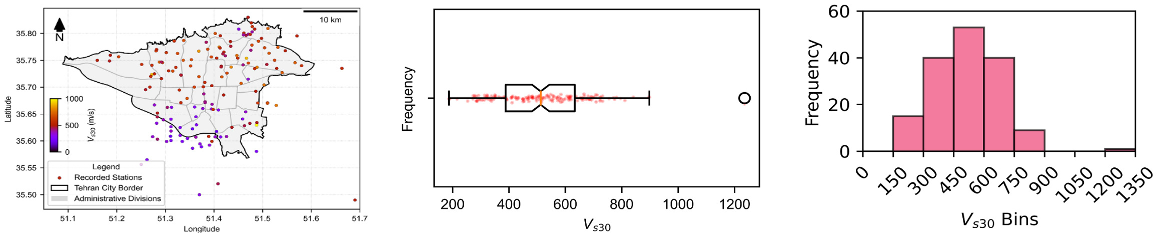


Figure 1: Location, Box plot and histogram of the collected sample points

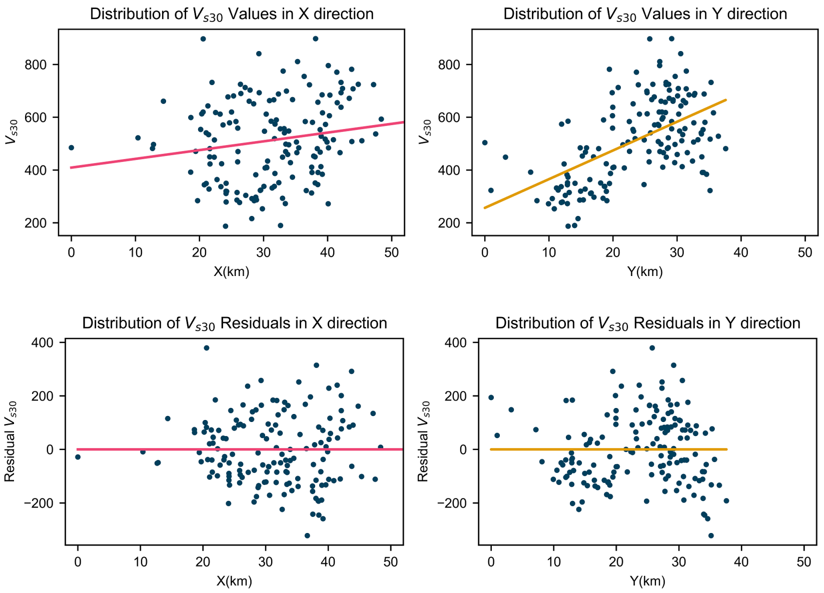


Figure 2: Scatter plot of the observed VS30 values (top) and detrended data (bottom) versus the relative distances between locations in the east-west and north-south directions

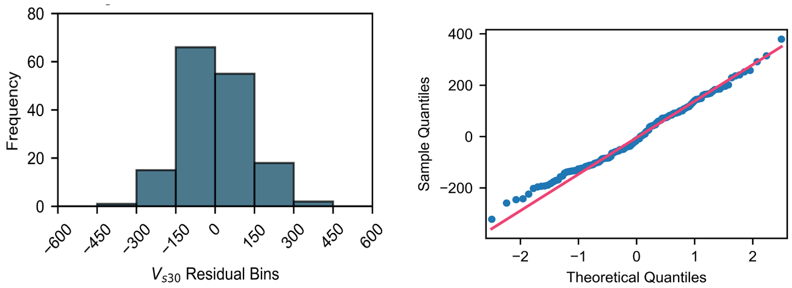


Figure 3: Histogram and Q-Q plot of detrended data

# 4 Determining the Bivariate Covariogram Function

In order to capture the anisotropic spatial correlation characteristics of random fields, it is possible to use valid bivariate semivariogram (or covariogram) functions. These types of functions lead us to a valid covariance matrix (a positive semi-definite matrix with valid Cholesky decomposition). See Abbasnejadfard et al. (2020) for more details. The current research work uses the covariogram function of equation (4.1), which was also utilized by Abbasnejadfard et al. (2019) and adapted from Apanasovich and Genton (2010).

$$C(\mathbf{h}) = \frac{\sigma}{|\gamma\omega^{\mathrm{T}}\mathbf{h}| + 1} \exp\left\{-\frac{a\,\|\mathbf{h}\|}{(|\gamma\omega^{\mathrm{T}}\mathbf{h}| + 1)^{1/2}}\right\} \tag{4.1}$$

In equation (4.1), $\mathbf{h} = \mathbf{s}_i - \mathbf{s}_j$ is the distance vector of points $i$ and $i$, $\sigma$ represents the covariance value at $\mathbf{h} = \mathbf{0}$, which is equal to the nugget value of the semivariogram models, $\gamma \geq 0$ is a parameter that includes the effects of anisotropy ratio, $\omega^{\mathrm{T}} = \{\omega_1, \omega_2\}^{\mathrm{T}}$ is a 2-dimensional vector that determines the anisotropy direction, and a is the range parameter. In order to obtain the mentioned parameters for the random field of VS30 values in Tehran, first, the empirical covariogram values of the normalized detrended VS30 values are calculated for different directions and distances using the gstat package of the R programming language. Then, the nonlinear least square regression with the least absolute residuals (LAR) method and the Trust-Reagion algorithm is utilized to fit the equation (4.1) to the calculated empirical covariogram values. The visual representation of the fitted covariogram function is shown in Figure 4. Moreover, the model parameters and coefficient of determination are presented in Table 2.

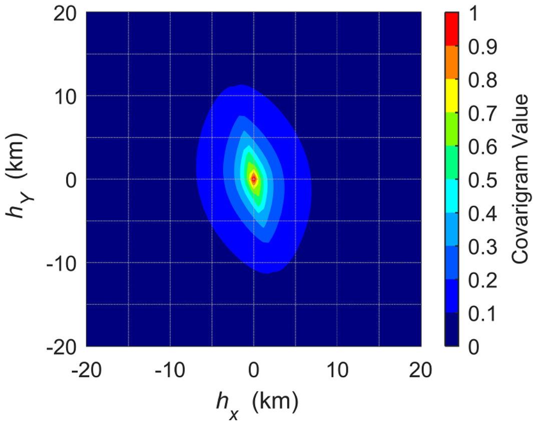


Figure 4: The fitted covariogram function

Table 2: Geometric parameters of the seismic source

| $\sigma$ | $a$ | $\gamma$ | $\omega^{\mathrm{T}}$ | $\mathrm{R}^2$ |
|---|---|---|---|---|
| 0.96 | 0.204 | 0.689 | [0.92, 0.39] | 0.794 |

# 5 Estimation of the Parameters

By obtaining the characteristics of the anisotropic spatial correlations of the VS30 values, it would be possible to employ the latent dimensions (LD) method, proposed by Abbasnejadfard et al. (2020), to calculate the anisotropic spatially correlated seismic hazard in the considered study are. According to the LD method, the marginal- and cross-covariance functions of equations (5.1) and (5.2) should be used.

$$C_{\alpha\alpha}(\mathbf{h}) = \frac{1}{|\gamma\omega^{\mathrm{T}}\mathbf{h}| + 1} \exp\left\{ -\frac{a\,\|\mathbf{h}\|}{(|\gamma\omega^{\mathrm{T}}\mathbf{h}| + 1)^{1/2}} \right\} \tag{5.1}$$

$$C_{\alpha\beta}(\mathbf{h}) = \frac{1}{|\upsilon_{\alpha\beta} - \gamma\omega^{\mathrm{T}}\mathbf{h}| + 1} \exp\left\{ -\frac{a\,\|\mathbf{h}\|}{(|\upsilon_{\alpha\beta} - \gamma\omega^{\mathrm{T}}\mathbf{h}| + 1)^{1/2}} \right\} \tag{5.2}$$

In equations (5.1) and (5.2), $\alpha$ and $\beta$ determine the earthquake intensity measures, $\upsilon$ is known as the latent distance value, and other parameters are defined under (2.1). The parameters of equations (5.1) and (5.2) ($a$, $\gamma$, $\omega$, and $\upsilon$) for different combinations of earthquake intensity measures are provided in Table 3. The values provided in the mentioned tables can be used to conduct anisotropic spatially correlated earthquake hazard assessment in the Tehran region using the LD method. More details about the calculation of these parameters based on the characteristics of anisotropic spatial correlations of local VS30 values are presented in Abbasnejadfard et al. (2020).

Table 3: Geometric parameters of the seismic source

| | Model 5 | | | | Model 6 | | |
|---|---|---|---|---|---|---|---|
| $\alpha$ | $a$ | $\gamma$ | $\omega^{\mathrm{T}}$ | $\beta$ | $a$ | $\gamma$ | $\omega^{\mathrm{T}}$ |
| PGA | 1.751 | 0.572 | [0.92, 0.39] | PGV | 1.751 | 0.572 | [0.92, 0.39] |
| PGA | 2.001 | 0.937 | [0.92, 0.39] | PGD | 2.001 | 0.937 | [0.92, 0.39] |
| PGV | 2.472 | 0.565 | [0.92, 0.39] | PGD | 2.472 | 0.565 | [0.92, 0.39] |
| SA(T=0.5s) | 2.212 | 1.130 | [0.92, 0.39] | SA(T=1s) | 2.212 | 1.130 | [0.92, 0.39] |
| SA(T=0.5s) | 2.212 | 0.744 | [0.92, 0.39] | SA(T=2s) | 2.212 | 0.744 | [0.92, 0.39] |
| SA(T=1s) | 2.472 | 0.503 | [0.92, 0.39] | SA(T=2s) | 2.472 | 0.503 | [0.92, 0.39] |

# Discussion and Results

This study focuses on collecting measured Vs30 values in Tehran and analyzing their spatial correlation characteristics. A total of 158 Vs30 values in Tehran are collected for this purpose. The analysis shows that the Vs30 values can be regarded as a realization of a non-stationary anisotropic random field with an anisotropic range of 1.45 and an anisotropy direction aligned with approximately the North-South direction. Furthermore, an anisotropic covariogram model is fitted to the data and its parameters are estimated. Based on the parameters of the anisotropic covariogram model, the parameters of the multivariate anisotropic spatial correlation model of earthquake intensity measures proposed by Abbasnejadfard et al. (2020) are obtained, which enable conducting spatially correlated seismic hazard and risk assessment in Tehran region.

# Acknowledgement

The authors gratefully acknowledge Dr. Mohammad Kazem Jafari and Dr. Asghar Azadi for providing the data that were utilized in this study.

# References

Abbasnejadfard, M., Bastami, M., Fallah, A. (2019), Application of Multivariate Spatial Correlation Model in Seismic Hazard Analysis Considering Anisotropy. *Proceedings of the 3rd Seminar on Spatial Statistics and Its Applications*, 1–12.

Abbasnejadfard, M., Bastami, M., Fallah, A. (2020), Investigation of Anisotropic Spatial Correlations of Intra-event Residuals of Multiple Earthquake Intensity Measures using Latent Dimensions Method. *Geophys J Int*, **222**, 14491469.

Abbasnejadfard, M., Bastami, M., Fallah, A., Garakaninezhad A. (2021), Analyzing the Effect of Anisotropic Spatial Correlations of Earthquake Intensity Measures on the Result of Seismic Risk and Resilience Assessment of the Portfolio of Buildings and Infrastructure Systems. *Bull. Earthq. Eng.*, **19**, 5791–5817.

Abbasnejadfard, M., Bastami, M., Fallah, A., Garakaninezhad, A. (2021), Significance of Anisotropic Spatial Correlation Considerations of Earthquake Intensity Measures on the Seismic Risk Assessment of Infrastructure Systems. *Proceedings of the 4th Seminar on Spatial Statistics and Its Applications*, **19**, 1–8.

Abbasnejadfard, M., Bastami, M., Jafari, M.K., Azadi, A. (2023), Spatial Correlation Models of VS30 Values: A Case Study of the Tehran Region. *Engineering Geology*,

Apanasovich, T., Genton, M. (2010), Cross-covariance Functions for Multivariate Random Fields based on Latent Dimensions, *Biometrika*, **97**, 15-30.

Du, W., Wang, G. (2013), Intra-event Spatial Correlations for Cumulative absolute Velocity, Arias Intensity, and Spectral Accelerations based on Regional Site Conditions, *Bulletin of the Seismological Society of America*, **103**, 1117-1129.

Garakaninezhad, A., Bastami, M. (2017), A Novel Spatial Correlation Model based on Anisotropy of Earthquake Ground-motion Intensitya Novel Spatial Correlation Model based on Anisotropy of Earthquake Ground-motion Intensity, *Bulletin of the Seismological Society of America*, **107**, 2809-2820.

Jayaram, N., Baker, J. W. (2009), Correlation Model for Spatially Distributed Ground-motion Intensities, *Earthquake Engineering & Structural Dynamics*, **38**, 1687-1708.

# Enhancing Suicide Mortality Prediction Using Spatially Informed Random Forest Models: A Comparative Study with Spatial Econometrics

Mohadeseh Alsadat Farzammehr*
Judiciary Research Institute, Tehran, Iran.

## Abstract:

Amidst the growing adoption of novel machine learning techniques like random forest, grasping the significance of spatial factors within these models is pivotal. This study introduces an innovative approach, crafting spatially informed classification random forest models by integrating spatially lagged variables, mirroring diverse spatial panel data econometric frameworks. Our investigation rigorously compares these models to traditional spatial and non-spatial regression methods in predicting suicide mortality rates across Iran's provinces from 2011 to 2022. Outcomes reveal a nuanced edge of spatial econometric models over random forest counterparts. Remarkably, the optimal spatial random forest model, infused with spatial lag parameters, attains an impressive 89.19% predictive accuracy for suicide mortality levels, surpassing both spatial econometric (46.51%) and non-spatial random forest (27.03%) models. Despite these variances, our conclusion underscores that random forest methods don't surpass traditional spatial econometric models in predicting suicide mortality rates. These findings offer vital insights into spatial considerations within predictive modeling, guiding researchers towards apt choices for spatial data analysis models.

**Keywords:** Court performance prediction; Data mining; Judicial data; Machine learning techniques; Artificial intelligence.

**Mathematics Subject Classification (2010):** 6207, 62H30, 62H11.

*Speaker: m-farzammehr@jri.ac.ir

# 1 Introduction

Suicide mortality poses a pressing public health challenge that reverberates across various societal realms. Tackling this issue necessitates accurate predictive models for timely interventions and resource allocation. Amid the emergence of machine learning techniques, such as random forest, their potential becomes evident in handling intricate datasets. However, these techniques often disregard spatial dimensions crucial for deciphering nuanced geographic patterns in suicide rates and risk factors.

Conventional machine learning models in spatial contexts can obscure underlying relationships, particularly when spatial interactions are pivotal. Spatial econometrics steps in to grapple with complexities tied to spatial data. Nevertheless, a comparative analysis between spatial econometrics and machine learning, such as random forest, remains elusive.

This study pioneers an innovative fusion, harnessing the strengths of spatial econometrics and random forest. We craft novel spatially informed classification random forest models, seamlessly integrating spatial lagged variables to mirror the structures of spatial panel data econometrics. Our assessment comprehensively contrasts these hybrids with spatial econometrics and non-spatial regression. Importantly, we predict suicide rate categories across Iran's provinces from 2011 to 2022.

The research centers around two pivotal questions: a comparison between predictive random forest and spatial econometrics, and an exploration of spatially explicit random forest's potential to surpass conventional methods in suicide rate prediction. Unraveling these inquiries offers insights into the synergy between machine learning and spatial econometrics, providing guidance for model selection in spatial data analysis, particularly in the realm of suicide prediction. In essence, this paper seamlessly integrates the urgency of addressing suicide concerns with state-of-the-art predictive analytics and spatial insights. Through a skillful amalgamation of spatial econometrics and random forest, our study heralds a new era of informed predictive modeling. It provides a guiding beacon for researchers and policymakers, ushering in an era of heightened spatially conscious data analysis.

# 2 Data Collection and Preparation

Iran's suicide data originates from the Iranian Forensic Medicine Organization (IFMO), an entity under the Iranian Judicial Authority. IFMO operates a comprehensive suicide registry and conducts autopsies for documented cases. Suicide rates (per 100,000) were computed for each province. Socio-demographic and economic data spanning 2011 to 2022 for all 31 provinces were collated from the Statistical Center of Iran. The dataset includes variables such as unemployment rate (X1), labor force participation rate (X2), ln(population aged 15 and over) (X3), consumer price index (CPI) (X4), literacy rate (X5), and ln(gross domestic product) (X6). These variables serve as inputs for an econometric model that accounts for spatial correlations, elucidating factors influencing suicide rates (y) per 100,000 population.

Table 1: Descriptive Summary and Data Transformation.

| Variable | Minimum | Mean | Maximum | SD | Skewness | Kurtosis |
|---|---|---|---|---|---|---|
| y | 1.70 | 6.24 | 19.70 | 3.57 | 1.44 | 4.74 |
| X1 | 5.80 | 11.20 | 21.70 | 3.07 | 0.90 | 4.16 |
| X2 | 32.30 | 41.38 | 50.20 | 3.62 | -0.28 | 3.01 |
| X3 | 12.38 | 14.13 | 16.46 | 0.77 | 0.12 | 2.63 |
| X4 | 39.20 | 139.40 | 401.00 | 97.43 | 1.18 | 3.26 |
| X5 | 70.80 | 84.49 | 92.90 | 4.36 | -0.62 | 3.52 |
| X6 | 10.30 | 11.77 | 14.39 | 0.85 | 0.59 | 3.23 |

The dataset encompasses both independent and dependent variables in Table 1, offering insights into the study. Over an 11-year span, suicide mortality rates averaged 6.24 per 100,000 residents, with a standard deviation of 3.57. Certain dataset variables show skewedness and broad value ranges, typically addressed through logarithmic transformation. Here, we applied such transformation to the population aged 15 and over and GDP variables, effectively simplifying analysis and interpretation. Our dataset comprises 341 instances, with the inclusion of a new 'suicide category' attribute to enable prediction. This categorical attribute classifies instances based on the percentage of suicide mortality: 'Low' below 33%, 'Medium' between 33% and 66%, and 'High' above 66%. Calculation precision and multiple author cross-checks ensure attribute accuracy.

Spatial correlations among provinces stem from neighborhood relationships, necessitating spatial considerations for accuracy. A tailored queen-contiguity weight matrix, designed for polygonal data, captures spatial links through shared vertices. Connecting provinces that share at least one vertex enhances spatial relationship understanding. This matrix exploration unveils Iran's 31 provinces' spatial interdependencies, providing a comprehensive perspective on suicide mortality dynamics across regions.

In spatial econometrics, model selection is pivotal. Empirical specification tests, using the specificity-to-generality approach like Lagrange multiplier (LM) tests by Anselin(1988), stand as robust tools against diverse spatial dependence sources. Spatial dependencies signify mutual influence among neighbors, ignoring which biases estimates. LM tests assess spatial autocorrelation in non-spatial models' residuals (e.g., OLS), identifying systematic spatial patterns. A significant outcome suggests model misspecification, indicating unaccounted spatial dependencies. Conducting LM tests reveals spatial autocorrelation presence and extent. Significance underscores a spatial econometric model's suitability, recognizing interconnected neighboring observations. Integrating geographical attributes mitigates positive spatial autocorrelation effects.

Table 2: LM Tests Confirm Strong Spatial Autocorrelation.

| Test | Statistic | P-value |
|---|---|---|
| LMerr | 417.06 | ¡ 2.2e-16 |
| LMlag | 247.38 | ¡ 2.2e-16 |
| RLMerr | 176.3 | ¡ 2.2e-16 |
| RLMlag | 6.6259 | 0.01005 |
| SARMA | 423.68 | ¡ 2.2e-16 |

In Table 2, the LMerr, LMlag, and RLMerr tests yield highly significant p-values (¡ 0.05), indicating robust spatial autocorrelation evidence in residuals and the dependent variable. RLMlag also supports this, emphasizing dependent variable spatial autocorrelation. The SARMA model reinforces significant spatial autocorrelation. In conclusion, LM tests crucially detect spatial autocorrelation, advocating spatial lag term inclusion. Addressing spatial relationships enhances accuracy and robustness, vital for interpreting data's spatial patterns.

# 3 Methodology

This study is designed to undertake a comparative analysis of predictive accuracy between conventional spatial econometric models and novel random forest models in projecting suicide mortality levels within Iran's provinces. To accomplish this objective, our methodology encompasses seven distinctive model specifications, each shedding light on the intricate interplay between spatial dependencies and the dynamics of suicide mortality.

Our spatial econometric models address spatial dependence, where proximity strengthens relationships. Building on Ordinary Least Squares (OLS), these models integrate a spatial weights matrix ($W$) ingeniously (LeSage,2009). The matrix's placement adapts across scenarios, finely tuning for distinct spatial autocorrelation patterns. In contrast, the random forest, an ensemble learning hallmark, amalgamates outputs from decision trees. Consider the Classification and Regression Tree (CART) algorithm, yielding categorical assignment probabilities or average predictions (Breiman,2001).

The crux of our endeavor pivots around the calibration and scrutiny of the following seven model configurations: spatial lag (autoregressive) (SAR), spatially lagged $X$ (SLX), spatial Durbin (SDM), random forest (RF), random forest with the spatial lag of $y$ included (RFSAR), random forest with spatial lags of both $X$ and $y$ included (RFSDM), and random forest with only the spatial lag of $X$ included (RFSLX). Each distinct configuration provides a unique vantage point into unraveling the complex tapestry of spatial dependencies and their role in shaping the trajectory of suicide mortality projection across Iran's provinces.

# 4 Results

Three spatial econometrics models (SAR, SLX, SDM) and four data mining models (RF, RFSAR, RFSLX, RFSDM) were applied for classifying outcomes into Low, Medium, or High categories. Cross-validation evaluated each model's performance rigorously. Table 3 comprehensively assesses various models based on metrics like accuracy, precision, sensitivity, F-score, and specificity. RFSDM stands out as superior.

Table 3: Model Performance Metrics: A Comparative Analysis of Prediction Models for Suicide Mortality Levels.

| Model | Accuracy | Precision | Sensitivity | Fscore | Specificity |
|---|---|---|---|---|---|
| SAR | 0.3659 | 0.3750 | 0.2727 | 0.3158 | 0.4737 |
| SLX | 0.4651 | 0.4500 | 0.4286 | 0.4390 | 0.5000 |
| SDM | 0.4651 | 0.4211 | 0.4000 | 0.4103 | 0.5217 |
| RF | 0.2703 | 0.0769 | 0.4000 | 0.1290 | 0.2500 |
| RFSAR | 0.8250 | 0.4615 | 1 | 0.6316 | 0.7941 |
| RFSLX | 0.8684 | 0.4444 | 1 | 0.6154 | 0.8529 |
| RFSDM | 0.8919 | 0.4286 | 1 | 0.6000 | 0.8824 |

Accuracy measures overall correctness via correctly predicted instances divided by total. Higher values mean better predictions; e.g., RFSLX's accuracy is 0.8684 (86.84%). Precision is the ratio of true positive predictions among all positives. Higher values reduce false positives; RFSDM's precision is 0.4286 (42.86% true positives). Sensitivity (Recall) gauges true positive predictions among actual positives. Higher values lower false negatives; RFSAR's sensitivity is 1 (100% true positives). F-score balances precision and sensitivity. A higher F-score suggests better balance; e.g., RFSAR's F-scores are 0.6316, signifying balanced performance. Specificity assesses true negative predictions among actual negatives. Higher values mean fewer false positives; RFSDM's specificity is 0.8824 (88.24% correct negatives).

The confusion matrix and the out-of-bag (OOB) error plot evaluate random forest performance. The confusion matrix breaks down predictions per class, aiding accuracy, precision, sensitivity, specificity assessment. The OOB error plot displays OOB error against trees. This metric estimates unseen data prediction error. The plot depicts error changes with tree count, helping assess overall performance and optimal tree number.

The OOB error rate trend gradually decreases until it levels off, indicating limited gains beyond a certain point. In our dataset, stability is reached at approximately 0.12 (1.2%) after around 70 trees, beyond which the error remains steady. This trend is visualized in the OOB error plot, aiding the selection of the optimal tree count for RFSDM (Figure 1). Identifying the stabilization point allows confident selection of around 70 trees, striking a balance between capturing patterns and avoiding overfitting. This choice ensures the effectiveness of RFSDM with unseen data while avoiding unnecessary complexity. This analysis offers valuable insights for performance assessment, guiding informed research

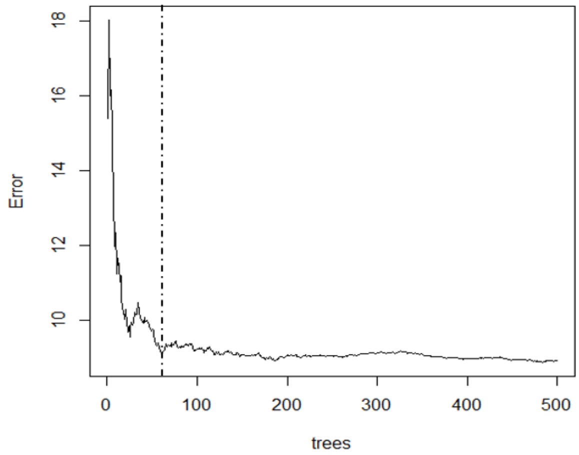


Figure 1: Out-of-Bag (OOB) Error Rate Trend for RFSDM.

decisions.

Table 4 displays confusion matrices for seven models predicting suicide mortality levels (Low, Medium, High). Each model's predicted outcomes are compared to actual observations. Instances per level, correct/incorrect predictions are detailed. SAR, SLX, SDM, RF, RFSAR, RFSLX, and RFSDM models are evaluated, revealing RFSDM's superiority in prediction accuracy and overall performance. It correctly predicts 92 instances30 Low, 29 Medium, 33 Highyet makes 12 incorrect predictions across levels.

Table 4: Confusion Matrix Using the Seven Model.

| Model | Level | Low | Medium | High | Correct | Incorrect |
|---|---|---|---|---|---|---|
| SAR | Low | 9 | 6 | 13 | 33 | 71 |
| | Medium | 10 | 15 | 12 | | |
| | High | 16 | 14 | 9 | | |
| SLX | Low | 11 | 9 | 9 | 36 | 68 |
| | Medium | 11 | 11 | 12 | | |
| | High | 12 | 15 | 14 | | |
| SDM | Low | 12 | 8 | 9 | 37 | 67 |
| | Medium | 11 | 11 | 12 | | |
| | High | 12 | 15 | 14 | | |
| RF | Low | 8 | 2 | 0 | 56 | 48 |
| | Medium | 24 | 25 | 12 | | |
| | High | 3 | 8 | 23 | | |
| RFSAR | Low | 27 | 6 | 0 | 86 | 18 |
| | Medium | 7 | 26 | 2 | | |
| | High | 0 | 3 | 33 | | |
| RFSLX | Low | 29 | 4 | 0 | 90 | 14 |
| | Medium | 5 | 28 | 2 | | |
| | High | 0 | 3 | 33 | | |
| RFSDM | Low | 30 | 3 | 0 | 92 | 12 |
| | Medium | 4 | 29 | 2 | | |
| | High | 0 | 3 | 33 | | |

Table 5 displays the cross tabulation of RFSDM Model's predictions against actual suicide mortality levels. Table cells contain diverse information: observations (N), Chi-square contribution, proportions relative to row total (N / Row Total), column total (N / Col Total), and overall total (N / Table Total). This cross table encompasses 104 dataset observations. Rows represent model-predicted suicide mortality levels, columns actual levels. The rightmost column tallies predicted level observations, while the bottom row provides actual level totals. To glean insights from the table, we delve into each cell's values. Consider the first row: the model forecasted 30 instances as Low, 3 as Medium, and none as High. The row total of 33 signifies total Low predictions. Chi-square contribution values gauge cells' impact on the overall goodness-of-fit Chi-square statistic. By applying similar analysis to other rows and columns, we evaluate the model's accuracy in predicting actual suicide mortality levels. Altogether, this table unveils the model's predictive prowess, showcasing alignment between its predictions and actual data.

Table 5: Confusion Matrix: Predicted vs. Actual Suicide Mortality Levels.

| Categories | Low | Medium | High | Row Total |
|---|---|---|---|---|
| Low | 30 | 3 | 0 | 33 |
| | 34.211 | 5.916 | 11.106 | |
| | 0.909 | 0.091 | 0.000 | 0.317 |
| | 0.882 | 0.086 | 0.000 | |
| | 0.288 | 0.029 | 0.000 | |
| Medium | 4 | 29 | 2 | 35 |
| | 4.841 | 25.178 | 8.118 | |
| | 0.114 | 0.829 | 0.057 | 0.337 |
| | 0.118 | 0.829 | 0.057 | |
| | 0.038 | 0.279 | 0.019 | |
| High | 0 | 3 | 33 | 36 |
| | 11.769 | 6.858 | 36.001 | |
| | 0.000 | 0.083 | 0.917 | 0.346 |
| | 0.000 | 0.086 | 0.943 | |
| | 0.000 | 0.029 | 0.317 | |
| Column Total | 34 | 35 | 36 | 104 |
| | 0.327 | 0.337 | 0.337 | |

# 5 Conclusion

This study delved into the intricate realm of predictive modeling to forecast suicide mortality levels across Iranian provinces. Through a comprehensive analysis, we embarked on a comparative journey, pitting the strengths and challenges of spatial econometric models against those of random forest models. Our exploration encompassed SAR, SLX, and SDM spatial econometric models, each showing potential in capturing certain facets of spatial autocorrelation. However, their overall predictive prowess fell shy of expectations, hinting at the need for refinement and the integration of supplementary spatial insights.

In contrast, the performance of the non-spatial random forest model proved underwhelming, underscoring the indispensability of spatial considerations in predictive modeling. Emerging as potent contenders, the spatial econometric random forest models stepped into the spotlight, overshadowing traditional spatial econometric and non-spatial random forest counterparts. Particularly, the RFSDM model shone brightly, showcasing impressive accuracy, sensitivity, and specificity across diverse metrics.

In parallel with the findings of Credit(2022), our research emphasizes the pivotal role of spatial awareness in predictive modeling. Yet, we venture into uncharted territory by examining panel data and harnessing classification random forest models, enriched with spatial lag. This innovative approach unearths the predictive prowess of these models in classification tasks, unveiling the intricate spatial relationships intrinsic to the data. Our discoveries hold crucial implications for public health policymakers and spatial data analysts alike. The spotlight firmly on RFSDM as the prime model for predicting suicide mortality rates carries immense potential for steering targeted interventions and data-driven public health policies against this pressing concern. However, amid these illuminating findings, our study does acknowledge limitations. The choice of predictor variables and data quality could sway model accuracy. Furthermore, the omission of temporal dynamics in suicide mortality rates warrants dedicated exploration in forthcoming research endeavors.

# References

Anselin, L. (1988). Spatial Econometrics: Methods and Models. Springer Science & Business Media.

Breiman, L. (2001). Random Forests. Machine Learning, **45**, 532.

Credit, K. (2022). Spatial Models or Random Forest? Evaluating the Use of Spatially Explicit Machine Learning Methods to Predict Employment Density Around New Transit Stations in Los Angeles. *Geographical Analysis*, **54(1)**, 58-83.

LeSage, J. P., and R. K. Pace. (2009). Introduction to Spatial Econometrics. New York, NY: CRC Press.

دانشگاه بین المللی امام خمینی (ره) با همکاری
دانشگاه تربیت مدرس و انجمن آمار ایران برگزار می‌کند:
conf.ikiu.ac.ir/spatial5
پنجمین سمینار
آمار فضایی و کاربردهای آن
5th Seminar on Spatial Statistics and Its Applications
۳ و ۴ آبان‌ماه ۱۴۰۲
Oct. 2023 25-26

# Spatio-Temporal Functional Data Analysis of Traffic Offenses in Iran from 2016 to 2023

Mohammad Fayaz*
Allameh Tabataba’i University, Tehran, Iran.

**Abstract:**

Traffic accident in Iran is one of the most important causes of losing years of life and studying risky traffic behavior helps to control and manage it in a proactive way. We estimate the spatio-temporal functional structure of traffic behavior and risky driving patterns of four indices 1) total traffic, 2) speeding, 3) unsafe distance and 4) illegal overtaking in Iran from 2016 to 2023. In this regard, we collect data from more than 2500 count stations near roads. The sandwich smoother for spatio-temporal functional data with hero R package are used in 5 steps 1) Initial smoothing preparation, 2) Assembling spline information, 3) Prepare the data, 4) Enhance the fit and 5) Estimate and Smooth. The results are presented in various maps with quarterly data and summary statistics such as mean squared error (MSE) and correlation (COR) are presented in tables for three resolution scenarios. The best scenario according to them consists of five resolutions 30,60,90,120 and 150 knots.

**Keywords:** Traffic Offenses, Spatio-Temporal, Functional Data, Iran .
**Mathematics Subject Classification (2010): 62H11, 62M30, 91D25**

## 1 Introduction

According to the World Health Organization Report, every 24 seconds someone dies on the road and the speed of vehicles is at the core of road traffic injury problems. They consider Speed management, Leadership in road safety, Infrastructure design, and improvement, Vehicle safety standards, Enforcement of traffic laws, and Survival after a crash are key elements of the ”Save Lives” technical package. The estimated number of injuries and

*Speaker: Mohammad.Fayaz.89@gmail.com

death on roads and highways were about 500,000 and 15,300 people in 2020 in Iran. (WHO Team , 2018)

A lot of studies focused on the deaths and injuries on Iranian roads but only a few of them focus on traffic offenses as a prevention factor such as (Fayaz et al , 2020) estimates the traffic offenses near some important locations, for example, airports and (Fayaz et al , 2022) analysis unusual traffic behavior in holidays like Iranian New year holidays (Noruz) with Functional Data Analysis (FDA) (Ramsay and Silverman , 2005), bivariate generalized additive models (GAM) (Wood , 2017) and Integrated Nested Laplace Approximation (INLA) (Moraga , 2019). The limitation of dates or locations of these studies from one side and the complex structure of spatiotemporal data need new methodologies that work well and fast with large spatio-temporal data. Recently, the hero methodology also worked fast and well with such large datasets. (French and Kokoszka , 2021)

In this regard, we study and estimate the spatio-temporal structure of traffic behavior and risky driving patterns in Iran from 2016 to 2023 with hero methodology.

## 2 Material and Methods

The datasets are recorded hourly and daily and we collected and integrated them from more than 2,500 count stations near the roads (Figure 1). The road lengths are about 197,770 Kilometers in 2020. We summarize data to average quarterly for each station and at least more than 40% of all count stations with no missing values of each province are considered. Four indices are 1) Total traffic, 2) Speeding, 3) Unsafe Distance, and 4) Illegal Overtaking. Each count station may count a vehicle, therefore the numbers are not presented the unique vehicles. For example, on a road from A to B, there exits three count stations A1B, A2B, and A3B. A vehicle that goes from A to B has one count in each count station. Therefore, it can show the traffic congestion points and risky points.

The dataset has a complex pattern and we model it as spatio-temporal (SP) data.(Wikle, Zammit-Mangion and Cressie , 2019) Among the SP methods, we consider a new class of them that has a combination with Functional Data Analysis (FDA) methods. In the FDA, we worked with the functional and curve data instead of each observation and we considered the underlying structures with smoothing methods like B-Spline and dimension reduction methods such as Functional principal component analysis (FPCA) (Ramsay and Silverman , 2005). The non-parametric FDA is also introduced without previous assumptions (Ferraty and Vieu , 2006) and many statistical R packages were developed to do both them, for example, (Febrero-Bande and De La Fuente , 2012) The collection of new FDA methods with spatial and geographical data are published (Mateu and Giraldo , 2021) and we use the generalization of the sandwich smoother for spatio-temporal functional data with hero R package. (French and Kokoszka , 2021). The hero methodology used FDA ideas to reduce the dimension of data in both spatial and temporal dimensions, separately. In this regard, it represents them with basis functions such as B-Splines for temporal patterns and radial basis functions with Wendland covariance function for spatial patterns. The developed penalized spline is Originally from Sandwich Smoother (OSS) (Xiao et al , 2013) and Spatio-Temporal Sandwich Smoother (STSS) was presented in hero. The radial basis function is used instead of the tensor product for smoothing bivariate data. The main benefits of B-splines are Compact support, Easy-to-compute derivatives, and Specifiable parameters related to smoothness (Ramsay and Silverman , 2005). In this regard, the Wendland covariance function was used. (French and Kokoszka , 2021). The Wendland covariance function (French and Kokoszka , 2021):

$$r(h) = \begin{cases} \sum_1^N a_j h^j & 0 \leq h^j \leq \phi \\ 0 & \phi < h \end{cases}$$

$h$ is distance between two points in d-dimensional space, $N$ is the desired degree of the polynomial (Smoothness), $\phi$ defines the support of the function, $\{a_j, j = 1, 2, ..., n\}$ are a set of non-zero coefficients,

The hero methodologies have 1) Initial smoothing preparation, 2) Assembling spline information, 3) Preparing the data, 4) Enhance the fit, and 5) Estimate and Smooth (French and Kokoszka , 2021). We consider three scenarios with five resolutions. The number of knots are 30, 60, 90, 120, and 150 in resolutions 1 to 5, respectively. The first scenario is a combination of resolutions 1,2 and 3, the second scenario is a combination of resolutions 1,2,3, and 4, and the third scenario is a combination of resolutions 1,2,3,4, and 5. The results are compared with Mean Squared Error (MSE) and correlation (COR).

# 3 Results

The model comparisons are presented in Table-1. The best results are obtained with five resolutions. The estimated maps are presented in Figure-2, Figure-3, Figure-4, and Figure-5 for total traffic, speeding, unsafe distance, and illegal overtaking, respectively.

In Table 1, two indices mean squared error (MSE) and correlation (COR) between observations and predictions are presented in three resolution scenarios (A, B, and C) for all four variables. It also compares the two statuses for the response variable: 1) without transformation and 2) with logarithmic transformation. According to the MSE, the best results are obtained in Scenario C and no transformation with five resolutions for Total traffic (61,485,635.36), Speeding (1,858,893.64), Unsafe Distance (7,072,722.76), and Illegal Overtaking (24,685.62). One of the reasons for the high value of MSE is that the numbers itself are very large. Therefore, other summary indexes in percentage are calculated but they were not presented in this paper. But the highest correlation are for responses with logarithmic transformation in scenario C with five resolutions: Total Traffic (80.7%) , Speeding (65.0%), Unsafe Distance (77.9%) and Illegal Overtaking (68.7%).

Table 1: The hero Spatio-Temporal Results

| **Variabes** | **Index** | **Y Transformations** | | | | | |
|---|---|---|---|---|---|---|---|
| | | **No Transforamtion** | | | **Log** | | |
| | | **Resolutions*** | | | **Resolutions*** | | |
| | | **A** | **B** | **C** | **A** | **B** | **C** |
| Total Traffic | MSE | 72,577,477.16 | 66,955,879.15 | **61,485,635.36** | 85,508,904.54 | 78,555,179.85 | 72,989,626.44 |
| | COR | 64.4% | 67.8% | 71.0% | 70.3% | 76.9% | **80.7%** |
| Speeding | MSE | 1,986,524.26 | 1,918,335.87 | **1,858,893.64** | 2,452,734.21 | 2,407,946.87 | 2,364,804.68 |
| | COR | 44.2% | 47.3% | 49.7% | 51.4% | 58.5% | **65.0%** |
| Unsafe Distance | MSE | 8,110,095.41 | 7,599,598.58 | **7,072,722.76** | 10,575,791.63 | 10,044,393.01 | 9,427,749.30 |
| | COR | 58.7% | 62.1% | 65.4% | 69.7% | 74.1% | **77.9%** |
| Illegal Overtaking | MSE | 28,948.15 | 27,080.74 | **24,685.62** | 32,601.03 | 31,994.15 | 30,368.31 |
| | COR | 29.2% | 38.2% | 47.2% | 56.1% | 62.7% | **68.7%** |

*Resolutions: A = 1,2,3 , B = 1,2,3,4 , C = 1,2,3,4,5

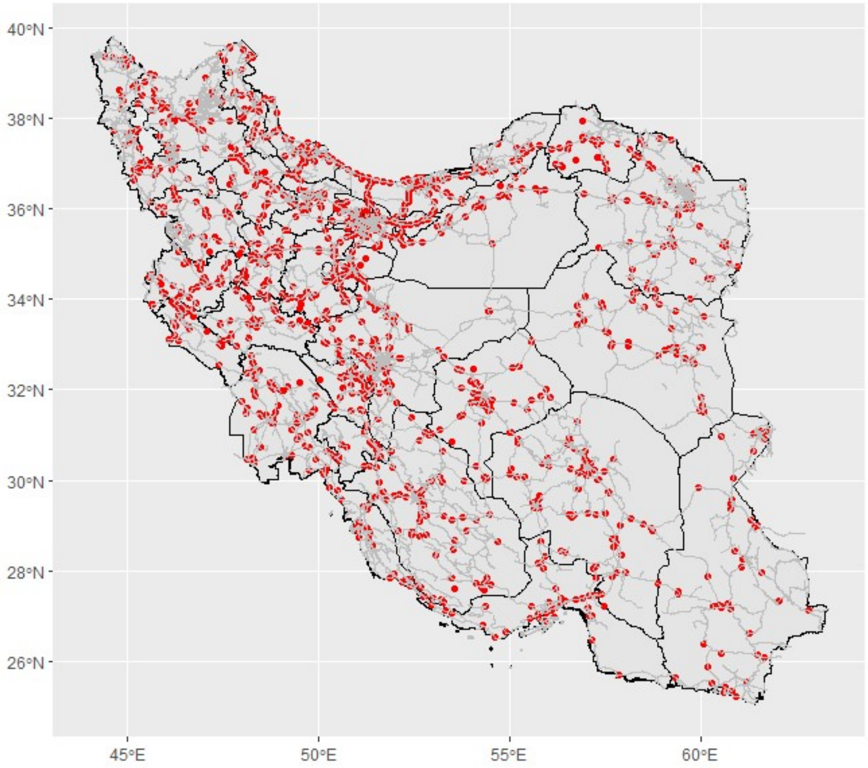


Figure 1: Count Stations (red dots), Roads and Highways (Grey lines) in Iran

In Figure 1, the location of count stations and road and highways and border of provinces are plotted in red dots, Grey lines and black lines, respectively. The prediction

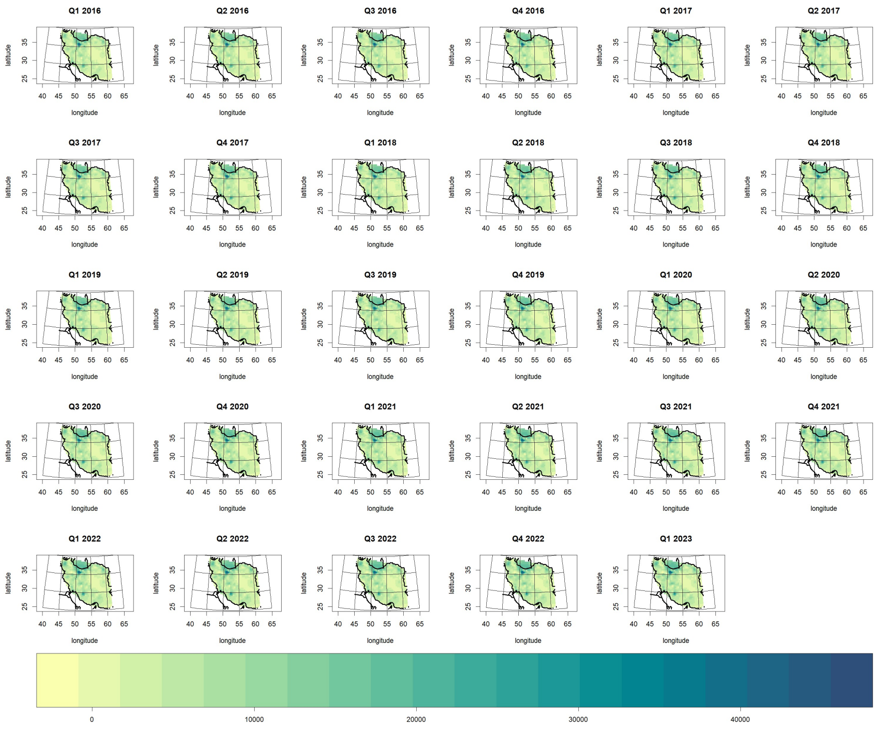


Figure 2: Total of Transportation.

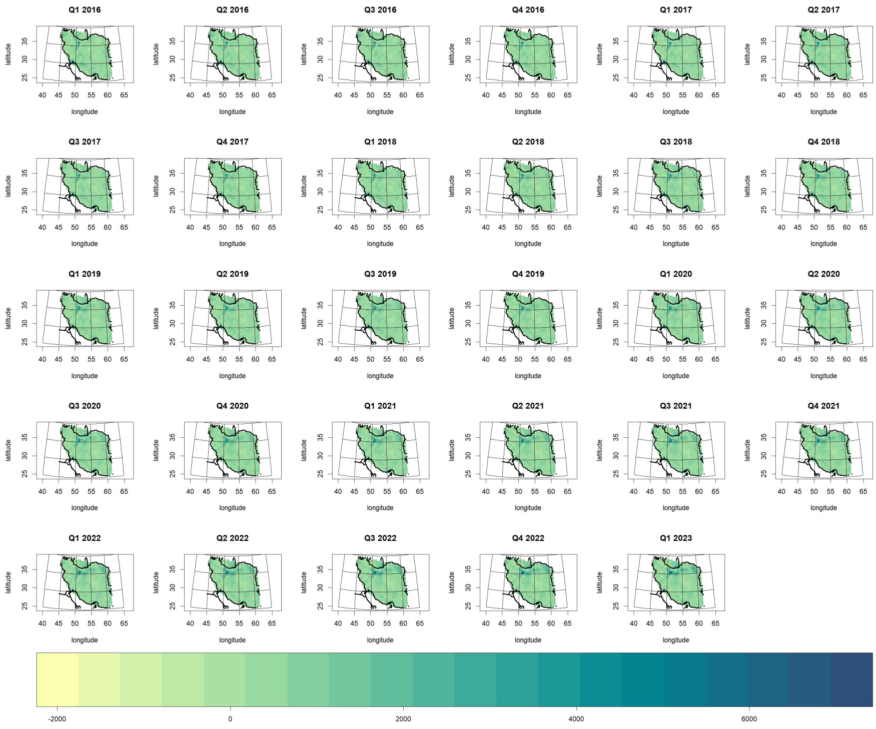


Figure 3: Speeding

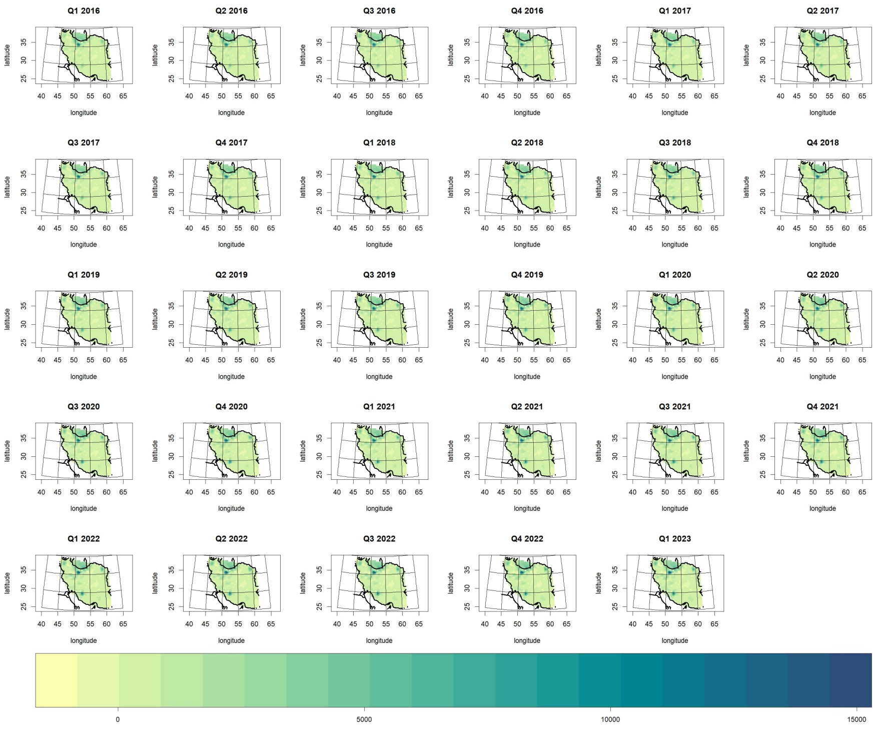


Figure 4: Unsafe Distance

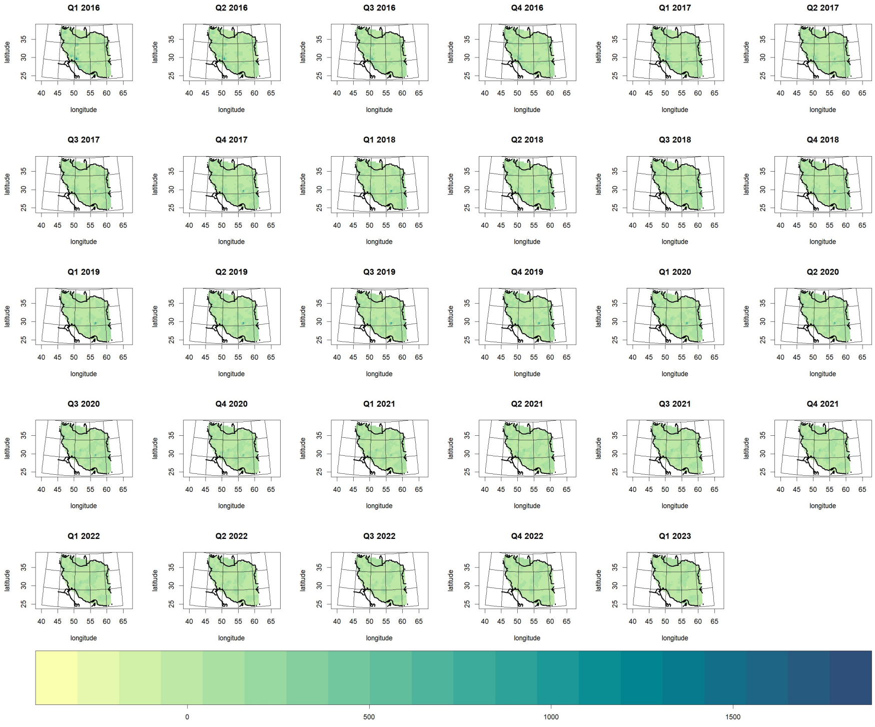


Figure 5: Illegal Overtaking.

results for all four indices are presented in the 2, 3, 4 and 5 from Q1 of 2016 to Q1 of 2023. The color range are presented from low (yellow) to high (dark blue).

## Conclusion

he hero methodology also work fast and well with Iranian traffic behavior dataset and the comparison with other new methods such as Spatio-temporal DeepKriging (Nag et al , 2023), Generalized Spatio-temporal Regression with PDE Penalization (Arnone et al , 2023) etc. are one of the future direction of this research. The driver behavior data is not available and it is the main limitation that can obtained with Surveys, Questionnaires, Mobile Applications, Internet of Things and Vehicle Telematics. Traffic accident in Iran is one of the most important causes of losing years of life and studying risky traffic behavior helps to control and manage it in a proactive way. (Saadat et al , 2022)

## References

Arnone E, Elia C, Sangalli LM(2023). Generalized Spatio-temporal Regression with PDE Penalization. În Classification and Data Science in the Digital Age, Springer. **1:6**.

Fayaz, M.; Abadi, A.R.; Khodakarim, A.; Hoseini, M.R.; Razzaghi, A.R. (2020), THE DATA-DRIVEN Pattern For Healthy Behaviors of Car Drivers Based on Daily Records of Traffic Count Data From 2018 to 2019 Near Airports: A Functional Data Analysis, *JP Journal of Biostatistics*, **17**, 539557.

Fayaz M., Abadi A.R., Razzaghi A.R., Khodakarim S, Hosseini M. (2022), Investigation of the Hourly and Spatial Patterns of Traffic Offenses During March-April 2019 in Iran Using Bivariate Generalized Additive Models and Integrated Nested Laplace Approximation. *International Journal of High Risk Behaviors and Addiction.*, **3**, 11.

Febrero-Bande M., De La Fuente MO. (2012), Statistical Computing in Functional Data Analysis: The R package fda.usc. *Journal of statistical Software*, **51**, 1:28.

Ferraty F. , Vieu P. (2006), Nonparametric Functional Data Analysis, Theory and Practice. *Springer New York.*

French JP, Kokoszka PS. (2021), A Sandwich Smoother for Spatio-Temporal Functional Data. *Spatial Statistics*, **42**, 100413.

Mateu J., Giraldo R. (2021), Geostatistical functional data analysis. *John Wiley and Sons.*

Moraga P. (2019), Geospatial health data: Modeling and visualization with R-INLA and shiny. *CRC press.*

Nag P, Sun Y, Reich BJ (2023), Spatio-temporal DeepKriging for Interpolation and Probabilistic Forecasting. *arXiv preprint arXiv:2306.11472.*

Ramsay J., Silverman B. (2005), Functional Data Analysis. *Springer.*

Saadat S, Yousefifard M, Asady H, Jafari AM, Fayaz M, Hosseini M (2015), The most important causes of death in Iranian population; a retrospective cohort study. *Emergency.*, **3**, 16.

Wikle CK., Zammit-Mangion A., Cressie N. (2019), Spatio-Temporal Statistics With R. *CRC Press.*

World Health Organization (WHO) Team, Social Determinants of Health (2018), Global status report on road safety. *WHO.*

Wood SN. (2017), Generalized Additive Models: An Introduction With R. *CRC press.*

Xiao L, Li Y, Ruppert D (2013), Fast bivariate P-splines: the sandwich smoother. *Journal of the Royal Statistical Society Series B: Statistical Methodology.* **75** 3.

دانشگاه بین المللی امام خمینی (ره) با همکاری
دانشگاه تربیت مدرس و انجمن آمار ایران برگزار می‌کند:
conf.ikiu.ac.ir/spatial5
پنجمین سمینار
آمار فضایی و کاربردهای آن
5th Seminar on Spatial Statistics and Its Applications
۳ و ۴ آبان‌ماه ۱۴۰۲
Oct. 2023 25-26

# Application of Adaptive Lasso Sparisity Identification in Mixed Effects Quantile Regression Models

Forouzan Jafari Maryaki*, Mousa Golalizadeh
Department of Statistics, Tarbiat Modares University, Tehran, Iran.

**Abstract:**

With the development of experimental techniques, one can collect complex structure data in many fields and informations provided by these data are becoming more complicated. A common property of these data sets is that they come from a population with inter-class correlation, which refers to the mixed effects data; the other one is in which the number of variables greatly exceeds the number of samples, then we have high dimensional data. This paper proposes an adaptive lasso approach for the simultaneous selection of mixed effects and also regression coefficients. It is a new approach in variable selection in the mixed effects quantile regression model context by considering the sparsity. Therefore, the present paper proposes a new optimization problem process in this field to shrink the mixed effects and regression coefficients simultaneously. Our simulation experiments show the superiority of the presented method in comparison with lasso penalty in mixed effects quantile regression models.

**Keywords:** Quantile Regression, Adaptive Lasso, Regularization, Variable Selection.
**Mathematics Subject Classification (2010):** 62G08, 62J05, 62J07.

## 1 Introduction

Sometimes, to invoke the regression tools, there is an obligation to utilize the quantiles rather than the mean while analyzing the data from the real-life phenomenon. This can be done with the quantile regression (Koenker and Bassett , 1978), due to the known drawbacks of the regression based upon the mean, i.e., the ordinary regression.

Similar to the ordinary regression, the ordinary quantile regression applied in high dimensional data has a low bias but large variance, leading to the low accuracy. The

*Speaker: f.jafari@modares.ac.ir

regularization is, then, an option to balance between the bias and variance. In the regularization framework, many different types of penalties have been introduced to achieve variable selection. The most famous methods of regularization with convex penalty include the nonnegative garotte (Briman, 1995), ridge regression (Horel and Kennard , 1970) and the lasso (Tibshirani, 1996). The adaptive lasso presented by Zou (2006) is a cherished development of the lasso that allocates adaptive weights for different coefficients in the $L_1$ penalty; enjoying some interesting properties too. Historically, Wang and *et al.* (2007) used the $L_1$ penalty in the median regression model. Adaptive weight in the penalty term, known as the adaptive lasso, was used by Wu and Liu (2009) in the quantile regression and was called the adaptive lasso quantile regression.

As expected, constructing an effective variable selection method in a mixed effects quantile model is a challenging topic. By inducing the role of random effects in the quantile regression models, the within-subject variability is included in these model. This trick prevents obtaining biased estimates for model parameters (Diggle and *et al.* , 2002). On the other hand, if the model contains unnecessary random effects, it will make the covariance matrix singular, which is not conducive to the estimation of unknown parameters (Li and *et al.* , 2020). Therefore, taking into account the impact of the random effects for estimating and selecting the fixed effects is a crucial problem in a mixed effects model. There are very few methods that handle the selection of random effects directly. Scientific reports show that Koenker (2004) is the first to propose the $L_1$ penalized quantile regression model for longitudinal data analysis. However, his proposed method cannot regression modelling in tackle high-dimensional data.

Bondell and *et al.* (2010) proposed an adaptive lasso approach for the simultaneous selection of random and fixed effects. However, their method relied on the mean regression framework. Li and *et al.* (2020) suggested a new algorithm to simultaneously obtain estimates of fixed and the random effects based on lasso penalty by combining the technique introduced by Bondell and *et al.* (2010), Koenker (2004).

This paper proposes an adaptive lasso approach for the simultaneous selection of random and fixed effects. That is a new approach in variable selection in the mixed effects quantile regression model context. This idea with applying adaptive lasso penalty to both fixed and random effects simultaneously in mixed effects quantile regression model can jointly estimate parameters and random effects. Moreover, we present a process to estimate both fixed and random coefficients by considering the sparsity too.

The rest of this paper is organized as follows. In Section 2, we describe mixed effects quantile regression. In Section 3 our proposed method is described in more detail. Also, we provide an algorithm to utilize this new method in the application. The results of the simulation study on comparing the the adaptive lasso with the lasso are reported in

Section 4.

## 2 Model Specification

Suppose that we have $n$ subjects, where the i-th subject has $n_i$ observations. Based on the specific quantile, let's say $\tau$, chosen from (0,1), the mixed effects quantile regression model is often written as

$$y_{ij} = \mathbf{x}_{ij}{}^T\boldsymbol{\beta}_\tau + \mathbf{z}_{ij}{}^T\mathbf{u}_{i\tau} + \varepsilon_{ij\tau}, \qquad j = 1, \dots, n_i, \;\; i = 1, \dots, n, \;\; N = \sum_i n_i, \tag{2.1}$$

where $p$-dimensional covariate vectors $\mathbf{x}_{ij}^T = (x_{ij1}, x_{ij2}, \dots, x_{ijp})$ are row of known design matrix $\mathbf{X}_i$ and $y_{ij}$, is $j$-th observation of continuous random variable on the $i$-th subject. Moreover, in the model (2.1), a $p$-dimensional vector of fixed regression coefficients is $\boldsymbol{\beta}_\tau = (\beta_{0_\tau}, \beta_{1_\tau}, \dots, \beta_{p_\tau})^T$, $\mathbf{z}_{ij}^T = (z_{ij1}, z_{ij2}, \dots, z_{ijq})$ is row of covariate matrix associated with random effects $\mathbf{Z}_i$ and $\mathbf{u}_{i\tau} = (u_{i1\tau}, u_{i2\tau}, \dots, u_{iq\tau})^T$ is a $q \times 1$ vector of random effects. Without loss of generality, we can proceed our discussion through considering a quantile regression model with no intercept via centering the covariates. Throughout this manuscript, the quantile, i.e., $\tau$ is taking its value in (0,1). Hence, to ease repetition, we omit this statement. So, we write $\boldsymbol{\beta}_\tau$, $y_{ij\tau}$ and $\mathbf{u}_{i\tau}$ respectively as $\boldsymbol{\beta}$, $y_{ij}$ and $\mathbf{u}_i$ when there is no confusion.

Here, $\varepsilon_{ij\tau}$ is the model errors usually following the Asymmetric Laplace Distribution $(ALD)$ written as $\varepsilon_{ij\tau} \sim ALD(\mathbf{0}, \sigma_{ij}, \tau)$, for $i = 1, \dots, n$. ALD is comprehensively treated by Koenker and Machado (1999). Considere the following model for the conditional quantile functions of the response of the $j$-th observation on the $i$-th subject:

$$G_{y_{ij}}(\tau|\mathbf{x}_{ij}, \mathbf{u}_i) = \mathbf{x}_{ij}{}^T\boldsymbol{\beta} + \mathbf{z}_{ij}{}^T\mathbf{u}_i, \qquad j = 1, \dots, n_i, \;\; i = 1, \dots, n. \tag{2.2}$$

To estimate the parameters in the model (2.2), one should solve the optimization problem

$$\underset{(\beta, \mathbf{u})}{\arg\min} \sum_{i=1}^{n}\sum_{j=1}^{n_i} \rho_\tau(y_{ij} - \mathbf{x}_{ij}{}^T\beta - \mathbf{z}_{ij}{}^T\mathbf{u}_i) \tag{2.3}$$

where $\rho_\tau(\nu) = \Big[(1-\tau)I(u \leq 0) + \tau I(\nu > 0)\Big]|\nu|$ is called check function and $I(.)$ is indicator function. Let write $\mathbf{y_i} = (y_{i1}, \dots, y_{in_i})^T$, $\mathbf{X}_i = (\mathbf{x}_{_{i1}}, \mathbf{x}_{_{i2}}, \dots, \mathbf{x}_{_{in_i}})^T$, $\mathbf{Z}_i = (\mathbf{z}_{_{i1}}, \mathbf{z}_{_{i2}}, \dots, \mathbf{z}_{_{in_i}})^T$ and also consider the vector of errors as $\boldsymbol{\varepsilon}_i = (\boldsymbol{\varepsilon}_{_{i1}}, \boldsymbol{\varepsilon}_{_{i2}}, \dots, \boldsymbol{\varepsilon}_{_{in_i}})^T$. Further, assumes that $\boldsymbol{u}_i \sim N(\mathbf{0}_{\boldsymbol{q}}, \Sigma_{u_i})$, $i = 1, \dots, n$. The model in (2.1) for $i$-th subject is often written in the following form:

$$\mathbf{y_i} = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i, \qquad i = 1, \dots, n. \tag{2.4}$$

Now, suppose $\boldsymbol{X}_i^* = [\boldsymbol{X}_i, \boldsymbol{Z}_i]$ is a $n_i \times (p+q)$ matrix and $\boldsymbol{\beta}_i^* = \left[\boldsymbol{\beta}^T, \boldsymbol{u}_i^T\right]^T$ is a $(p+q)-$vector. We can re-write the equation (2.4) as

$$\boldsymbol{y}_i = \boldsymbol{X}_i^* \boldsymbol{\beta}_i^* + \boldsymbol{\varepsilon}_i, \qquad i = 1, \dots, n. \tag{2.5}$$

So to achive the estimate of $\boldsymbol{\beta}_i^*$ in (2.5), say $\hat{\boldsymbol{\beta}}_i^*$, one can rewrite (2.3) in the same structure as (2.5) and then solve the optimization problem

$$\arg\min_{\boldsymbol{\beta}^*} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \rho_\tau (y_{ij} - {\mathbf{x}^*}_{ij}{}^T \boldsymbol{\beta}_i^*). \tag{2.6}$$

The estimate of other parameter are done accordingly.

# 3 Penalized Mixed Quantile Regression

## 3.1 Lasso MixedQuantile Regression

In the penalized mixed quantile regression approach, introduced by Koenker (2004), the lasso penalty function is considered as the loss function leading to the general optimization problem

$$\min_{(\mathbf{u},\boldsymbol{\beta})} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \omega_{ij} \rho_\tau \left( y_{ij} - {\mathbf{x}_{ij}}^T \boldsymbol{\beta} - u_i \right) + \lambda \sum_{i=1}^{n} |u_i|. \tag{3.1}$$

As seen, the penalty used in this approach is only a function of random effects and shrinkage is not done based on the fixed parameter. Therefore, we suggest to use the model (2.5), and according to Li and Zhu (2008), define the lasso mixed quantile regression estimates i.e., $\hat{\boldsymbol{\beta}}^*$ as:

$$\arg\min_{\boldsymbol{\beta}^*} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \rho_\tau (y_{ij} - {\mathbf{x}_{ij}^*}^T \boldsymbol{\beta}_i^*) + \lambda_n \sum_{k=1}^{p+nq} |\beta_k^*|. \tag{3.2}$$

where $\boldsymbol{\beta}^* = (\beta_1, \beta_2, ..., \beta_p, u_{11}, u_{12}, ..., u_{nq})^T$ for $\tau \in (0, 1)$.

### 3.1.1 Adaptive Lasso Quantile Regression

It is known that the lasso ignores the effect of randomness of variables in the penalty term. The adaptive lasso, instead, overcomes this drawback and therefore has better performance in the statistical sense (Zou, 2006). Also, it enjoys the oracle properties and also efficiently follows the same algorithm as the lasso does. Mathematically, adaptive weights are determined using the initial estimates, already derived via invoking the ordinary regression method to estimate the regression coefficients. Those weights lead to high precision, turning the adaptive lasso very popular. Let us recall the weighted lasso mixed quantile regression based on a known weights vector $\boldsymbol{w} = (w_1, ..., w_{p+nq})^T$ defined as:

$$\arg\min_{\boldsymbol{\beta}^*} \sum_{i=1}^{n}\sum_{j=1}^{n_i} \rho_\tau (y_{ij} - {\mathbf{x}_{ij}^*}^T \boldsymbol{\beta}_i^*) + \lambda_n \sum_{k=1}^{p+nq} w_k |\beta_k^*|, \tag{3.3}$$

where $\boldsymbol{\beta}_i^* = (\beta_1, \beta_2, ..., \beta_p, u_{i1}, u_{i2}, ..., u_{iq})^T$ and $\boldsymbol{\beta}^* = (\beta_1, \beta_2, ..., \beta_p, u_{11}, u_{12}, ..., u_{nq})^T$. We use idea that is suggested by Zou (2006) and Wu and Liu (2009) and in (3.3), define the weights as $w_k = \frac{1}{|\tilde{\beta}_k^*|^\gamma}$, $k = 1, \ldots, p+nq$ where $\gamma > 0$ and $\tilde{\boldsymbol{\beta}}^*$ is the vector of the initial coefficients taken from the ridge or unpenalized mixed quantile regression.

## 3.2 The Algorithm of The Adaptive Lasso

In this section, we are going to provide an algorithm suggested by Wu and Liu (2009) and Koenker and Mizera (2014) for deriving the estimates arising from the adaptive lasso in mixed effects quantile regression.

**Algorithm 1:**

**Step 1.** Define $\mathbf{x}_{ij}^{**} = {\mathbf{x}_{ij}^*}^T \boldsymbol{w}_i^{-1}$, where $i = 1, 2, ..., n$, $j = 1, 2, ..., n_i$ where $\boldsymbol{w}_i$ is $p+q$-vector that its elements are the weights in adaptive lasso penlty structure.

**Step 2.** Solve the lasso problem for all $\lambda_n$, i.e.,

$$\hat{\boldsymbol{\beta}}_{lasso}^* = \arg\min_{\boldsymbol{\beta}_{\tau_1}^*} \sum_{i=1}^{n}\sum_{j=1}^{n_i} \rho_{\tau_1} (y_{ij} - \mathbf{x}_{ij}^{**T} \boldsymbol{\beta}_{i,\tau_1}^*) + \lambda_n \sum_{k=1}^{p+nq} |\beta_{k,\tau_1}^*|.$$

**Step 3.** Report the estimate as $\hat{\beta}_{alasso}^* = {\boldsymbol{w}^{-1}}^T \hat{\beta}_{lasso}^*$ where $\boldsymbol{w} = (w_1, ..., w_{p+nq})^T$ is $p+nq$-vector.

Below, we sketch a simple proof on why the **Algorithm 1** guarantees a solution to the optimization problem that appeared in **Step 2** We write:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{alasso}^* &= \arg\min_{\boldsymbol{\beta}_{\tau_1}^*} \sum_{i=1}^{n}\sum_{j=1}^{n_i} \rho_{\tau_1} (y_{ij} - {\mathbf{x}_{ij}^*}^T \boldsymbol{\beta}_{i,\tau_1}^*) + \lambda_n \sum_{k=1}^{p+nq} w_k |\beta_{k,\tau_1}^*| \\
&= \arg\min_{\boldsymbol{\beta}_{\tau_1}^*} \sum_{i=1}^{n}\sum_{j=1}^{n_i} \rho_{\tau_1} (y_{ij} - {\mathbf{x}_{ij}^*}^T \boldsymbol{w}_i^{-1} {\boldsymbol{w}_i}^T \boldsymbol{\beta}_{i,\tau_1}^*) + \lambda_n \sum_{k=1}^{p+nq} w_k |\beta_{k,\tau_1}^*| \\
&= \arg\min_{\boldsymbol{\beta}_{\tau_1}^*} \sum_{i=1}^{n}\sum_{j=1}^{n_i} \rho_{\tau_1} (y_{ij} - \mathbf{x}_{ij}^{**T} {\boldsymbol{w}_i}^T \boldsymbol{\beta}_{i,\tau_1}^*) + \lambda_n \sum_{k=1}^{p+nq} w_k |\beta_{k,\tau_1}^*| \\
&= \arg\min_{\boldsymbol{\beta}_{\tau_1}^*} \sum_{i=1}^{n}\sum_{j=1}^{n_i} \rho_{\tau_1} (y_{ij} - \mathbf{x}_{ij}^{**T} \boldsymbol{\beta}_{i,\tau_1}^{**}) + \lambda_n \sum_{k=1}^{p+nq} |\beta_{k,\tau_1}^{**}|.
\end{aligned}$$

The tunning parameter plays a crucial role in penalization problems determining. This is also true while implementing in the mixed effects quantile regression. For selecting the optimal pair of $(\gamma, \lambda_n)$ in the adaptive lasso quantile regression, we use the idea proposed by Zou (2006), which will be discussed in detail in the simulation section.

## 4 Simulation Study

In this section, we evaluate the performance of our method in the mixed effects quantile regression indicated by alasso-MQR and compare it with lasso-MQR methods. To conduct our simulation study, we consider the model given in (2.1) and suppose we have 10 subjescts that every subject has 20 observations i.e., $n = 10$ , $m = 20$. set $\boldsymbol{\beta}_\tau = (3, 1.5, 0, 0, 2, 0, 0)^T$ and the covariates, $\mathbf{x}_{ij}^T = (x_{ij1}, x_{ij2}, \dots, x_{ij8})$ and $\mathbf{z}_{ij}^T = (x_{ij1}, x_{ij2}, \dots, x_{ij5})$, $i = 1, 2, \dots, n$, $j = 1, 2, \dots, m$ are generated as $i.i.d$ samples from the multivariate normal density with the mean vector zero and the correlation between each pairs of the predictor variables, $x_{ijk}$ and $x_{ijl}$ through the expression $cor(x_{ijk}, x_{ijl}) = (0.5)^{|l-k|}$, $1 \leq l, k \leq 8$. We also set $\sigma$ equal to 1, 3 and 6 where the corresponding $SNR$s are 21.25, 2.35 and 0.59 respectively. Finally, we use $\mathbf{u}_{i\tau} = (u_{i1\tau}, u_{i2\tau}, \dots, u_{iq\tau})^T \overset{iid}{\sim} N_5\left(\mathbf{0}, D\right)$ and $D = diag(2, 2, 2, 0, 0)$ and $\varepsilon_{ij} \overset{iid}{\sim} N\left(0, \sigma\right)$.
We consider nine different scenarios via altering the relevant and influential parameters. To compare different methods and scenarios, we use the Relative Prediction Error (RPE) based on a distance constructed by invoking the check function $RPE = \frac{E[\rho_\tau(\hat{y} - \mathbf{X}^{*T}\beta^*_\tau)]}{\sigma^2}$. As is common in invoking the linear quantile regression models, we use the estimates of coefficients offered by the unpenalized quantile regression model as the initial values for the weights of the adaptive lasso. We obtain the estimates after fitting two methods using the algorithm proposed by Koenker and Mizera (2014), then extended by Sherwood and *et al.* (2017) and freely available in the *rqPen* package. We consider a set of feasible values for each method for $\lambda_n$, $\gamma$ appeared in the adaptive lasso. These values for $\lambda_n$ and $\gamma$ are $\{0.001, 0.002, ..., 2\}$ and $\{0.1, 0.2, ..., 2\}$, respectively. Also, to concentrate on particular quantile, we set $\tau$ to 0.25, 0.5 and 0.75 in each individual investigation. To evaluate the accuracy of the RPE, its standard errors were also computed through a bootstrap scheme. The standard deviation of these medians, let us call this the Monte Carlo *sd*, was reported as the estimated standard error of the RPEs. In Table 1, we show the values of RPEs and their standard errors (in bracket), after fitting the model set in the simulation setup using the lasso, adaptive lasso (alasso) on the simulated data. In each scenario, the selected methods correspond to the columns highlighted by the bold faces. According to the results reported in Table 1, it can be seen that our proposed method outperforms two alternatives in terms of the RPE measure in this particular simulation

setting. But, it might not suffice to make a decision just in terms of the RPE. Hence, we reported the results in terms of the bias computed as discussed above. The results are also summarized in Table 1. As seen, the adaptive lasso has the better performance than the lasso in some cases here. It sounds that it is weak when $\sigma$ is 6. But, generally, we can assert that in the mixed effect quantile regression considered in this paper, the adaptive lasso penalty is doing better than the lasso penalty in our simulation study.

Table 1: The values of RPEs and their standard errors (in bracket) in lasso and adaptive lasso penalty using the model and scenarios discribed in the text.

| $\tau$ | $n$ | *Lasso* | *ALasso* |
|---|---|---|---|
| 0.25 | 1 | 2.666(0.0058) | **2.653**(0.0052) |
| | 3 | 0.263(0.0011) | **0.261**(0.0015) |
| | 6 | **0.082**(0.0004) | 0.084(0.0005) |
| 0.5 | 1 | 2.178(0.0071) | **2.175**(0.0081) |
| | 3 | 0.258(0.0014) | **0.257**(0.0012) |
| | 6 | **0.082**(0.0004) | 0.083(0.0006) |
| 0.75 | 1 | 2.212(0.0055) | **2.210**(0.0054) |
| | 3 | 0.253(0.0015) | **0.252**(0.0011) |
| | 6 | **0.082**(0.0005) | 0.083(0.0005) |

# Conclusion

The purpose of this paper was to use adaptive lasso penalty in the mixed effects quantile regression models, which is our research innovation. The efficiency of lasso and adaptive lasso penalty was compared and it was shown that adaptive lasso penalty compared to lasso in these models has a better performance. A proposed technique was used to select and estimate the fixed and random effects simultaneously, which has not been done in previous researches. It can be investigated as a topic for future research to use a penalty, which is a function of the check function in the mixed effects quantile regression models that can work better than the adaptive lasso penalty.

# References

Bondell, H. D, Krishna, A. and Ghosh, S. K. (2010). Joint Variable Selection for Fixed and Random Effects in Linear Mixed-Effect Models. *Biometrics*, **66**, 10691077.

Breiman, L. (1995), Better Subset Selection Using Nonnegative Garrote,*Techonometrics*, **37**, 373384.

Diggle, P. J., P. Heagerty, K. Y. Liang, and S. L. Zeger. (2002). *Analysis of Longitudinal Data*, 2nd ed. Oxford: Oxford University Press.

Hoerl, Arthur E., and Kennard, Robert W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, **56(12)**, 5567.

Koenker, R., and Bassett, G. (1978). Regression Quantiles. *Econometrica*, **46**, 33-50.

Koenker, R. (2004). Quantile Regression for Longitudinal Data. *Journal of Multivariate Analysis*, **91**, 74-89.

Koenker, Roger., and Mizera, Ivan. (2014). Convex Optimization in R, *Journal of Statistical Software*, **60(5)**, 123.

Koenker, Roger., and Machado, Jose AF. (1999). Goodness of Fit and Related Inference Processes for Quantile Regression, *Journal of the American Statistical Association*, **94(448)**, 1296-1310.

Li, Youjuan., and Zhu, Ji. (2008). L1-norm Quantile Qegression, *Journal of Computational and Graphical Statistics*, **17(1)**, 163185.

Li, H., Liu, Y. and Luo, Y. (2020). Double Penalized Quantile Regression for the Linear Mixed Effects Model. *Journal of Systems Science and Complexity*, **33**, 20802102.

Sherwood, B., and Maidman, A. (2017). rqPen: Penalized quantile regression. *R package version*, **2**.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, **58**, 26788.

Wang, Hansheng., Li, Guodong., and Jiang, Guohua. (2007). Robust Regression Shrinkage and Consistent Variable Selection through the LAD-Lasso, *Journal of Business and Economic Statistics*, **25(3)**, 347355.

Wu, Yichao., and Liu, Yufeng. (2009). Variable Selection in Quantile Regression, *Statistica Sinica*, **19(2)**, 801817.

Zou, H. (2006). The Adaptive Lasso and its Oracle Properties, *Journal of the American Statistical Association*, **101**, 14181429.

# Predicting the Performance of Trial Court Administration Using Machine Learning

Meisam Moghimbeygi[1] *, Mohadeseh Alsadat Farzammehr[2]
[1]Department of Mathematics, Faculty of Mathematics and Computer Science,
Kharazmi University, Tehran, Iran.
[2]Iranian Judiciary Research Institute, Tehran, Iran.

**Abstract:**

Traditionally, empirical indicators have been generated through methods like expert surveys, document reviews, administrative data analysis, and public surveys. However, this paper utilizes machine learning techniques to predict trial court performance using key indicators for trial case processing. The study uses a dataset collected from 18 civil branches within a trial court in Tehran, Iran, with a sample size of 119 case management data. Logistic Regression was found to be the most effective data mining model, achieving an area under the curve (AUC) of 98.5% and classification accuracy (CA) of 95.0%. The logistic regression analysis revealed that the probability of positive performance evaluation was influenced by factors such as the number of resolved cases. In contrast, the number of pending cases at the beginning of a period had minimal impact. Evaluating trial court administration is crucial for identifying and addressing negative performance issues early on, which helps build public trust and confidence in the justice system. Regular performance evaluations can also contribute to developing a decision support system that enhances overall court performance.

**Keywords:** court performance prediction; data mining; judicial data; machine learning techniques; artificial intelligence.

**Mathematics Subject Classification (2020):** 62P99, 62H30.

## 1 Introduction

Court administration performance and reliable performance indicators are very important in a well-functioning justice system. Court administration performance includes the

*Speaker: m.moghimbeygi@khu.ac.ir

efficiency and effectiveness of court processes, while reliable performance indicators are crucial in evaluating and monitoring court administration performance. These indicators provide accountability, transparency, and opportunities for continuous improvement in court procedures.

Using machine learning to predict court administration performance is significant for several reasons. Machine learning can analyze large amounts of data quickly and accurately, identifying patterns and trends that may not be apparent to human analysts. It can automate the performance measurement and prediction process, reducing the workload of court administrators and facilitating informed decision-making. Machine learning can also improve the accuracy of performance predictions, leading to more efficient resource allocation and better outcomes for court users. Additionally, machine learning can build predictive models that explore the relationships between court administration performance and other factors, helping court administrators address key factors affecting performance. Several performance indicators are used to measure court administration performance in the Iranian court system, particularly in trial case processing. These indicators include the average time for a case to be heard and decided, clearance rate, pending caseload, number of cases resolved, and trial duration. Analyzing these indicators allows court administrators to assess the efficiency and effectiveness of the court system and identify areas for improvement. Machine learning can be employed to predict court administration performance based on these indicators, optimizing resource allocation in the court system (DeMatteo et al. , 2010; Islam et al. , 2017; Martin , 2019).

## 1.1 Materials and Methods

The Iranian justice system collects data on various aspects, which can be communicated using justice indicators. These indicators are effective tools for assessing performance, identifying issues, setting benchmarks, monitoring progress, and evaluating policy effectiveness. Using justice indicators and other monitoring mechanisms ensures transparency and accountability in the functioning of the Iranian justice system while offering policymakers and reformers essential feedback to inform decision-making.

Table 1 outlines the study objectives implemented to forecast trial court performances efficiently. This research primarily focuses on performance prediction by utilizing state-of-the-art machine learning algorithms for a judicial complex in Tehran, Iran. By incorporating advanced techniques, we aim to provide in-depth insights into the performance of the Iranian justice system, enabling more informed decisions to improve its functioning. Ultimately, this study aims to contribute to the betterment of the Iranian justice system through data-driven analysis and informed policy implementations.

We leveraged the power of machine learning to formulate a system that accurately

Table 1: Variability Analysis of Civil Branches in Trial Court.

| Variable | Average | Min | Max | SD | CV |
|---|---|---|---|---|---|
| The number of working days in a month (trial courts typically have around 20 to 22 working days in a month; In Tehran, courts generally maintain a five-day workweek with around 20 working days in a month.) | 19.89 | 18 | 21 | 1.21 | 0.06 |
| Pending cases at the beginning of a time period (courts maintain records of the number of cases pending at the beginning of each month or quarter.) | 599.59 | 0 | 1154 | 282.46 | 0.47 |
| The number of cases referred to a court judge during a particular period of time (the total number of cases that have been formally submitted to the judge for a decision during a specific period.) | 180.25 | 8 | 246 | 42.93 | 0.24 |
| The number of resolved cases during a period (the total number of cases that have been brought to a conclusion during a specific period.) | 171.15 | 14 | 277 | 61.10 | 0.36 |
| The pending trial caseload (the number of cases that have been committed for trial but have not yet been finalized or resolved.) | 135.10 | 3 | 251 | 43.62 | 0.32 |
| The number of precautionary/ monitoring time (the length of time that a case is placed on hold or paused by a judge while additional investigation, evidence gathering, or legal procedures are carried out. This period can also be called a pre-trial period or trial adjournment.) | 158.97 | 0 | 778 | 101.15 | 0.64 |
| Processing time (in a legal context, refers to the period elapsed between a case being ready to be listed for trial, and the earliest date it can be scheduled for trial or expedited hearing.) | 87.80 | 4 | 1589 | 199.99 | 2.28 |
| Precautionary/monitoring time period (a temporary pause of the trial process by the judge for a particular case. During this period, the judge temporarily suspends the proceedings for the purpose of allowing sufficient time for conducting further investigation, gathering additional evidence, or completing legal procedures to guarantee that the decision is based on a comprehensive understanding of the case facts and information.) | 87.53 | 7 | 2624 | 255.25 | 2.92 |
| The final decision number (the number of verdicts, orders, judgments, or rulings issued by a court or a judge within a specified period.) | 168.59 | 14 | 277 | 60.43 | 0.36 |
| The average entry processing time (the total time taken from the last case registered for trial to its finalization in a court proceeding.) | 66.72 | 8 | 203 | 34.86 | 0.52 |

predicts court administration performance based on historical judicial data. Implementation of advanced techniques like support vector machines, k-nearest neighbours, and naive Bayes were utilized to achieve our objective. However, traditional machine learning classifiers' encoding methods fail to capture the intricate relationships between predictor variables in a machine learning-based dataset, which can limit their ability to forecast judicial courts effectively using judicial data only. Our system adopts the widely used supervised learning approach, which requires the system to receive input data and their corresponding labels during the training phase. During this stage, the system detects patterns and relationships between the input and output data to make accurate predictions. With this system, we aim to enhance decision-making, streamline resource allocation and improve court administration by leveraging accurate predictions and valuable insights from historical data. After training, our system undergoes an evaluation phase, assessed with similar, non-utilized data. The model predicts labels for each document, and because each label in this instance represents a court administration performance, our system's purpose is to forecast court performance. With the aid of pattern recognition in historical data, our system can provide data-driven insights that enable informed decision-making and performance evaluation in trial court administration. The typical method of assessing a classification system's performance is by using accuracy or the F1-score. Accuracy measures the number of correctly classified labels, while the F1-score measures the harmonic mean of precision and recall. Precision evaluates the accuracy of the assigned court performance, while recall measures the proportion of correctly classified cases with a specific outcome. These metrics facilitate the effective measurement and identification of areas that need further improvement to enhance overall accuracy and forecasting capability.

# 2 Result

Ten different data mining models were employed to classify the outcome into positive or negative using ten independent variables (detailed in Table 1). The models included Neural Network (NN), Naive Bayes (NB), Adaptive boosting (AdaBoost), Gradient Boosting (GraBoost), Random Forest (RF), Classification Tree (Tree), k-nearest neighbours (kNN), Stochastic Gradient Descent (SGD), Support Vector Machine (SVM), and Logistic Regression (LR).

Several performance metrics were employed to evaluate the models' performance in terms of classification accuracy, including area under the curve (AUC), classification accuracy (CA), F1-score, precision, and recall. Linear regression was not considered, given that the analysis focuses on classification rather than regression. Based on AUC and CA, the best-performing models are ranked in decreasing order, as supplied in Table 2.

Table 2: Performance metrics of the ten data mining models

| Model (average over classes) | AUC | CA | F1 | Precision | Recall |
|---|---|---|---|---|---|
| Logistic Regression | 0.985 | 0.950 | 0.950 | 0.950 | 0.950 |
| Gradient Boosting | 0.955 | 0.899 | 0.899 | 0.899 | 0.899 |
| Neural Network | 0.933 | 0.882 | 0.881 | 0.883 | 0.882 |
| SGD | 0.932 | 0.941 | 0.941 | 0.944 | 0.941 |
| SVM | 0.929 | 0.824 | 0.820 | 0.827 | 0.824 |
| Random Forest | 0.848 | 0.748 | 0.747 | 0.747 | 0.748 |
| Tree | 0.834 | 0.849 | 0.849 | 0.850 | 0.849 |
| kNN | 0.825 | 0.765 | 0.763 | 0.763 | 0.765 |
| Naive Bayes | 0.816 | 0.731 | 0.732 | 0.733 | 0.731 |
| AdaBoost | 0.780 | 0.773 | 0.775 | 0.785 | 0.773 |

Additionally, Table 2 shows the outputs when the target class is averaged over classes.

In summary, the statistical analysis used ten different data mining models to classify outcomes as positive or negative using ten separate variables. Through the application of several performance metrics, including AUC, CA, F1-score, precision, and recall, the models were assessed, and the best-performing models were identified. Linear regression was ruled out, given the analysis's emphasis on classification rather than regression.

The analysis indicated that Logistic Regression was the best-performing model across all three cases, achieving an AUC and CA of 98.5% and 95.0%, respectively. However, when the target classification is 'negative', the sensitivity is lower when compared to classifying 'positive'. Notably, the model performed better for the 'positive' target class and worse for the 'negative,' possibly due to unequal class sizes. While the AUC and CA values were uniform across all three cases, there were notable differences in F1-score, precision, and recall values. Similar differences were observed with the other nine models evaluated in this study. Overall, the results suggest that Logistic Regression is the best model for the classification task, with consistent AUC and CA performance across all three target classes. However, there were variations in the model's ability to predict 'positive' and 'negative' target classes, indicating the need to further explore the class imbalance.

In evaluating the model's effectiveness, a confusion matrix was employed to classify instances of classification and presented in Table 3. The matrix comprises True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN). The Logistic Regression and SGD models correctly classified 113 out of 119 instances, with only 6 out of 119 misclassified. Generally, the number of models in which the number of false positives is lower than the number of false negatives is equal to the number of models in which the number of false positives is more or equal to the number of false negatives. It means that Type I errors and Type II errors are almost equal.

Regarding the best-performing model, Logistic Regression's success can be attributed

Table 3: Classification instances

| Model | TP | FP | FN | TN | Correct | Incorrect |
|---|---|---|---|---|---|---|
| SGD | 69 | 5 | 1 | 44 | 113 | 6 |
| Logistic Regression | 67 | 3 | 3 | 46 | 113 | 6 |
| SVM | 64 | 15 | 6 | 34 | 98 | 21 |
| Naive Bayes | 53 | 15 | 17 | 34 | 87 | 32 |
| Neural Network | 65 | 9 | 5 | 40 | 105 | 14 |
| KNN | 58 | 16 | 12 | 33 | 91 | 28 |
| Gradient Boosting | 64 | 6 | 6 | 43 | 107 | 12 |
| Tree | 60 | 8 | 10 | 41 | 101 | 18 |
| Random forest | 58 | 12 | 12 | 37 | 95 | 24 |
| Adaptive Boosting | 52 | 9 | 18 | 40 | 92 | 27 |

to its utilization of the new feature introduced in Orange Software, as demonstrated in Fig. 1. In summary, the confusion matrix assessed the model's effectiveness and classified instances into TP, FP, FN, and TN. Logistic Regression outperformed other models, achieving high classification accuracy and minimal misclassification. All models demonstrated lower false positive rates than false negative rates, correlating with fewer Type I and more Type II errors.

As depicted in Fig. 1, the key predictor of positive court performance was the number of resolved cases during a specified period. The number of cases referred to a court judge also played a significant role in predicting positivity. Since, red colour represents higher feature value, while blue colour is a lower value and the positive points (points right from the centre) in Fig. 1 are feature values with the impact toward the prediction for the selected class, Obviously, Increasing the number of resolved cases and reducing the number of referred cases leads to an increase in the performance of the courts. Fig. 2 can help us to determine which features most contributed to the prediction (features with longer tape length) and how they affect it. So, the number of resolved cases during the specified period emerged as the key contributor to increase the probability of positive court performance. In other word, as the number of resolved cases increases, the probability of positive court performance also tends to increase. Also, the probability of positive court performance tends to decrease with increasing the number of referred cases. The average probability of positive court performance in this dataset (baseline probability) is 0.52

# 3 Conclusion

In summary, our paper utilized machine learning techniques to classify court performances by analyzing past court behaviour. Although the broad field of automatic legal analysis has a lengthy history, we focus solely on machine learning in this study. Our results

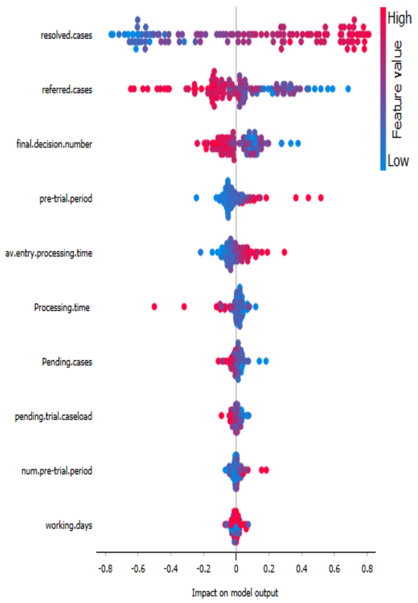


Figure 1: The ranking of the impact of the variables obtained using logistic regression model.

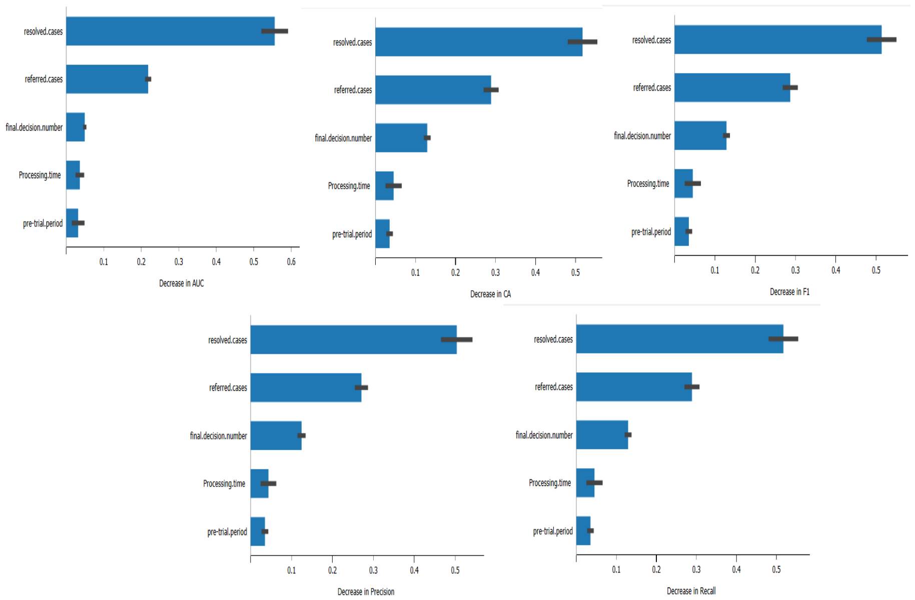


Figure 2: Features Importance based on all AUC, CA, F1-score, precision, and recall scores in logistic regression model.

show that machine learning techniques, specifically logistic regression, Gradient Boosting, neural network, and Stochastic Gradient Descent (SGD) models, effectively predict court performance positivity using ten indicators for data collected from a trial court department of the Judiciary of Tehran jurisdictions, Tehran, Iran. The data mining models showed varying levels of classification accuracy, with Logistic Regression outperforming the others.

Our study adds to the growing support for data mining models' use in detecting court performance, which could be used to develop decision support systems that enhance the positivity rate of monitoring and evaluating court performance in Iran, increasing public confidence in the judicial system. However, it's worth noting that disparities in the data could affect the results' accuracy, such as differences in features, variables, and model type used. While the precision of our data mining models could have been better with a more significant amount of data, this study was limited by the quantity available to us. Properly managing historical data in jurisdictions is highly recommended for researchers seeking to improve the precision of data mining models' court performance predictions. In conclusion, our study contributes to the ongoing effort to enhance the monitoring and evaluation of court performance in Tehran, Iran. We hope that our findings will inspire further research on the range of possible applications of machine learning techniques in the legal field, which could provide deeper insights into predicting court performance with greater precision, ultimately aiding in the administration of justice.

# References

DeMatteo, D., Forst, B., Rose, V., & Mellgren, A. (2010), Performance measurement in court administration: Developing a balanced scorecard for court leaders, *Journal of court innovation*, 3(**3**), 39-59.

Islam, H., Sharghi, M., & Gharibpour, F. (2017), An international comparison of judicial case management: Lessons for Iran, *Journal of Judicial Administration Studies*, 21(**3**), 15-40.

Martin, E. C. (2019), The use of predictive analytics in court decision-making, *Executive Journal*, 19(**2**), 41-57.

# Penalized Pairwise Likelihood Estimation for Spatial GLM Models

Mohsen Mohammadzadeh*, Leyla Salehi
Department of Statistics, Tarbiat Modares University, Tehran, Iran.

**Abstract:**

In this article, we utilized pairwise and weighted pairwise likelihood functions to estimate the parameters of Spatial Generalized Linear Mixed (SGLM) models. Subsequently, we applied the penalized pairwise likelihood function to enhance the accuracy of parameter estimation for the model. In a comprehensive simulation study, we assessed and compared the accuracy of parameter estimations achieved through the pairwise, weighted, and penalized pairwise likelihood, using the mean squared error as the evaluation criterion. Next, we employed the penalized pairwise likelihood method to analyze a real dataset. Finally, the discussion and results are presented.

**Keywords:** Generalized linear mixed model, Penalized pairwise likelihood, Composite likelihood.
**Mathematics Subject Classification (2020):** 62H11, 62M30, 62J12.

## 1 Introduction

Generalized linear models were first introduced by Nelder and Wedderburn (1972), while McCullagh (1989) employed these models to modeling discrete response variables. To represent the correlation of spatial responses, a Spatial Generalized Linear Mixed (SGLM) model can be used. Unlike linear models, the likelihood functions for SGLM models do not offer a closed form owing to the non-Gaussian nature of the response variable; the parameters can not, hence, be estimated using the maximum likelihood method. Therefore, most articles accept the assumption of latent variables' normality and provide a solution to estimate model parameters and latent variables by maximizing likelihood functions,

*Speaker: mohsen_m@modares.ac.ir

penalized quasi-likelihood, or hierarchical likelihood using numerical methods. Among others, McCulloch (1997) utilized maximum likelihood algorithms for GLM models with non-spatial random effects using numerical methods such as Monte Carlo Expectation Maximization (MCEM). Recent studies have examined other approximate methods, which are not based on the complete likelihood of observations. Compared to the approximate likelihood methods, the advantage of these methods is the presumed lack of need for simultaneously modeling all the observations. Wedderburn (1974) employed quasi-likelihood functions, a subclass of composite likelihood methods. Varin *et al.* (2005) used the pairwise composite likelihood method for SGLM models, in which a novel EM algorithm that uses numerical quadrature was introduced. Bevilacqua *et al.* (2010) used the weighted likelihood function for the spatio-temporal data and proved it is a good approximation of maximum likelihood. We employed the pairwise likelihood function for SGLM models, following which the weighted pairwise likelihood function was developed and hence used to estimate the parameters of the models. Moreover, the penalized pairwise likelihood function was used to increase the accuracy of the model parameters estimations. In a simulation study, the accuracy of the model parameter estimations using pairwise likelihood, weighted pairwise likelihood, and penalized pairwise likelihood was evaluated and compared using the Mean Squared Error parameter. Finally, the penalized pairwise likelihood method was used to analyze two real data sets.

## 2 Spatial Generalized Linear Mixed Models

Let $Y(s)$ be a discrete spatial response variable, $Z_1(s), \ldots, Z_p(s)$ are covariates and $\{X(s), s \in \mathbb{R}^2\}$ is a latent spatial random field, where $X(s)$ is a random effect at location $s$. Diggle *et al.* (1998) defined a SGLM model as:
(a) Let $\{X(s), s \in \mathbb{R}^2\}$ be a zero mean stationary Gaussian random field with spatial covariance function $C(h;\theta) = Cov(X(s+h), X(s))$, where $\theta \in \mathbb{R}^k$ is the correlation parameter.
(b) Given $\{X(s), s \in \mathbb{R}^2\}$, $Y(s)$ is a set of independent random variables and the distribution of $Y(s)$ characterized by the conditional mean $E[Y(s)|X(s)]$.
(c) For every link function $g$ and regression parameters, we have $g\{E[Y(s)|X(s)]\} = \sum_{j=1}^{p} Z_j(s)\beta_j + X(s)$.
(d) Conditional distribution of $[Y(s)|X(s)]$ belongs to the exponential family.

## 3 Weighted Pairwise Composite Likelihood Function

The composite likelihood function is obtained by multiplying a set of likelihood components, in which each likelihood component represents a subset of observations. Inference based on the likelihood function for particularly voluminous data is associated with integrating and inverting high-dimensional matrices, which may be difficult to solve even for more powerful computers. Using the composite likelihood function, these multiple integrals can be converted into the sum of integrals with lower dimensions. An example of the composite likelihood function is the weighted pairwise composite likelihood function. Bevilacqua *et al.* (2010) applied the weighted likelihood function to space-time data and showed that composite likelihood weighted estimators are consistent and asymptotically Gaussian with a variance equal to the inverse of the Godambe function. They also showed with simulation studies that this estimate approximates the maximum likelihood estimate and requires far fewer calculational overloads than the maximum likelihood and composite likelihood estimates. Joe and Lee (2009) applied the weighted composite likelihood function to categorise data with a variable number of categories and examined the asymptotic relative efficiency measure for different weights. The results implied that weighting the composite likelihood function increases the asymptotic relative efficiency. The logarithm of the weighted pairwise composite likelihood function can be represented as follows

$$\ell_w(\theta; y) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_{ij} \log f(y_i, y_j; \theta). \tag{3.1}$$

where

$$w_{ij} = \begin{cases} 1 & ||s_i - s_j|| \leq d_s, \\ 0 & \text{O.W}, \end{cases}$$

here $d_s$ represents the distance between spatial points. As such, the weighted pairwise composite maximum likelihood estimation of $\theta$ that maximizes the function (3.1) under regularity conditions is equal to the unique solution of the equation $u(\theta; y) = \bigtriangledown_\theta \ell_w = 0$. The pairwise likelihood function for SGLM models is as follows:

$$L_W(\boldsymbol{\eta}; \mathbf{y}) = \prod_{(i,j)\in\chi} L(\eta|y_i, y_j) \propto \prod_{(i,j)\in\chi} \int\int f(y_i|x_i) f(y_j|x_j) f(x_i, x_j|\eta) dx_i dx_j, \tag{3.2}$$

where $\chi$ is the pairwise neighborhood set of $(y_i, y_j)$.

## 3.1 Penalized Pairwise Likelihood Function

The maximum likelihood method in parameter estimation in various problems is sometimes plagued with overfitting, low accuracy, or high variance of the estimators. Penalization of the likelihood function is a solution to alleviate the behavior of estimators mentioned above by the usual maximum likelihood method, called the penalized maximum likelihood method (Azzalini and Valle , 2013). For extensive spatial data, where obtaining the likelihood function analytically is extremely difficult, the composite likelihood function can be employed to estimate the parameters. The composite likelihood function approximates the likelihood function of observation, and hence estimators based on this function may have low accuracy. The penalized likelihood function can be, as such, used to improve the estimation accuracy in SGLM models, in which a penalty function is embedded in the logarithm of the likelihood of observations. For SGLM models, the penalized pairwise likelihood function is given by $\ell(\eta; y) = \sum_i \sum_{j>i} \log f(y_i, y_j; \eta) - \lambda J(\eta)$, where $\lambda$ is the smoothness parameter and $J(\eta)$ is the penalty function, which can be selected using various methods. For example, Tibshirani (1996) presented the Lasso penalty as $J_\lambda(\eta) = \lambda\eta$ for estimation in linear models. Here, Lasso and Green (1990) $(2\eta\eta^T)$ functions were used for penalization where $\eta = (\beta_0, \beta_1, \sigma^2, \phi)$ is the vector of model parameters.

## 3.2 Expectation Maximization Algorithm

Varin *et al.* (2005) presented a pairwise EM algorithm for maximizing the likelihood. Based on this algorithm, in the E step, the value of conditional expectation is selected as follows

$$Q(\eta|\eta^{(m)}) = \sum_{(i,j)\in\chi} \int\int \log\{f(x_i, x_j, y_i, y_j; \eta)\} f(x_i, x_j|y_i, y_j; \eta^{(m)}) dx_i dx_j. \tag{3.3}$$

In the M step, the value $\eta^{(m+1)}$ is selected such that $Q(\eta^{(m+1)}|\eta^{(m)}) \geq Q(\eta^{(m)}|\eta^{(m)})$. If the conditional expectation cannot be expressed in a closed form, it can be approximated numerically. Varin *et al.* (2005) presented the Quadrature Pairwise EM (QPEM) algorithm

Approximate EM Algorithm:

- Step 1: Approximate E step: the conditional expectation value of (3.3) in the penalized EM algorithm is approximated with the value of $\hat{Q}(\eta; \eta^{(m)})$.
- Step 2: Generalized M step: the value of $\eta^{(m+1)}$ is chosen such that $\hat{Q}(\eta^{(m+1)}|\eta^{(m)}) \geq \hat{Q}(\eta^{(m)}|\eta^{(m)})$.
- Step 3: Reiterate Steps 1 to 3 of the algorithm until convergence.

for SGLM models and showed that its speed is more than the MCEMG algorithm. To solve the double integral of (3.3) the vector $(x_i, x_j)^T$ is transformed into the standardized components of $(\nu_i, \nu_j)$, where $\nu_i = \frac{x_i}{\sigma}$ and $\nu_j = \frac{x_j - \rho_{ij} x_i}{\sigma\sqrt{1-\rho_{ij}^2}}$ and $\rho_{ij} = \rho(s_i - s_j; \alpha)$. Now the approximate value of $Q(\eta;\eta^{(m)})$ is as follows

$$\hat{Q}(\eta;\eta^{(m)}) = \sum_{i,j=1}^{n}\sum_{k_1,k_2=1}^{k} \log f(x_i(h(k_1)), x_j(h(k_1),h(k_2)), y_i, y_j; \boldsymbol{\eta}) w_{ij}(k_1,k_2;\eta^{(m)}),$$

$$w_{ij}(k_1,k_2;\eta^{(m)}) = \frac{f(y_i|x_i(h(k_1);\eta^{(m)})f(y_j|x_j(h(k_1),h(k_2));\eta^{(m)})\ell(k_1)\ell(k_2)}{\sum_{k_1,k_2} f(y_i|x_i(h(k_1);\eta^{(m)})f(y_j|x_j(h(k_1),h(k_2));\eta^{(m)})\ell(k_1)\ell(k_2)}$$

Here, $h(k)$ are the nodes and $\ell(k)$ are the weights.

# 4 Simulation Study

In this study, composite likelihood, pairwise composite likelihood, and penalized composite likelihood functions are used, and the accuracy of each of these functions is checked through the MSE criterion. A neighbourhood with a radius of 4 is used for each observation. If we want to use all 48 neighbours for each point in the model, there are $48n = 10800$ pairs, which is far less than all possible ordered pairs, i.e., $n(n-1)/2 = 25200$. The QPEM algorithm with $M = 4 \times 4$ nodes of Gauss-Hermite quadrature was used to estimate the parameters. Now 15 pairs from radius four neighbours are randomly selected for each observation, which is shown in Figure 1. This would, in turn, reduce the number of observations to $15n = 3375$. The parameters were estimated using pairwise likelihood and weighted pairwise likelihood function, and the accuracy was compared through the MSE criterion. Results were obtained for 100 datasets. Different values of $d_s$ were inputted in the function (3.1) to obtain the optimal value of the weight function. Consider a matrix of Euclidean distance between spatial points and sort them from smallest to largest, then $d_s$'s are function of the quantiles of this values. For this research we consider $d_s = q(0.4), q(0.6), q(0.8), q(0.9)$. The results of the simulations are presented in Table 1.

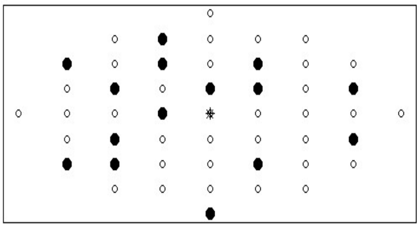


Figure 1: Sampling pairs within a neighborhood of radius 4. Here, × is the observation location and the filled circles are 15 neighbors sampled at random without replacement. The contributing pairs consist of × and each of the 15 sampled neighbors.

The effect of weighting the pairwise likelihood function on the accuracy of SGLM model was also examined, with Poisson response and logarithm link function. The data is generated from a $25 \times 25$ regular grid with nodes $\{(s_1,s_2) : s_1, s_2 = 0, 0.04, \ldots, 1\}$. To generate the spatial latent variables $X$, the normal distribution $N(0, \Sigma_\theta)$, the isotropic exponential covariance function $C(h) = \sigma^2 \exp(-3h/\phi), h > 0$, and the values $\sigma^2 = 1.5$, $\boldsymbol{\beta} = (\beta_0, \beta_1) = (1, 0.5)$ and $\phi = 6$ are considered. The explanatory variable in each position $s = (s_1, s_2)$ is considered as $Z_s = s_1$. The response variable, $Y_s$ is also generated by conditioning on spatial latent variables from distribution $Y_s \sim Poisson(n, \exp(\beta_0 + \beta_1 z_s + x_s))$ where $n = 25\times25$ is the number of samples. To avoid singularity, logarithm transformation was used to maximize the parameters of random effects. That is, the parameters were $\sigma^2$ and $\phi$ inputted in the model as $\log \sigma^2$ and $\log \phi$. The relation $\max_{\mathrm{i}}|\eta^{(m+1)} - \eta^{(m)}|/|\eta^{(m)}| < 0.0005$ was used as the criterion of convergence. To obtain the initial values, the regression parameters $\beta_0$ and $\beta_1$are estimated without considering the random variable and using a simple GLM model. Then, using the link function, the observed values are converted and the remaining values are estimated in the form of $\hat{r}(s_i) = g(y_i) - x_i\hat{\boldsymbol{\beta}}^{(0)}, i = 1, \ldots, n$. According to the results in Table 1, the weighted pairwise composite likelihood outperforms the pairwise composite likelihood function when $d_s$ is equal to $q(0.8)$ or $q(0.9)$ of the maximum distance between spatial points. Moreover, the outputted values for these two inputs were similar, implying that excluding the outlying pairs from the likelihood function would not result in substantial information loss and the model parameters can be estimated with acceptable accuracy even with fewer pairs.

Table 1: Estimation of SGLM model based on pairwise and weighted pairwise likelihood functions

| Likelihood | Weight | Parameter | Estimate | MSE | SE |
|---|---|---|---|---|---|
| Pairwise | 1 | $\beta_0$ | 0.842 | 0.1580 | 0.0377 |
| | | $\beta_1$ | 0.565 | 0.0934 | 0.0305 |
| | | $\sigma^2$ | 1.230 | 0.1137 | 0.0177 |
| | | $\phi$ | 5.640 | 3.4111 | 0.1826 |
| Penalized | $q(0.4)$ | $\beta_0$ | 1.552 | 0.8563 | 0.0374 |
| | | $\beta_1$ | 0.327 | 0.5946 | 0.0307 |
| | | $\sigma^2$ | 1.196 | 0.9116 | 0.0176 |
| | | $\phi$ | 6.105 | 14.054 | 0.3756 |
| | $q(0.6)$ | $\beta_0$ | 0.807 | 0.1702 | 0.0336 |
| | | $\beta_1$ | 0.474 | 0.1170 | 0.0312 |
| | | $\sigma^2$ | 1.470 | 0.1042 | 0.0187 |
| | | $\phi$ | 5.830 | 6.8114 | 0.3756 |
| | $q(0.8)$ | $\beta_0$ | 0.883 | 0.1052 | 0.0380 |
| | | $\beta_1$ | 0.469 | 0.1167 | 0.0300 |
| | | $\sigma^2$ | 1.554 | 0.1111 | 0.0184 |
| | | $\phi$ | 5.907 | 3.3143 | 0.3677 |
| | $q(0.9)$ | $\beta_0$ | 0.922 | 0.1120 | 0.0327 |
| | | $\beta_1$ | 0.519 | 0.0900 | 0.0343 |
| | | $\sigma^2$ | 1.511 | 0.0970 | 0.0166 |
| | | $\phi$ | 6.041 | 3.4819 | 0.1849 |

Two penalization functions, Lasso and $2\eta\eta^T$, were used for $\eta^T = (\sigma^2, \phi)^T$. QPEM algorithm estimates the parameters, maximizing $Q(\eta|\eta^m) - \lambda J(\eta)$ in the M step. Table 2 indicate that MSE and SE criteria in estimating the model parameters using the penalized pairwise likelihood function for both penalty functions are lower than those of the pairwise likelihood function. That is, the pairwise likelihood function can significantly increase parameter estimation accuracy. Also, the results indicate that the lasso penalty function outperforms that of the Green (1990).

Table 2: Estimation of SGLM models with pairwise and penalized pairwise likelihood functions

| Likelihood | Parameter | Estimate | MSE | SE |
|---|---|---|---|---|
| Pairwise | $\sigma^2$ | 1.483 | 0.114 | 0.0177 |
| | $\phi$ | 5.640 | 3.411 | 0.1826 |
| Penalized with Lasso | $\sigma^2$ | 1.506 | 0.071 | 0.0169 |
| | $\phi$ | 1.895 | 1.925 | 0.1504 |
| Penalized with Green | $\sigma^2$ | 1.489 | 0.093 | 0.0169 |
| | $\phi$ | 1.895 | 2.252 | 0.1504 |

# 5 Data Analysis

A real dataset was employed to check the performance of composite likelihood functions based on the QPEM algorithm. The dataset contains counts of Rhizoctonia root rot disease in barley collected at 100 sampling sites at Cunningham Farm in the northwestern United States. For each sampling site, 15 plants were pulled out from the ground for examination. A binomial SGLM model with a logit link function, a constant mean, $\beta_0$, and an exponential correlation function $C(h) = (1 - \tau^2)\sigma^2 \exp(-\frac{h}{\phi})$ was used for the spatial random effect. Furthermore, the penalized composite likelihood function with Lasso and Green penalty functions was implemented to estimate the SGLM model. Also, the method mentioned in Section 4 was used to obtain the initial values, the results of which are shown in Table 3.

Table 3: Estimation of SGLM models with pairwise and penalized pairwise likelihood functions

| | Method | | |
|---|---|---|---|
| Parameter | Lasso penalty | Green penalty | Weighted pairwise likelihood |
| $\beta_0$ | −1.69 | −1.75 | −1.73 |
| $\sigma^2$ | 0.15 | 0.09 | 0.18 |
| $\phi$ | 149.08 | 152.65 | 148.4 |
| $\tau^2$ | 0.58 | 0.61 | 0.46 |

# Discussion and Resuls

This research used pairwise likelihood, weighted pairwise likelihood, and penalized pairwise likelihood functions for spatial generalized linear mixed models. Furthermore, the

QPEM algorithm was used to maximize these functions. The simulation study showed that the weighted pairwise likelihood function and the penalized pairwise likelihood function outperformed the pairwise likelihood function in estimating the parameters of the SGLM model. The findings further established that the penalized pairwise likelihood function was more accurate than other functions in estimating the correlation parameters. Compared to the weighted pairwise likelihood function, the penalized pairwise likelihood function has higher accuracy in estimating $\sigma^2$ and $\phi$. Also, Lasso penalized function exhibited better results than the Green penalized function among penalized composite likelihood functions. The results were analyzed and compared for a real data set with binomial distribution. Overall, the results of the current study showed that the proposed functions outperformed other functions in estimating the parameters of the SGLM model.

# References

Azzalini, A., and Arellano-Valle, R. B. (2013), Maximum Penalized Likelihood Estimation for Skew-Normal and Skew-t Distributions, *Journal of Statistical Planning and Inference*, **143**, 419-433.

Bevilacqua, M., Mateu, J., Porcu, E., Zhang, H., and Zini, A. (2010), Weighted Composite Likelihood-Based Tests for Space-Time Separability of Covariance Functions, *Statistics and Computing*, **20**, 283-293.

Diggle, P., Tawn, J. A. and Moyeed, R. A. (1998), Model-Based Geostatistic, *Royal Statistical Society, Series C*, Applied Statistics, **47**, 299-350.

Green, P. J. (1990). On Use of the EM for Penalized Likelihood Estimation. *Journal of the Royal Statistical Society, Series B* (Methodological), 443-452.

Joe, H., and Lee, Y. (2009). On Weighting of Bivariate Margins in Pairwise Likelihood, *Journal of Multivariate Analysis*, **100**, 670-685.

McCulloch, C. (1997), Maximum Likelihood Algorithms for Generalized Linear Mixed Models, *Journal of the American Statistical Association*, **92**, 162-170.

McCullagh, C. E., and Nelder, J. A. (1989), *Generalized Linear Models*, Chapman and Hall. London.

Nelder, J. A., and Wedderburn, R. W. M. (1972), Generalized Linear Models, *Journal of the Royal Statistical Association, Series A*, **135**, 370-384.

Tibshirani, R. (1996), Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society, Series B*(Methodological), 267-288.

Varin, C., Host, G., and Skare, O. (2005), Pairwise Likelihood Inference in Spatial Generalized Linear Mixed Models, *Computational Statistics and Data Analysis*, **49**, 1173-1191.

Wedderburn, R. (1974), Quasi-Likelihood Functions, Generalized Linear Models and the Gaussnewton Method, *Biometrica*, **61**, 973-981.

دانشگاه بین المللی امام خمینی (ره) با همکاری
دانشگاه تربیت مدرس و انجمن آمار ایران برگزار می‌کند:

conf.ikiu.ac.ir/spatial5

پنجمین سمینار
آمار فضایی و کاربردهای آن
5th Seminar on Spatial Statistics and Its Applications
۳ و ۴ آبان‌ماه ۱۴۰۲
Oct. 2023 25-26

# Analysis of Survival Data with Spatial Survival Tree

Kiomars Motarjem[*]
Department of Statistics, Tarbiat Modares University, Tehran, Iran.

**Abstract:**

Spatial Survival Tree is a modeling approach used to analyze time-to-event data in the presence of spatial dependency and predictive covariates. This method is capable of dividing the data into subgroups, each associated with a relevant survival curve, and calculating the probability of survival for individuals in each group over time, taking into spatial-temporal survival correlations. Additionally, it utilizes criteria such as spatial location and event timing to further partition the data into smaller groups. The structure of the tree enables the identification of subgroups of individuals or spatial locations that possess unique survival characteristics while facilitating the selection of influential predictive variables on survival time. Simulation results conducted in this study demonstrate that the Spatial Survival Tree exhibits a higher efficacy in analyzing survival data with spatial structure, contributing significantly to improved accuracy and efficiency in the analysis of spatial survival data.

**Keywords:** Survival Data, Tree-based Algorithm, Spatial Survival Tree.
**Mathematics Subject Classification (2010):** 62M30, 62H30, 62N05.

## 1 Introduction

A spatial survival tree is a statistical method used to model the relationship between spatially dependent survival times and a set of independent variables on a spatial reference dataset. It utilizes a recursive binary partitioning algorithm to divide the study area into multiple regions or ”nodes” with distinct survival characteristics. Separate survival models are then fitted for each node. This method is particularly valuable when survival time is influenced by both spatial and non-spatial variables, and when the spatial structure of the data is important for analysis. It can identify spatially distinct regions with

[*] Speaker: k.motarjem@modares.ac.ir

different survival features and determine the variables that have the strongest association with survival time in each region (Breiman , 2001). The applications of spatial survival trees span various fields such as medicine, geology, and environmental science (De'ath and Fabricius , 2000). This method offers a novel approach to analyzing survival data, considering spatial dependency, and can aid in identifying regions with different survival characteristics. Notably, it can be applied to modeling survival in contagious diseases like COVID-19, which are influenced by spatial variables. By utilizing spatial survival trees, healthcare professionals can identify regions with different survival features within a specific geographical area and identify variables strongly associated with survival time in each region. This information can assist in selecting appropriate treatments for patients with similar survival characteristics. Additionally, spatial survival trees have applications in geology and environmental science, contributing to the identification of regions with similar characteristics and the identification of variables associated with survival time in each region.

## 2 Structure of Spatial Survival Tree

The structure of a spatial survival tree is similar to a decision tree, with the difference that spatial information is also taken into account in the analysis of survival data. This means that the root of the tree represents all individuals or spatial locations in the study area. Then, the tree is divided into smaller partitions recursively based on a splitting rule. This splitting rule usually includes one or more predictor variables, such as age, gender, and information related to spatial-temporal correlations in survival. The division process continues until a stopping condition is met, such as a minimum number of individuals in each sub-region or a minimum level of homogeneity within each sub-region. At this point, each sub-region is assigned a unique label or ”terminal node” corresponding to a specific combination of predictor variables and geographic location.

After constructing the tree, predictions can be made for any new individual or location in the tree and assigned to the appropriate terminal node. Ultimately, the probability of survival for that individual or location is estimated based on the survival function associated with the corresponding terminal node. Calculate the survival probability for each terminal node, typically depends on a specific type of survival analysis model, such as the Cox proportional hazards model or the accelerated failure time model. In a spatial survival tree, the calculation of survival probability for a new individual or location is based on the survival function associated with the terminal node it belongs to. To calculate this probability, the survival function relevant to the desired terminal node needs to be computed. The Cox model and accelerated failure time model are examined for two scenarios. (Bou-Hamad and Benoit , 2013). The calculation of survival probability for each terminal node usually relies on a specific type of survival analysis model, such as the Cox proportional hazards model or the accelerated failure time model. (Therneau and Grambsch , 2000). In a spatial survival tree, the calculation of survival probability for each new individual or location is based on the survival function associated with the corresponding terminal node. To calculate this probability, the survival function relevant to the desired terminal node needs to be computed. In the following, the Cox model and accelerated failure time model are examined for two scenarios.

1- The Cox proportional hazards model with spatial random effect defines the survival function associated with a terminal node as follows:

$$S(t|X,Z(s)) = S_0(t)^{exp(\beta_1 x_1+\beta_2 x_2+...+\beta_p x_p+Z(s))} \tag{2.1}$$

In this equation, $S(t|X,Z(s))$ represents the probability of survival of an individual at time $t$ and in position $s$. Also, $Z(.)$ represents a spatial random field. Given the values of the predictor variables $X$, $S_0(t)$ is the baseline survival function chosen by the user. $\beta_1, \beta_2, ..., \beta_p$ are the coefficients of the model, which are determined using estimation and fitting methods with training data (Motarjem et al. , 2020).

2- The accelerated failure time model with spatial random effect defines the survival function associated with a terminal node as follows:

$$S(t|X) = S_0(\frac{t}{exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + Z(s))}) \tag{2.2}$$

In this equation, $S(t|X,Z(s))$ and $S_0(t)$ represent the probability of survival of an individual at time $t$ in position $s$ given the values of the predictor variables $X$ and the baseline survival function, respectively. $\beta_1, \beta_2, ..., \beta_p$ are the coefficients of the model, which are determined using estimation and fitting methods with training data. By calculating the survival function associated with a terminal node, the probability of individual survival at time $t$ can be computed based on the predictor values. Additionally, $Z(.)$ represents a spatial random field. Generally, the probability of survival in a terminal node is calculated using the corresponding survival function of that node and the desired time.

To construct a spatial survival tree, the data containing predictor variables and survival time need to be collected. Then, based on this data, the spatial survival tree is built. The algorithm used for constructing the spatial survival tree is typically the ”Greedy” algorithm (Cormen et al. , 2009). In this algorithm, a base node is first selected for the tree, and then additional nodes are gradually added to the tree until a complete spatial survival tree is constructed. To select the next node, a stopping criterion should be defined. The stopping criterion serves as a condition under which the selection of the next node is halted (Hastie et al. , 2009). There are various criteria for this purpose, but two common criteria for stopping the spatial survival tree are:

- Partitioning Event: In this method, the tree is recursively partitioned to determine a stopping condition for adding new nodes. For example, the tree may be recursively divided into two sub-trees if the survival probability of an individual at different times varies significantly based on different predictor variables. In this case, the tree is recursively partitioned until the survival probability of an individual at each time is approximately the same.
- Number of Nodes: In this method, the number of nodes in the tree is used as a stopping criterion. For example, the tree may continue recursively as long as the number of nodes in the tree is less than a specified maximum value.

The mentioned criteria are used in constructing spatial survival trees. Below is a concise statement of the important partitioning theorem (Breiman et al. , 1984):

**Theorem 2.1.** *Partitioning Theorem: Let the data consist of a set of predictor vectors and survival time vectors represented as follows:*

$$D = (x_1, t_1), (x_2, t_2), ..., (x_n, t_n)$$

*Let $S$ be a subset of $D$ containing $m$ points. Also, let $S_L$ and $S_R$ be two other subsets of $S$ defined as follows:*

$$S_L = (x_i, t_i) \in S | x_i \leq x_L$$

$$S_R = (x_i, t_i) \in S | x_i > x_L$$

*Here, $x_L$ is a fixed value used for partitioning the data, so we can say that $S$ is divided into two subsets, $S_L$ and $S_R$, based on $x_L$. Now, if we define the sum of squared errors for each of these two subsets as follows:*

$$RSS_L = \sum_{(x_i,t_i) \in S_L} (t_i - \hat{t}_L)^2$$

$$RSS_R = \sum_{(x_i,t_i) \in S_R} (t_i - \hat{t}_R)^2$$

*where $\hat{t}_L$ and $\hat{t}_R$ are the predicted mean survival time for individuals in $S_L$ and $S_R$, respectively. Then, if $S$ is split into $S_L$ and $S_R$, the following quantity is minimized:*

$$RSS_{total} = RSS_L + RSS_R$$

*Therefore, the splitting stopping criterion is defined as:*

$$\Delta RSS = RSS_{parent} - RSS_{total}$$

*Here, $RSS_{parent}$ is the sum of squared errors for all data points in $S$. Based on this definition, the partitioning theorem can be stated as follows:*

*To achieve the best split on the data, we need to calculate $\Delta RSS$ for each possible partition and select the one with the largest $\Delta RSS$.*

After constructing a spatial survival tree, the partitioning and stopping criteria can be used for analyzing and predicting new data. Using the spatial survival tree, the probability of an individual's survival at different times can be calculated based on the values of predictor variables, and this information can be used for predicting other outcomes.

# 3 Spatial Survival Tree

To construct a spatial survival tree, we begin by partitioning the data based on the values of predictive variables and spatial coordinates. If D is the set of observations and S is a subset of D, we can express the partition as follows:

$$D = S_1 \cup S_2$$

where S1 and S2 are disjoint subsets of D. To determine the optimal partition between S1 and S2, we employ the logarithmic rank test statistic, which measures the difference in survival between the two subsets (Lehmann , 1959). The logarithmic rank test statistic is defined as follows:

$$LR(S_1, S_2) = \frac{(O_1 - E_1)^2}{V_1} + \frac{(O_2 - E_2)^2}{V_2}$$

In this equation, O1 and O2 represent the observed event counts in S1 and S2, while E1 and E2 represent the expected event counts in S1 and S2 under the assumption of no difference in survival, and V1 and V2 denote the variances of O1 and O2, respectively. After identifying the optimal partition, a proportional hazards model is fitted for each subset. This involves estimating the coefficients $\beta$ for each subset, which capture the impact of the predictive variables on the hazard function within that subset. The process of data partitioning based on predictive variables and spatial coordinates continues recursively and persists until a stopping criterion, such as a minimum number of observations in each subset, is reached. The resulting tree can be used to predict the survival time of new observations based on their predictive variable values and spatial coordinates.

# 4 Simulation

In this study, the objective is to compare the performance of a classical survival tree model with a spatial survival tree model. The classical survival tree considers only the predictive variables, while the spatial survival tree incorporates information related to spatial correlation alongside the predictive variables. By examining their performance based on common criteria, we can evaluate the advantages of considering spatial information in survival analysis. The steps involved in this simulation are as follows:

1. Variable Generation: We simulate two independent random variables, $X_1$ and $X_2$. Variable $X_1$ follows a normal distribution with a standard deviation of one and a mean of zero, while variable $X_2$ follows a uniform distribution between zero and one.

2. Spatial Correlation Structure: To introduce spatial correlation among the observations, we generate a spatial survival variable, $S$, using a spatial autocorrelation model. We consider a spatial exponential correlation model based on the distance between the observations. The spatial survival variable $S$ for observation $i$ is calculated as follows:

$$S_i = \sum_{j=1}^{n} \exp(-\alpha \cdot d_{ij}) \cdot X_{2j}$$

Here, $S_i$ represents the spatial survival variable for observation $i$, $d_{ij}$ denotes the distance between observations $i$ and $j$, $X_{2j}$ represents the value of variable $X_2$ for observation $j$, and $\alpha$ is a spatial decay parameter that controls the strength of spatial correlation. In this simulation, we set $\alpha = 0.5$.

3. Survival Time Generation: Survival times, $T$, are calculated using a Cox proportional hazards model that incorporates variables $X_1$, $X_2$, and $S$. We use a Cox proportional hazards model with relative risks:

$$h(t|X_1, X_2, S) = h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \beta_3 S)$$

Here, $h(t|X_1, X_2, S)$ represents the hazard function at time $t$ considering the values of variables $X_1$, $X_2$, and $S$, $h_0(t)$ is the baseline hazard function, and $\beta_1$, $\beta_2$, and $\beta_3$ are the coefficients associated with variables $X_1$, $X_2$, and $S$, respectively, all set to 1 for simplicity.

The table below compares the performance of the spatial survival tree and the classical survival tree:

Based on common criteria of accuracy, sensitivity, and precision, the spatial survival tree model outperforms the classical survival tree model. The spatial survival tree model

Table 1: Performance Comparison of Spatial Survival Tree and Classical Survival Tree

| Model | Accuracy | Sensitivity | Precision |
|---|---|---|---|
| Spatial Survival Tree | 0.85 | 0.80 | 0.90 |
| Classical Survival Tree | 0.70 | 0.65 | 0.75 |

achieves an accuracy of 0.85, correctly predicting survival outcomes in 85% of cases, and has a sensitivity of 0.80, accurately identifying the presence of survival in 80% of positive cases. Furthermore, the precision of the spatial survival tree model is 0.90, indicating that when it predicts a positive survival outcome, it is correct in 90% of cases. In contrast, the classical survival tree model exhibits poorer performance. It has an accuracy of 0.70, correctly predicting survival outcomes in 70% of cases, and a sensitivity of 0.65, accurately identifying the presence of survival in 65% of positive cases. Additionally, the precision of the classical survival tree model is 0.75, indicating that when it predicts a positive survival outcome, it is correct in 75% of cases.

## Conclusion

Generally, in this study, a spatial survival tree was compared to a classical survival tree in the context of survival data with spatial correlation. The results demonstrate that in the presence of spatial correlation, the spatial survival tree performs better. This finding was also confirmed in our sampling, where the spatial survival tree showed superior performance in identifying the desired spatial structure and its coherence with survival analysis. The ability of the spatial survival tree to consider spatial dependencies in survival data provides reliable insights and more accurate predictions, describing it as a promising tool for analyzing survival data with spatial correlation.

## Acknowledgement

We are grateful to the organizers and esteemed reviewers for their diligent evaluation and thorough review of the paper submitted for this fifth seminar on spatial statistics and its applications.

# References

Bou-Hamad, I., and Benoit, D. (2013). A review of tree-based methods for the analysis of survival data, *Statistical methods in medical research*, **22(4)**, 379-408.

Breiman, L. (2001). Random forests. *Machine learning*, **45**, 5-32.

Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J. (1984), *Classification and regression trees*, Wadsworth and Brooks, Monterey, CA.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2009). *Introduction to algoritheorems.* MIT press.

De'ath, G., and Fabricius, K. E. (2000). Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, **81(11)**, 3178-3192.

Hastie, T., Tibshirani, R. and Friedman, J. (2009), *The elements of statistical learning: data mining, inference, and prediction*, New York: springer.

Lehmann, E. L. (1959). Some principles of the theory of testing hypotheses. *The annals of mathematical statistics*, **30(2)**, 327-335.

Motarjem, K., Mohammadzadeh, M., and Abyar, A. (2020). Geostatistical survival model with Gaussian random effect. *Statistical Papers*, **61**, 85-107.

Therneau, T. M., and Grambsch, P. M. (2000). *Modeling survival data: extending the Cox model*, Springer, New York.

دانشگاه بین المللی امام خمینی (ره) با همکاری
دانشگاه تربیت مدرس و انجمن آمار ایران برگزار می‌کند:
conf.ikiu.ac.ir/spatial5
پنجمین سمینار
آمار فضایی و کاربردهای آن
5th Seminar on Spatial Statistics and Its Applications
۳ و ۴ آبان‌ماه ۱۴۰۲
Oct. 2023 25-26

# A Bayesian Semi-Parametric Spatial Count Model for Analysing Lung Cancer Mortality

Mahsa Nadifar[1*], Andritte Bekker[1], Mohammad Arashi[2]
[1]Department of Statistics, University of Pretoria, South Africa.
[2]Department of Statistics, Ferdowsi University of Mashhad, Iran.

**Abstract:**

The issue of treating unbalanced count data distributions in spatial count analysis prompts inquiries over the appropriateness of the Poisson model. Furthermore, more than traditional methods are required when straightforward parametric models do not capture the associations between variables because of the inclusion of covariates with unclear functional forms and complex or unspecified spatial patterns. To tackle these issues, we propose the implementation of an innovative Bayesian hierarchical modeling approach. This methodology combines non-parametric methods with a modified dispersed count model based on renewal theory, enabling us to effectively address challenges associated with count data exhibiting non-equivalent dispersion, nonlinear connections between variables, and intricate spatial patterns. In order to showcase the adaptability and efficacy of our proposed approach, we employ it to examine empirical lung cancer data obtained from Pennsylvania, United States.

**Keywords:** Bayesian spatial model, Count data, Semi-parametric, Over-dispersion, Under-dispersion, INLA.

**Mathematics Subject Classification (2010):** 62J12, 62F15, 62H11.

## 1 Introduction

Spatial count data play a pivotal role in diverse fields, encompassing disease mapping, environmental studies, ecology, sociology, crime analysis, and public health. While spatial generalized linear mixed models (SGLMMs) have become a popular choice for analyzing

[*] Speaker: u23027917@tuks.co.za

such data, certain scenarios necessitate greater flexibility due to the intricate relationships between variables and the presence of complex or unkown spatial patterns. In these instances, conventional parametric models often face difficulties in capturing nuanced characteristics of the data. To address these challenges, non-parametric methods offer a promising alternative. They enable us to adapt to unknown relationships, intricate dependency patterns, and circumvent the limitations of predefined functional forms (Illian et al. , 2013). By embracing non-parametric techniques, we empower the data itself to dictate the relationships, thus providing a more robust approach that minimizes misspecification bias. Additionally, when confronted with complex spatial patterns, non-parametric models frequently outperform their parametric counterparts. Consequently, adopting models with sufficient flexibility becomes imperative. Our utilization of a structured additive regression (STAR) (Kneib et al. , 2009) framework equips us with the capability to capture non-linear covariate relationships and various types of spatial dependencies.

Count data analysis often relies on the Poisson regression model, a component of generalized linear models. However, real-world count data often exhibit over- or under-dispersion, making the Poisson model unsuitable. Various extensions and alternatives have emerged to address these challenges, such as generalized linear mixed models and the negative binomial regression.Additionally, hurdle models, alternative distribution models (e.g., COM-Poisson, Poisson-Tweedie distribution) have been employed (Cameron , 2013). Winkelmann (1995) proposed an alternative approach to model non-equivalent-dispersed counts using renewal theory (Cox , 1962). This approach employs a less restrictive, non-exponential distribution with a non-constant hazard function for durations (waiting times) between events. Winkelmann linked models for counts and durations, thereby relaxing the equi-dispersion assumption at the expense of an additional parameter. Furthermore, he observed that a decreasing (increasing) hazard function results in negative (positive) duration dependence, which explains how negative duration dependence leads to over-dispersion and positive duration dependence leads to under-dispersion. Recent work has extended this approach to spatial modeling, as evidenced by Nadifar et al. (2023), who expanded the GC model for analyzing spatially correlated count data. Furthermore, Nadifar et al. (2021) introduced the GC STAR model, which offers enhanced applicability in spatial contexts.

In this paper, we address the challenge of modeling spatial count data with varying levels of dispersion by uniting renewal theory with popular semi-parametric approaches within a Bayesian framework. Our dual objectives are to model count responses with non-equivalent dispersion, particularly when covariate-response relationships are uncertain and spatial patterns exhibit complexity or uncertainty, and to explore the practical implications of this approach.

The remainder of this paper is organized as follows: Section 2 provides a brief overview of the spatial GC regression model with semi-parametric approaches, while Section 3 elaborates on the methodology for the Bayesian semi-parametric spatial GC model. In Section 4, we apply the methodology to analyze clinical data related to lung cancer mortality in Pennsylvania, USA. The paper concludes with a concise discussion in Section 5.

# 2 GC Structured Additive Regression Model

## 2.1 Gamma-Count Distribution

We provide a concise overview of the fundamental characteristics of the GC distribution. From a statistical standpoint, there exists a unique relationship between the distribution of cumulative waiting times and count distributions, which can be leveraged to develop novel count data distributions. The inception of the GC distribution traces back to Winkelmann (1995), and it is grounded in the concept of waiting times being gamma-distributed.

Consider a sequence of waiting times $\{u_k, k \geq 1\}$ representing the intervals between the $(k-1)$th and $k$th events. Consequently, the arrival time of the $n$th event can be expressed as $\vartheta_n = \sum_{k=1}^{n} u_k$, for $n = 1, 2, \ldots$. Let $Y_t$ denote the total count of events occurring between time 0 and $t$. Thus, $\{Y_t,\ t > 0\}$ forms a counting process, and for any fixed $t$, $Y_t$ represents a count variable. The statistical properties of this counting process (and subsequently of the count variable) are completely defined once we have access to the joint distribution function of the waiting times, $\{u_k,\ k \geq 1\}$. In essence, $Y_t < n$ if and only if $\vartheta_n > t$. As a result, the probability mass function of $Y_t$ can be expressed as $f_{Y_t}(n) = F_n(t) - F_{n+1}(t)$, where $F_n(T)$ denotes the distribution function of $\vartheta_n$. In general, $F_n(t)$ entails a complex convolution of the underlying densities of $u_k$, rendering it analytically intricate. Nevertheless, a notable simplification emerges when we assume that $\{u_k,\ k \geq 1\}$ comprises independently and identically gamma-distributed random variables with parameters $Gamma(\alpha, \gamma)$. In this scenario, where the mean is $\mathrm{E}(u_k) = \alpha/\gamma$ and the variance is $\mathrm{Var}(u_k) = \alpha/\gamma^2$, it can be demonstrated that if $Y_t$ signifies the count of events within the interval $(0, t)$, it follows a GC distribution with parameters $\alpha$ and $\gamma$, denoted as $Y_t \sim \mathrm{GC}(\alpha, \gamma)$. The probability mass function of $Y_t$ takes the form:

$$f_{Y_t}(y) = G(y\alpha, \gamma t) - G((y+1)\alpha, \gamma t), \qquad y = 0, 1, 2, \ldots, \tag{2.1}$$

where $G(n\alpha, \gamma t) = \frac{1}{\Gamma(n\alpha)} \int_0^{\gamma t} v^{n\alpha - 1} e^{-v} dv$, and $G(0, \gamma t) = 1$. Furthermore, the mean of the GC distribution can be computed as $\mathrm{E}(Y_t) = \sum_{k=1}^{\infty} G(k\alpha, \gamma t)$.

It's important to note that for non-integer values of $\alpha$, closed-form expressions are unavailable for $G(y\alpha, \gamma t)$, and consequently for $f_{Y_t}(y)$ and $\mathrm{E}(Y_t)$. In the case of $\alpha = 1$, the distribution of $u_k$ reduces to the exponential distribution, leading to a simplification of (2.1) to the Poisson distribution with the parameter $\gamma t$. Significantly, when $\alpha > 1$, indicating positive duration dependence, the GC distribution exhibits under-dispersion. Conversely, for $0 < \alpha < 1$, representing negative duration dependence, the GC distribution tends towards over-dispersion.

## 2.2 Semi-Parametric Spatial GC Regression Model

To construct the semi-parametric spatial GC regression model, we delve into the realm of hierarchical spatial modeling for count data aggregated across spatially indexed units like regions, districts, or countries. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ denote the response vector, with $Y_i$ representing the response in the $i$th area, where $i = 1, \ldots, n$, sampled from the GC distribution. Estimating $\mathrm{E}(Y_i)$ directly is challenging due to its complexity, making it cumbersome to extend a regression model based on the mean. Thus, our semi-parametric regression model operates on the waiting times $u_{k_i}$ rather than $Y_i$, where $u_{k_i}$ is the generic representation of waiting times for the $i$th observation. Consequently, we express:

$$\begin{aligned} \mathrm{E}(u_{k_i}|\widetilde{\boldsymbol{x}}_i) &= \frac{\alpha}{\gamma_i} = \exp(-\eta_i) \\ \eta_i &= \beta_0 + \sum_{j=1}^{J} f_j(\widetilde{\boldsymbol{x}}_i) + f_s(\sigma_i), \qquad i = 1, \ldots, n \end{aligned} \tag{2.2}$$

where $\beta_0$ acts as an intercept term representing the overall predictor level. The functions $\{f_j\}$ model non-linear fixed effects with first- or second- order random walk (RW1 or RW2), $f(\cdot)$ encompasses a two-dimensional random walk (RW2D) term for spatial effects, and $\{\sigma_i = (s_{i1}, s_{i2}),\ i = 1, \ldots, n\}$ denotes the set of geographical centroids. RW2D models share similarities with thin-plate splines (TPS) models (Dupont et al., 2020) and can be considered as non-parametric counterpart of intrinsic Gaussian Markov random field (IGMRF) (Rue and Held, 2005). The joint distribution for $\mathbf{f}_s = (f_s(\sigma_1), \ldots, f_s(\sigma_n))'$ is defined as:

$$\mathbf{f}_s|\tau_s \sim \mathrm{N}\left(\mathbf{0}, (\tau_s \mathbf{Q}_s)^{-1}\right),$$

where $\tau_s$ and $\mathbf{Q}_s$ denote the precision parameter and precision matrix of a two-dimensional second-order polynomial IGMRF, respectively. Please refer to Nadifar et al. (2022) for a comprehensive understanding of constructing an RW2D model.

Taking into account the inverse relationship between gaps and the number of occurrences, the negative sign preceding $\eta$ in (2.2) signifies the reverse influence of fixed and random effects on waiting times, contrasting with their impact on counts. In simpler terms, a longer expected time interval leads to a lower number of occurrences. Consequently, the GC regression model emerges from inherent parametric assumptions that encompass the Poisson regression model through a singular parametric constraint. From (2.2), we can express $\gamma_i = \alpha \exp(\eta_i)$. Hence, the semi-parametric spatial GC regression model can be formulated as follows:

$$Y_i|\alpha, \boldsymbol{\beta}, \phi_i \sim \mathrm{G}\left(y_i\alpha, \alpha \exp\left(\eta_i\right)\right) - \mathrm{G}\left(\left(y_i+1\right)\alpha, \alpha \exp\left(\eta_i\right)\right), \quad i = 1, \ldots, n, \tag{2.3}$$

where $\alpha$ denotes the dispersion parameter. Assessing the marginal likelihood corresponding (2.3) often necessitates dealing with intractable integrals, posing a significant challenge for implementing likelihood-based inferences. To circumvent these challenges, we develop the proposed models within a Bayesian framework and leverage the INLA methodology.

# 3 Semi-parametric Bayesian Learning

To conduct Bayesian analysis for our proposed model, we need to specify appropriate prior distributions for the model parameters, namely $\alpha$, $\boldsymbol{\beta}$, $\mathbf{f}_x$, $\mathbf{f}_s$, $\tau_x$, and $\tau_f$. These priors should reflect our prior beliefs about these parameters. We choose penalized complexity (PC) prior for $\alpha$ (Nadifar et al., 2021),with its hyper-parameter set to 3, an option available in the INLA. Additionally, we employ PC priors for the precision parameters governing the spatial effect and $\widetilde{\boldsymbol{x}}$'s non-linear effect (Simpson et al., 2017). Moreover, we consider RW2d for spatial effect and RW1 or RW2 for fixed effects as described in Section 2. These prior distributions offer the flexibility to capture prior beliefs through the appropriate choice of hyper-parameters. We assume that these parameters are a priori independent, and by accepting these priors, the joint posterior density for the proposed model can be expressed as follows:

$$\pi(\alpha, \mathbf{f}_x, \mathbf{f}_s, \tau_s, \tau_x|\widetilde{\boldsymbol{y}}) \propto \prod_{i=1}^{n} \left\{\mathrm{GC}\left(\alpha, \alpha \exp\left(\eta_i\right)\right)\right\} \pi(\alpha)\mathrm{RW2}(\mathbf{f}_x)\pi(\tau_x)\mathrm{RW2D}(\mathbf{f}_s)\pi(\tau_s). \tag{3.1}$$

Traditionally, inference for models (3.1) is accomplished using Markov Chain Monte Carlo (MCMC) sampling. However, it is well-known that MCMC methods encounter issues related to both convergence and computational time when applied to such complex models (Rue et al., 2009). In particular, implementing Bayesian inference through MCMC algorithms for large spatial data could take several hours or even days to complete. To address this challenge, Rue et al. (2009) introduced the Integrated Nested Laplace Approximation (INLA) method, a deterministic algorithm capable of providing accurate results in seconds or minutes. INLA combines Laplace approximations and numerical integration efficiently to approximate posterior marginal distributions. Let $\theta$ represent the vector of hyper-parameters, which is $(\alpha, \tau_s, \tau_x)'$ for model (3.1). Also, let $\psi$ denote the $\ell \times 1$

vector of latent variables, which is $(\boldsymbol{\beta}, \mathbf{f}_x, \mathbf{f}_s)'$, with $\ell$ determined by the specific model. In practice, our primary interest lies in the marginal posterior distributions for the elements of the latent variables and hyper-parameters, given by:

$$\pi(\psi_j|\widetilde{\boldsymbol{y}}) = \int \pi(\psi_j, \theta|\widetilde{\boldsymbol{y}})d\theta = \int \pi(\psi_j|\theta, \widetilde{\boldsymbol{y}})\pi(\theta|\widetilde{\boldsymbol{y}})d\theta, \quad j = 1, \ldots, \ell,$$
$$\pi(\theta_k|\widetilde{\boldsymbol{y}}) = \int \pi(\theta|\widetilde{\boldsymbol{y}})d\theta_{-k}, \quad k = 1, 2, 3,$$

where $\theta_{-k}$ represents $\theta$ with the $k$th element removed. INLA's crucial success lies in its ability to compute model comparison criteria, such as the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002). Our proposed GC model is already implemented in the `R-INLA` package as a `family` argument with the name "`gammacount`."

# 4 Lung Cancer Mortality Modeling

Here, we reanalyze the lung cancer mortality data in Pennsylvania, USA, comprising 67 districts for the year 2002. We introduce the ecological covariate $\mathtt{x}$ to account for smoking consumption. This dataset is available in the R package `SpatialEpi`. We assume that the observed death counts $y_i$ in district $i = 1, \ldots, 67$ are conditionally independent, following a semi-parametric spatial GC regression model, where $\eta_i$ is defined as:

$$\eta_i = \log(\mathtt{E}_i) + \beta_0 + f(\mathtt{x}_i) + g(s_i), \quad i = 1, \ldots, 67, \tag{4.1}$$

where, $\mathtt{E}_i$, $f(\cdot)$, and $g(\cdot)$ represent the offset, a smoothing functional effect of smoking consumption, and spatial dependency, respectively, as described in Section 2. To investigate more closely, we also considered the parametric model corresponding to (4.1), as well as the Poisson model, as competing count models. Table 1 presents the Deviance Information Criterion (DIC) values for the proposed model and its alternatives. Table 1 shows that the semi-parametric spatial GC regression model outperforms other models. Consequently, we present posterior inference for the superior model in Table 2 and Figure 1 for parameters and smoothing effects of smoking and spatial dependency, respectively. Table 2 displays the estimated parameters with their 95% credible intervals. The estimates of $\alpha$ and its 95% credible intervals confirm that Pennsylvania data exhibit over-dispersion,

Table 1: Computed DIC values for the semi-parametric spatial model and its corresponding parametric models for GC and Poisson.

| | Parametric | | Semi-parametric | |
|---|---|---|---|---|
| | Poisson | GC | Poisson | GC |
| DIC | 104.04 | 101.42 | 109.12 | **98.51** |

suggesting that the Poisson model is inappropriate, as indicated in Table 1. The posterior results for the precision parameter of the spatial effect show that the spatial dependency is significant. The estimated smoothing precision of smoking consumption effect represents less uncertainty. Figure 1 presents the posterior inferences for the effect of smoking consumption, revealing an increasing non-linear effect of smoking consumption, which closely follows a linear behavior. There is a positive relationship between smoking consumption and lung cancer deaths, indicating that higher smoking consumption increases the risk of mortality. Furthermore, Figure 1 illustrates the estimated posterior mean of the spatial effect, showing that people in downtown areas have a lower risk of mortality.

Table 2: Posterior mean estimates and 95% credible intervals of parameters for the semi-parametric spatial GC regression model.

| $\beta_0$ | $\alpha$ | $\tau_x$ | $\tau_s$ |
|---|---|---|---|
| -0.035 (-0.064,-0.005) | 0.82 (0.52,1.18) | 21010.5 (1403.92,77243.34) | 165.65 (56.44,392.5) |

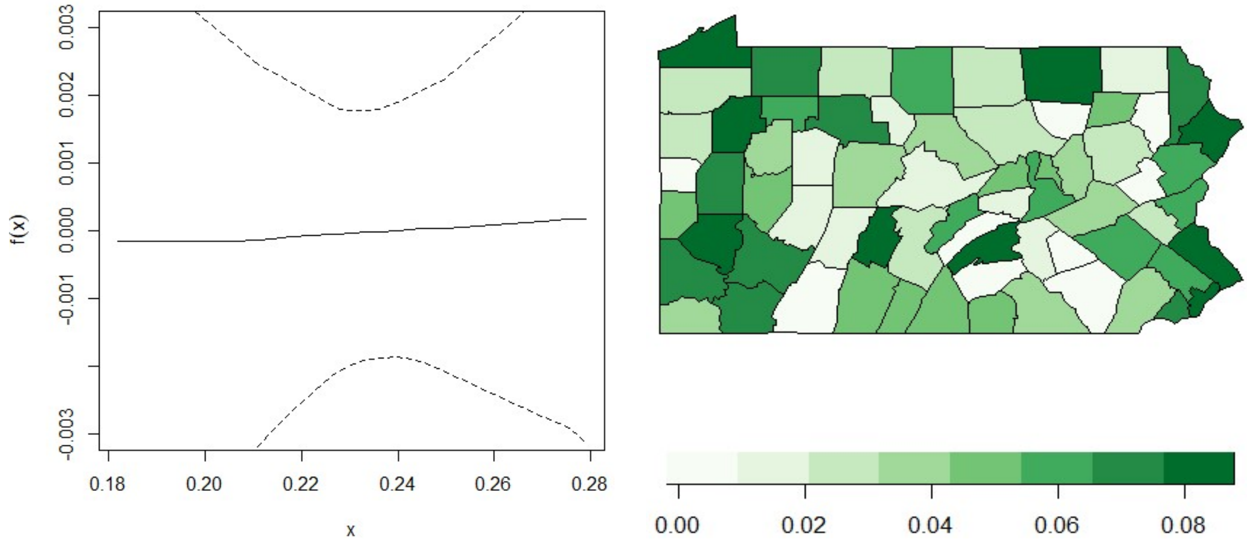


Figure 1: Posterior mean estimates of smoking consumption effect (line) and 95% credible bounds (dashed) for the semi-parametric spatial GC regression model.

## Conclusion

In this paper, we introduced a semi-parametric spatial Gamma-Count (GC) regression model, which we applied to analyze lung cancer mortality data in Pennsylvania, USA. Our model outperformed traditional Poisson and parametric GC models, confirming the data's over-dispersion and demonstrating the unsuitability of the Poisson model. We found a positive link between smoking consumption and lung cancer mortality, revealing a higher risk with increased smoking. Additionally, spatial analysis showed lower mortality risks in downtown areas, indicating geographical variations. This semi-parametric model offers a robust framework for analyzing count data with varied dispersion and complex spatial patterns, particularly relevant for public health research.

# References

Cameron, A. and Trivedi, P. (2013). *Regression Analysis of Count Data*. Cambridge University Press.

Cox, D. R. (1962). *Renewal Theory*. London: Methuen.

Illian, J. B., Martino, S., S∅rbye, S. H., GallegoFernández, J. B., Zunzunegui, M., Esquivias, M. P., and Travis, J. M. (2013). Fitting complex ecological point process models with integrated nested Laplace approximation. Methods in Ecology and Evolution, 4(4), 305-315.

Kneib, T., Hothorn, T., and Tutz, G. (2009). Variable selection and model choice in geoadditive regression models. *Biometrics*, **65(2)**:626-634.

Nadifar, M., Baghishani, H., Kneib, T. (2021). Flexible Bayesian Modeling of Counts: Constructing Penalized Complexity Priors, *Available from: arXiv: 2105.08686.*

Nadifar, M., Baghishani, H. and Fallah, A. (2023). A Flexible Generalized Poisson Likelihood for Spatial Counts Constructed by Renewal Theory, Motivated by Groundwater Quality Assessment. *JABES*. https://doi.org/10.1007/s13253-023-00550-5.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall/CRC Press, London.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society Series B*, **71**:319-392.

Spiegelhalter, D. J., Best, N., Carlin, B. P., and Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, **64**:1-34.

Winkelmann, R. (1995). Duration dependence and dispersion in count-data models. *Journal of Business & Economic Statistics*, **13(4)**:467-474.

دانشگاه بین المللی امام خمینی (ره) با همکاری
دانشگاه تربیت مدرس و انجمن آمار ایران برگزار می‌کند:
conf.ikiu.ac.ir/spatial5
پنجمین سمینار
آمار فضایی و کاربردهای آن
5th Seminar on Spatial Statistics and Its Applications
۳ و ۴ آبان‌ماه ۱۴۰۲
Oct. 2023 25-26

# Introducing a New Method in Determining the Fault Plane Using the Spatial Position of Aftershocks

Shahrokh Pourbeyranvand*
International Institute of Earthquake Engineering and Seismology, Tehran, Iran.

**Abstract:**

In this study, a new method for determining the fault plane using the spatial position of the hypocenter of aftershocks has been introduced. A conventional method that is usually used in seismological studies of aftershock sequences is to draw earthquakes on a map and design several cross-sections in such a way that the spatial distribution of events can be inferred. In this way, the fault plane can be visualized. In the proposed method however, the role of human observation and inference is almost eliminated. By using automatic fitting methods in the 3D environment, the spatial position of seismic events is used directly to create a surface and then a planar surface which is recognized and fault plane is fit to it. The geometrical characterization of the fault plane can be achieved by computer modeling hence.

**Keywords:** hypocenter, earthquake, focal mechanism, fault, seismicity, location.

## 1 Introduction

Spatial statistics play an important role in identifying fault planes in seismology. It helps to analyze the spatial distribution of seismic events and to understand the behavior of active fault systems https://earthquake.usgs.gov/data/finitefault/. In seismology, a fault plane is a surface that shows the location and direction of a fault, which is a break in the Earth's crust where rocks on either side have moved relative to each other. Identifying the fault plane is necessary to study the fault process and determine the geometry and movement of the fault https://www.sciencebase.gov/catalog/item/58da9d37e4b0543bf7fdaab3. Spatial statistical techniques are used to analyze seismic

*Speaker: beyranvand@iiees.ac.ir

catalogs and identify significant seismic aftershocks, independent background events, and clustersź. These techniques help to estimate the background seismicity rate, which is a critical parameter in seismic hazard analysis. One approach to identifying fault planes involves using a two-step clustering approach. This approach combines self-organizing map (SOM) and density-based temporal clustering methods for catalog analysis of earthquakes. The SOM phase identifies the main hotspots (SOM prototypes) in the region based on the event location and depth information. A density-based temporal clustering step then identifies events in the neighborhood of each SOM prototype, enabling efficient spatiotemporal analysis and identification of aftershock clusters and background events (Laesanpura et al. , 2019). The accuracy of fault plane identification can be verified using statistical parameters such as coefficient of variation (time domain) and m-Morisita index (spatial domain) (Sainoki et al. , 2023). These parameters help justify and validate the accuracy of the clustering approach (Sharma et al. , 2022). Using spatial statistical techniques, seismologists can gain insight into earthquake clusters, fault plane behavior, and seismic hazard analysis. This knowledge helps to better understand the complexity, occurrence patterns and behavior of earthquakes for successful risk reduction https://earthquake.usgs.gov/.

## 2 Earthquake in Damavand with a Magnitude of 5.1

A fairly strong earthquake shook the city of Damavand and its surrounding villages in Tehran province on the morning of Friday, May 19, 2019. This event was well felt in the city of Tehran and caused people to panic and leave the capital. The range of feelings of this earthquake reached the provinces of Mazandaran, Semnan, Alborz and Qom. The proximity of the epicenter of this earthquake to Tehran caused the attention of public opinion and media to this earthquake and the fault that caused it. According to the institute's report, the magnitude of this earthquake was at a depth of 7 kilometers and its epicenter was located near Masha village in the north of Damavand city. The National Earthquake Accelerometer Network recorded this earthquake in more than 40 accelerometer stations in Tehran, Semnan, Mazandaran and Qom provinces. The maximum acceleration due to the event of this earthquake at the station of Rudhen was around 120 cm/s2 (Figure 1) https://earthquake.usgs.gov.

## 3 Method

In the new approach presented here, an automatic method has been used instead of human visual judgment. First, the events are entered into a 3D modeling computer software

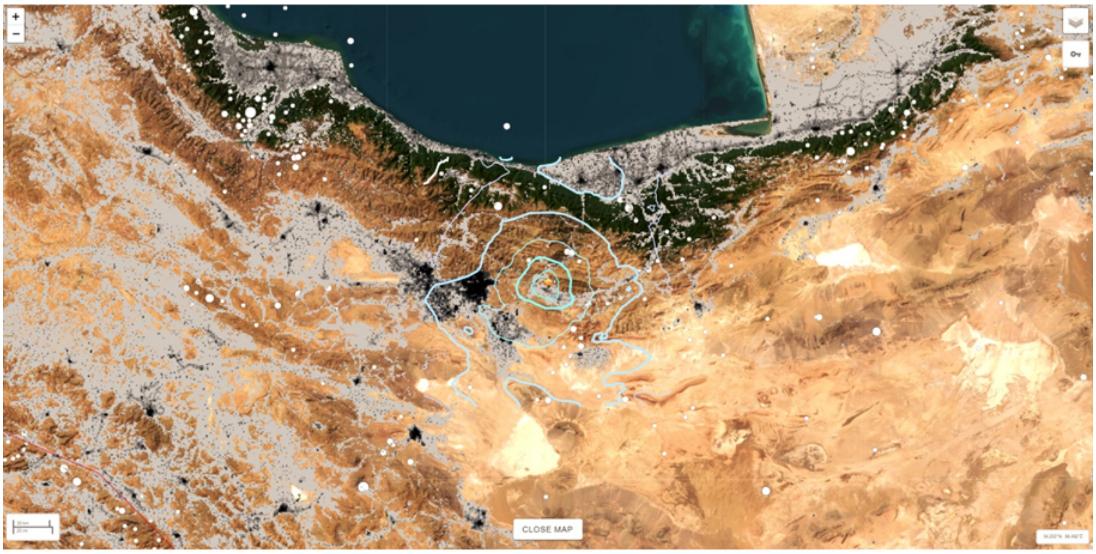


Figure 1: Earthquake shake map.

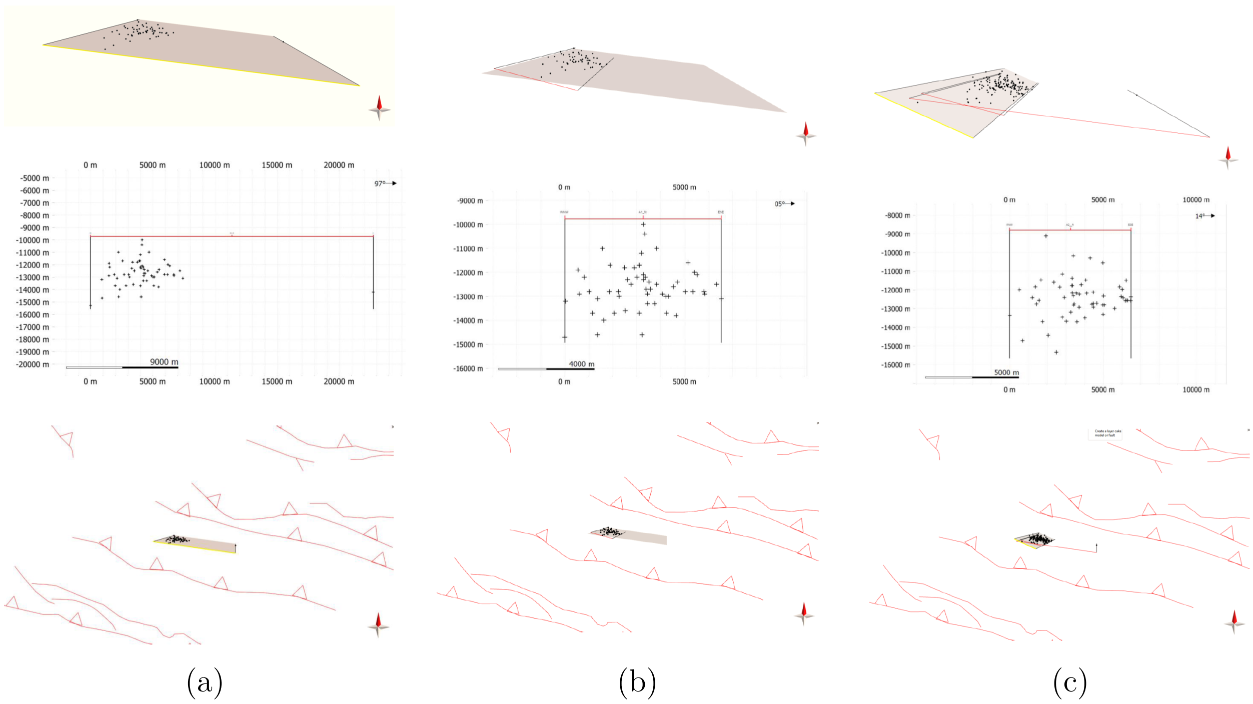
environment. A direct plane is then fitted to the data points in 3D space. In fact, a surface is created using the spatial position of the hypocenter of the earthquake, and then a planar surface is installed obliquely on the selected surface. The result of this quantitative approach is expected to be closer to reality compared to the previous qualitative method that used human visual perception instead of a mathematical description of the geometric features of the desired page. The hypocentral locations of the earthquakes can be seen in Figure 2a. A single seismic event is isolated from other events. The plane corresponding to the data after removing this outlier event is shown in Figure 2b. According to the higher accuracy of the hypocentral location of the events after the re-location of the earthquakes (Figure 2c.), the resulting plane is expected to be closer to reality.

The characteristics of the extracted fault planes in each of the three modes above are given in Table 1 and displayed in Figure 3.

Table 1: Geometrical description of the fault plane.

| No. | strike | dip |
|---|---|---|
| 1 | 277.42 | 88.6 1 |
| 2 | 278.48 | 81.61 |
| 3 | 279.21 | 89.81 |

In Figure 4, surfaces are created using spatial position plates of seismic events for the three discussed modes, and then a flat plate is fitted to these surfaces. This plane shows the same geometrical character as the actual fault page. The difference between the azimuth or alignment of the fault planes calculated by different research centers (Table 2) and this study is less than 15 degrees, which can be considered within the range of



(a) (b) (c)

Figure 2: Fitting a plane to the spatial data of aftershock events (a), Fitting a plane to the same data after removing outliers (b) and Fitting a plane to relocated events (c). top: the spatial position of the data and fitting a plane to them, middle: the cross-section of the events, bottom: the position of the created plane according to the faults in the area.

.

uncertainty of the focal mechanism data and is therefore acceptable. But the difference between the slopes of the fault planes among different studies reaches more than 37 degrees and seems quite variable. The observed differences are shown in Figure 5. The slope of the fault plane determined in this study is close to the result obtained by GFZ.

Table 2: Characteristics of the fault plane calculated for the earthquake in this study, Iranian Seismological Research Center, German Center for Geosciences and GCMT.

| No. | Date | Time | Lat. | Lon. | Depth | Mag. | S | D |
|---|---|---|---|---|---|---|---|---|
| New Method | 20200507 | 201821 | 35.777 | 52.041 | 12.1 | 5.1 (Ml) | 86.67 | 278.37 |
| IRSC | 20200507 | 201821 | 35.776 | 52.046 | 11.16 (centroid) | 4.9 (Mn) | 52 | 284 |
| GFZ | 20200507 | 201821 | 35.700 | 52.040 | 16 (centroid) | 4.9 Mw | 86 | 96 |
| GCMT | 20200507 | 201821 | 35.750 | 52.100 | 27 (centroid) | 5 Mw | 68 | 292 |

# 4 Calculation of earthquake focal mechanism

The focal mechanism of the above earthquake has been calculated in various international authorities. For example, the GCMT report about this earthquake in the study area is

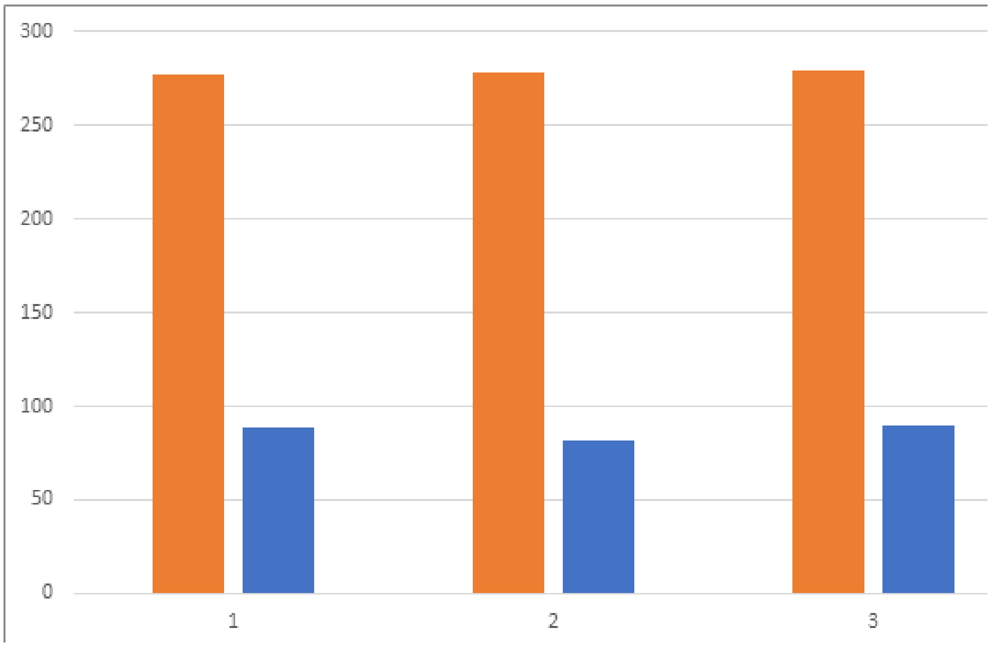


Figure 3: Fitting a plane to the spatial data of aftershock events, top: the spatial position of the data and fitting a plane to them, middle: the cross-section of the events, bottom: the position of the planes created according to the faults in the area.

presented in Figure 6. As can be seen in this table, the focal mechanism calculated for this earthquake has a dominant strike-slip movement with a dip-slip component. Therefore, the proposed fault plane shows a good agreement with the trend of the faults in the region. Also, the focal mechanism calculated for this earthquake in this study is in accordance with the result announced by GFZ.

## 5 Conclusion

Using the proposed method, the geometrical characteristics of the fault plane were calculated with acceptable accuracy. This fault plane is in accordance with the result announced by the GFZ and is consistent with the trend of the faults in the region. Considering the novelty of the proposed method, it is necessary to study more datasets with higher accuracy to implement the method and validate the results.

## 6 Acknowledgment

Dr. Farzam Yaminifard is appreciated for providing the seismicity of the studied area based on the local network data.

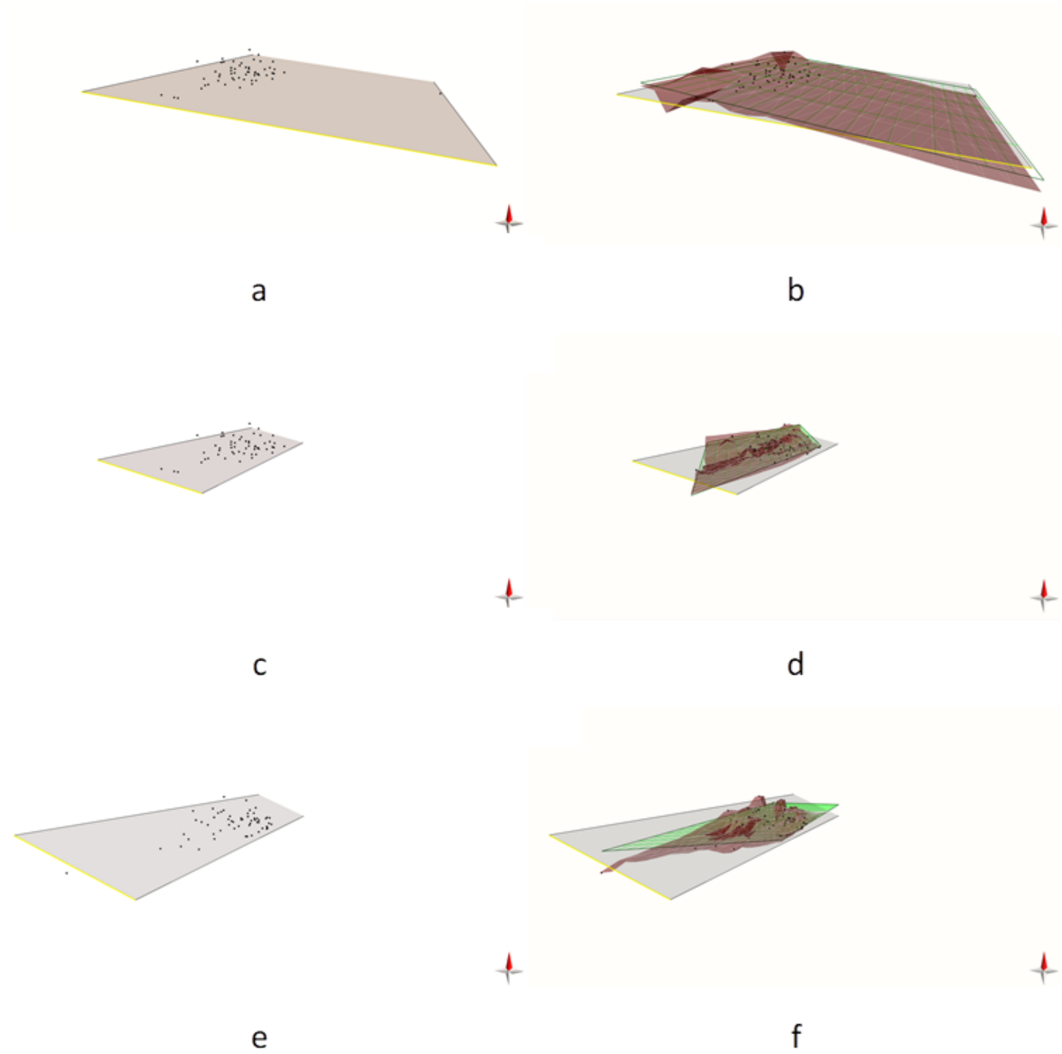


Figure 4: Fitting a plane to the spatial data of aftershock events, top: the spatial position of the data and fitting a plane to them, middle: the cross-section of the events, bottom: the position of the planes created according to the faults in the area.

# References

Laesanpura, A., Afnimar Dahrin, D. and Abdurachman, D. (2019), The fault identification by gravity and seismology in west Lembang segment, *IOP Conference Series: Earth and Environmental Science*, **311**(1), 012057.

Sainoki, A., Schwartzkopff, A. K., Jiang, L., and Mitri, H. (2023), Numerical modeling of spatially and temporally distributed on-fault induced seismicity: Implication for seismic hazards, *International Journal of Coal Science & Technology*, **10**(1), https://doi.org/10.1007/s40789-022-00560-7.

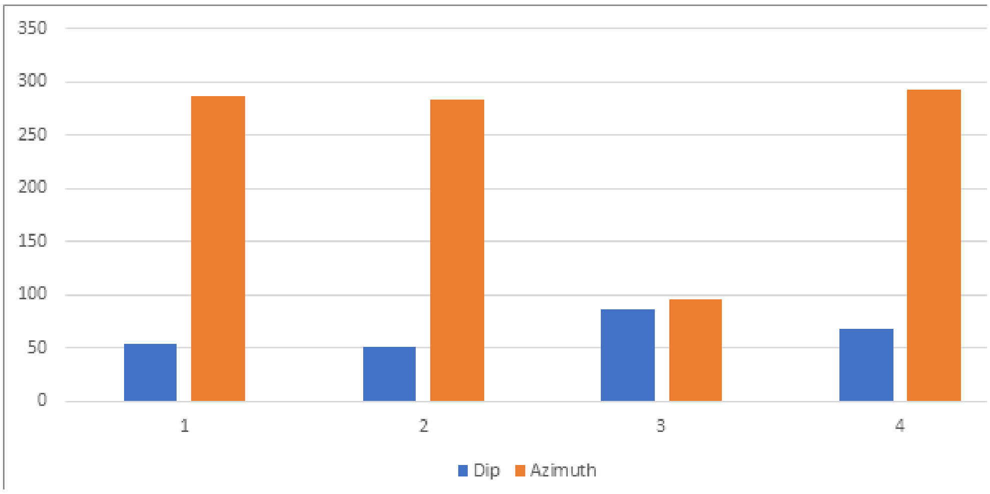
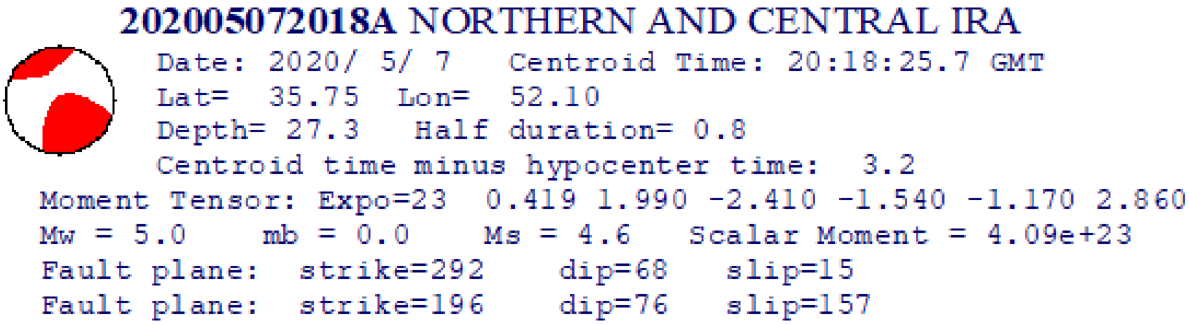


Figure 5: Fitting a plane to the spatial data of aftershock events, top: the spatial position of the data and fitting a plane to them, middle: the cross-section of the events, bottom: the position of the planes created according to the faults in the area.



```
            202005072018A NORTHERN AND CENTRAL IRA
             Date: 2020/ 5/ 7   Centroid Time: 20:18:25.7 GMT
             Lat=  35.75  Lon=  52.10
             Depth= 27.3   Half duration= 0.8
             Centroid time minus hypocenter time:  3.2
  Moment Tensor: Expo=23  0.419 1.990 -2.410 -1.540 -1.170 2.860
  Mw = 5.0    mb = 0.0    Ms = 4.6   Scalar Moment = 4.09e+23
  Fault plane:  strike=292    dip=68    slip=15
  Fault plane:  strike=196    dip=76    slip=157
```

Figure 6: Fitting a plane to the spatial data of aftershock events, top: the spatial position of the data and fitting a plane to them, middle: the cross-section of the events, bottom: the position of the planes created according to the faults in the area.

Sharma, A., Vijay, R. K., and Nanda, S. J. (2022), Identification and spatio-temporal analysis of earthquake clusters using SOMDBSCAN model, *Neural Computing and Applications*, , **35**(11), 8081-8108.

دانشگاه بین المللی امام خمینی (ره) با همکاری
دانشگاه تربیت مدرس و انجمن آمار ایران برگزار می‌کند:
conf.ikiu.ac.ir/spatial5
پنجمین سمینار
آمار فضایی و کاربردهای آن
5th Seminar on Spatial Statistics and Its Applications
۳ و ۴ آبان‌ماه ۱۴۰۲
Oct. 2023 25-26

# Bayesian Kriging Regression for Post-Earthquake Damage Prediction

Mahdi Rahmani-Qeranqayeh[1*], Morteza Bastami[1], Afshin Fallah[2]
[1]International Institute of Earthquake Engineering and Seismology, Tehran, Iran.
[2]Department of Statistics, Imam Khomeini International University, Qazvin, Iran.

**Abstract:**

This study utilizes Bayesian kriging regression to predict damage in earthquake-affected areas. This approach accounts for the spatial correlation between building damage and ensures that the predicted values remain within an acceptable range. It is also well-suited for handling damage data with measurement errors. These are aspects that are often overlooked in earthquake studies. The performance of the Bayesian kriging regression was compared with that of regression kriging and probit regression using both simulated and actual datasets from the Sarpol-e Zahab earthquake. The results showed that the Bayesian kriging regression model provided superior predictions of the damage ratio compared to the other models, exhibiting lower bias.

**Keywords:** Earthquake damage, Bayesian approach, Spatial correlation, Probit model.
**Mathematics Subject Classification (2010):** 62M30, 62H30, 62N05.

## 1 Introduction

Acquiring information on building damage is vital for effective crisis management in the aftermath of an earthquake. Numerous studies have employed spatial regression models, utilizing field-based data, to gather this crucial information. A kriging regression model, proposed by Lallemant and Kiremidjian (2013) leverages both remote-sensing and field-based data to predict damage. This model was later expanded by Loos et al. (2020) to incorporate multiple sources of damage data. Despite the ability of the model to consider the spatial correlation of building damage, thereby enhancing prediction accuracy

[*]Speaker: omid.karimi@semnan.ac.ir

beyond that of traditional regression models, it does not conform to the permissible range. Consequently, the predicted damage could potentially be negative or surpass acceptable values. The current study employed a Bayesian Kriging Regression (BKR) methodology as part of a framework proposed to predict damage ratios within acceptable limits in regions affected by earthquakes. This innovative approach not only takes into account the spatial correlation of building damage but also enables damage prediction based on a small sample size with measurement error using a Bayesian approach. This method proves to be timely, cost-effective, and accurate, making it an invaluable tool in the crisis management process following the occurrence of catastrophic events.The methodology section encompasses three predictive models for damage: Probit Regression (PR), Regression Kriging (RK), and BKR. Following this, both simulated and actual datasets from the Sarpol-e Zahab earthquake are presented to evaluate the effectiveness of these models. The final section offers a comprehensive conclusion, summarizing the key findings of the research.

# 2 Methodology

## 2.1 Probit regression

PR is a GLM with a probit link function that is frequently used in traditional regression models for predicting post-earthquake damage. In this model, the number of damaged buildings, $Z$, out of $n$ buildings is distributed according to the binomial distribution given by:

$$f(Z;\mu) = \binom{n}{Z}\mu^{Z}(1-\mu)^{n-Z} \tag{2.1}$$

where $\mu$ is the damage ratio (the number of damaged buildings divided by the total number of buildings). In the PR model, the damage ratio is connected to a linear predictor through the probit link function as (Dobson and Barnett , 2018):

$$\Phi^{-1}(\mu) = \beta_0 + \beta_1 \log(im) \tag{2.2}$$

The link function $\Phi^{-1}$ is the inverse of the cumulative distribution function of the standard normal distribution. log(.) represents the natural logarithm. $im$ is the intensity measure. $\beta_0$ and $\beta_1$ are the regression parameters that can be determined using the Maximum Likelihood Estimation (MLE) approach (Baker , 2015).

## 2.2 Regression kriging

PR predicts the damage ratio under the assumption that the response variable is independent, despite the fact that the damage ratio exhibits spatial correlation (Lallemant and Kiremidjian , 2013). To address this, Loos et al. (2020) utilized RK to predict the damage ratio. The KR model introduces a residual term to the damage ratio predicted by the PR model. In this context, the damage ratio predicted by PR is referred to as drift. The KR model makes separate predictions for drift and residuals, and subsequently integrates them to generate a prediction for damage ratio $D(\mathbf{s}_0)$ at coordinate $\mathbf{s}_0$ as:

$$D(\mathbf{s}_0) = \mu(\mathbf{s}_0) + e(\mathbf{s}_0) \tag{2.3}$$

where the drift, $\mu(\mathbf{s}_0)$, is obtained by the PR and the residual, $e(\mathbf{s}_0)$, at coordinate $\mathbf{s}_0$ is acquired by ordinary kriging. Therefore, equation (2.3) can be rewritten as:

$$D(\mathbf{s}_0) = \Phi(\beta_0 + \beta_1 \log(im(\mathbf{s}_0)) + \sum_{i=1}^{m} \alpha_i e(\mathbf{s}_i) \tag{2.4}$$

where $e(\mathbf{s}_i)$ is the residual at sample site $\mathbf{s}_0$, $m$ is the number of sample sites and $\alpha_i$ is the kriging weight at site $\mathbf{s}_0$ that minimizes prediction error. In the calculation of $\alpha_i$, semivariogram $\gamma(h_{ij})$ is defined to consider the spatial correlation between points $i$ and $j$ which can be obtained as:

$$\gamma(h_{ij}) = \sigma^2[1 - \exp(-\frac{\|h_{ij}\|}{b})] \tag{2.5}$$

where $h_{ij}$ represents the ordinary Euclidean distance norm between sites $i$ and $j$, $\sigma^2$ is known as sill and shows the variance of residuals and $b$ is the range of the model (a distance beyond which $e(\mathbf{s}_i)$ and $e(\mathbf{s}_j)$ are considered uncorrelated).

## 2.3 Bayesian kriging regression

The RK leads to negative values and more than one value for the damage ratio, both of which are unacceptable. Therefore, the current study proposes BKR, in which a spatial residual term is added to the linear predictor of the PR model that is expressed as:

$$\Phi^{-1}(D(\mathbf{s}_0)) = \beta_0 + \beta_1 \log(im(\mathbf{s}_0)) + e(\mathbf{s}_0) \tag{2.6}$$

where $D(\mathbf{s}_0)$ is the predicted damage ratio at site $\mathbf{s}_0$, and $im(\mathbf{s}_0)$ is the earthquake intensity measure at site $\mathbf{s}_0$. $e(\mathbf{s}_0)$ is a residual term at point s0 derived from a Gaussian random field with an $m$-by-1 zero-mean matrix and an $m$-by-$m$ covariance matrix $\mathbf{Q}$, and the covariance matrix $\mathbf{Q}$ is obtained as:

$$\mathbf{Q} = \begin{bmatrix} C(h_{11}) & \dots & C(h_{1m}) \\ \vdots & \ddots & \vdots \\ C(h_{11}) & \dots & C(h_{1m}) \end{bmatrix} \tag{2.7}$$

where $C(.)$ is the covariogram. The relationship between the covariogram and semivariogram can be expressed as:

$$\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h}) \tag{2.8}$$

Using Bayesian inference, the predictive distribution for the BKR model can be determined as:

$$\begin{aligned} &\pi(Z(\mathbf{s}_0)|\mathbf{Z}(\mathbf{s}), \mathbf{n}(\mathbf{s}), \mathbf{im}(\mathbf{s}), n(\mathbf{s}_0), im(\mathbf{s}_0)) \propto \\ &\quad \int p(Z(\mathbf{s}_0)|\mathbf{Z}(\mathbf{s}), n(\mathbf{s}_0), im(\mathbf{s}_0), \boldsymbol{\eta})\pi(\boldsymbol{\eta}|\mathbf{Z}(\mathbf{s}), \mathbf{n}(\mathbf{s}), \mathbf{im}(\mathbf{s}))d\boldsymbol{\eta} \end{aligned} \tag{2.9}$$

where $\mathbf{Z}(\mathbf{s})$, $\mathbf{n}(\mathbf{s})$ and $\mathbf{im}(\mathbf{s})$ are vectors representing the damaged buildings, the total number of buildings and the intensity measure at m observed site, respectively. Thre vector of model parameters is denoted as $\boldsymbol{\eta} = (\beta_0, \beta_1, \theta, \tau)$, where $\theta = 1/b$ and $\tau = 1/\sigma^2$ (Ribeiro and Diggle , 2010), $p(Z(\mathbf{s}_0)|\mathbf{Z}(\mathbf{s}), n(\mathbf{s}_0), im(\mathbf{s}_0), \boldsymbol{\eta})$ is the conditional binomial distribution and $\pi(\boldsymbol{\eta}|\mathbf{Z}(\mathbf{s}), \mathbf{n}(\mathbf{s}), \mathbf{im}(\mathbf{s}))$ is the posterior distribution of model parameters. To completely determine the prior distributions of parameters, the current study uses the empirical Bayes approach to estimate the hyper-parameters of the prior distribution from the data. As can be seen, the complexity of the posterior distributions of equation (2.9) precludes analytical posterior inferences about the interested parameters. Therefore, a Markov Chain Monte Carlo (MCMC) method should be employed to estimate the parameters via samples obtained from the posterior distribution. In this direction, a Gibbs sampler algorithm can be used to sample from the posterior distribution (S.Geman and Geman , 1984).

# 3 Material

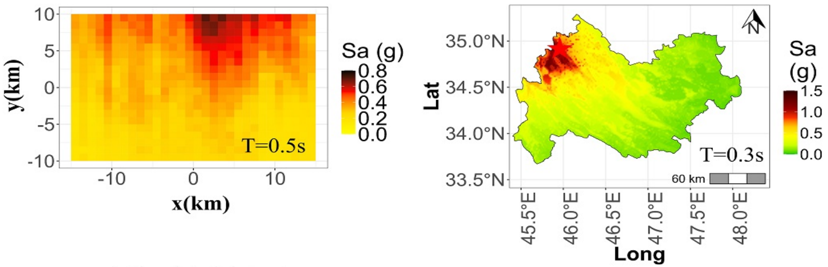
The proposed BKR approach was evaluated using both simulated and actual datasets from the Sarpol-e Zahab earthquake, which are described below.

### 3.1 Generating the datasets of the damage data

The simulated dataset incorporated the earthquake Intensity Measure (IM), along with the total number of buildings and the number of damaged buildings in the areas. A synthetic random field, representing the spatial distribution of the earthquake IM in terms of Sa(T=0.5s), was generated over a 30 km by 20 km area. This was based on a hypothetical M7.3 earthquake event (Figure 1). See (Abbasnejadfard et al. , 2020) for other characteristics of the source and site. Spatially correlated random numbers, $e(\mathbf{s}) \sim N(\mathbf{0}, \mathbf{Q})$, were produced following a normal distribution. The covariance matrix of $\mathbf{Q}$ was calculated using the covariogram function with $\sigma^2 = 2$ and $\theta = 0.1$. The damage ratio was calculated based on regression parameters of $\beta_0 = 1$ and $\beta_1 = 2$ and other variables were defined using equation (2.6).

### 3.2 Actual data from the Sarpol-e Zahab earthquake

The actual datasets included the earthquake IM, along with the total number of buildings and the number of damaged buildings in Kermanshah province, Iran. The shake map of Sa (T=0.3 s) was acquired from the United States Geological Survey (2017) and is depicted in Figure 1. The damage ratios were obtained using the damaged buildings (Management and Planning Organization , 2018) and the total number of buildings (Statistical Center of Iran , 2016).



a) Simulated dataset

b) Actual dataset

Figure 1: Earthquake intensity measures.

# 4 Results

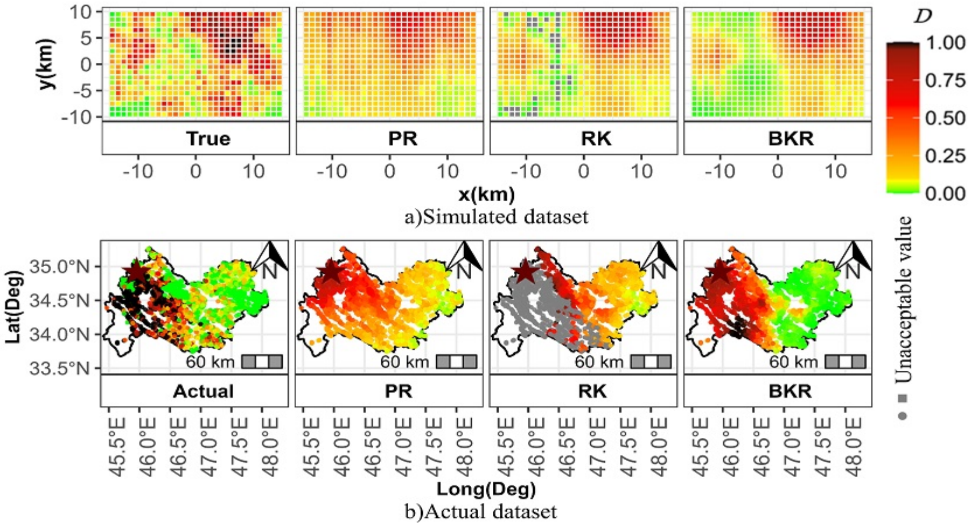
The BKR model has been compared with the other models. For this purpose, the regression parameters of $\beta_0$ and $\beta_1$ were estimated using the MLE method, and the correlation parameters $\tau$ and $\theta$ were determined by fitting an exponential variogram for the PR and RK. This was based on a sample size of 20 for simulated datasets and 65 for actual datasets. For the BKR, the estimated regression parameters of $\beta_0$ and $\beta_1$, along with expert opinion, were used to determine the prior distributions of parameters. The estimated

Table 1: Estimated parameters and prior distribution of parameters for models.

| Simulation | | | Actual | | |
|---|---|---|---|---|---|
| PR | RK | BKR | PR | RK | BKR |
| $\beta_0 = 1.15$ | $\beta_0 = 1.15$ | $\beta_0 \sim N(1.15, 10)$ | $\beta_0 = 1.26$ | $\beta_0 = 1.26$ | $\beta_0 \sim N(1.26, 10)$ |
| $\beta_1 = 1.76$ | $\beta_0 = 1.76$ | $\beta_1 \sim N(1.76, 10)$ | $\beta_1 = 1.53$ | $\beta_0 = 1.53$ | $\beta_1 \sim N(1.53, 10)$ |
| | $b = 7.78$ | $\theta \sim U(0.01, 1)$ | | $b = 27.84$ | $\theta \sim U(0.01, 0.5)$ |
| | $\sigma^2 = 0.06$ | $\tau \sim Ga(0.01, 0.01)$ | | $\sigma^2 = 0.3$ | $\tau \sim Ga(0.01, 0.01)$ |

N, U and Ga are normal, uniform and Gamma distribution, respectively.

parameters and prior distribution are reported in Table 1 for both simulated and actual datasets. After obtaining the model parameters, the damage ratios were predicted for the unsampled site in Figure 2. The analysis revealed that the damage ratios predicted by the



a)Simulated dataset

b)Actual dataset

Figure 2: True and predicted damage ratios.

PR models for the simulated dataset significantly deviated from their true values. This discrepancy arose because the PR model only incorporated the IM in its linear predictor, causing the damage ratio predictions to cluster similarly to the IM pattern. In contrast, the RK and BKR models obtained results that were closer to the true values. These models could predict the true value based on limited samples and taking into account the damage ratios of sampling buildings in the surrounding prediction area. They achieved this by utilizing interpolation between observed damage and spatial correlation between building damage. However, the RK model predicted multiple values for the damage ratio, which were not acceptable. The occurrence of these unacceptable values increased significantly in actual data, likely due to measurement errors in damage data. But the BKR maintained superior prediction performance compared to the RK because it used a Bayesian approach.

# 5 Sensitivity analyses

To ensure the superiority of the BKR model compared to the other models, the bias of prediction across all sites is calculated for 50 iterations as:

$$Bias = \frac{1}{N_{all}} \sum_{site}^{N_{all}} (\hat{D}_{site} - D_{site}) \tag{5.1}$$

where $N_{all}$ is the number of sites, $\hat{D}_{site}$ is the predicted damage ratio and $D_{site}$ is the true damage ratio at a specific site. Figure 3 illustrates the bias of the predicted damage ratio for all prediction models over 50 iterations of sampling. The BKR model exhibit lower bias and standard deviation compared to other models for both datasets. This is because of the use of correlation parameters and interpolations between observed damage in prediction. Furthermore, the bias of prediction for the RK model in the actual dataset increased due to measurement errors in damage data compared to other models.

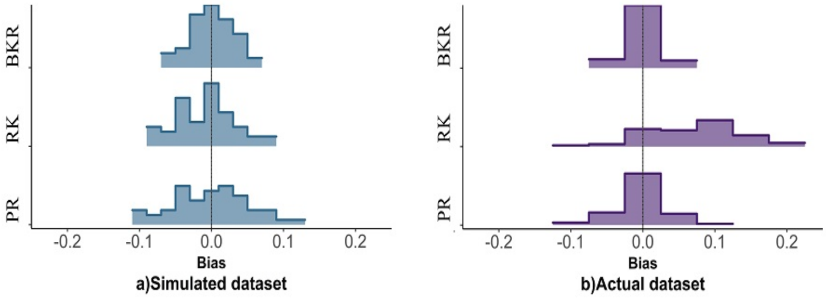


Figure 3: Distribution of Bias for predictions.

# Conclusion

The current study employed a BKR approach as part of a proposed framework to predict the damage ratio in earthquake-affected areas, taking into consideration the spatial correlation of building damage and ensuring the values remain within an acceptable range. The performance of this model was compared with other models using both simulated and actual datasets. Both the RK and BKR models were able to account for the spatial correlations between building damage, resulting in better predictions than the PR model. However, as the measurement error increased, BKR outperformed RK, providing more accurate predictions.

## Acknowledgement

The authors acknowledge the funding and support provided by the International Institute of Earthquake Engineering and Seismology (IIEES) under project No 793 and Iran National Science Foundation (INSF) under project No 4003212.

## References

Abbasnejadfard, M., Bastami, M. and Fallah, A. (2020), Investigation of Anisotropic Spatial Correlations of Intra-Event Residuals of Multiple Earthquake Intensity Measures Using Latent Dimensions Method, *Geophysical Journal International*, **222**, 14491469.

Baker, J. W. (2015), Efficient Analytical Fragility Function Fitting Using Dynamic Structural Analysis, *Earthquake Spectral*, **31**, 579599.

Dobson, A. J., Barnett, A. G. (2018). *An Introduction to Generalized Linear Models*, Taylor and Francis Group, Boca Raton.

Geman, S., Geman, D. (1984), Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images, *IEEE Transactions on Pattern Analysis and Machine Intelligencel*, **6**, 721741.

Lallemant, D., Kiremidjian, A. (2013), Rapid Post-Earthquake Damage Estimation Using Remote-Sensing and Field-Based Damage Data Integration, *In: Proceedings of the 11th International Conference on Structural Safety and Reliability*, New York.

Loos, S., Lallemant, D., Baker, J., McCaughey, J., Yun, S.-H., Budhathoki, N., Singh, R. (2020), G-DIF: A Geospatial Data Integration Framework to Rapidly Estimate Post-Earthquake Damage, *Earthquake Spectra*, **36**, 16951718.

Management and Planning Organization. (2018), *The 12 November 2017 Sarpol-e Zahab Earthquake Lessons Learned*, Kermanshah, Iran.

Ribeiro, P. J., Diggle, P. J. (2010), Bayesian Inference in Gaussian Model-Based Geostatistics, *TECHNICAL REPORT ST-99-08*, Lancaster University.

Statistical Center of Iran. (2016), *Population and Housing Censuses*, https://www.amar.org.ir/english/Population-and-Housing-Censuses.

United States Geological Survey. (2017), *ShakeMap*,https://earthquake.usgs.gov/earthquakes/eventpage/us2000bmcg/.

دانشگاه بین المللی امام خمینی (ره) با همکاری
دانشگاه تربیت مدرس و انجمن آمار ایران برگزار می‌کند:
conf.ikiu.ac.ir/spatial5
پنجمین سمینار
آمار فضایی و کاربردهای آن
5th Seminar on Spatial Statistics and Its Applications
۳ و ۴ آبان‌ماه ۱۴۰۲
Oct. 2023 25-26

# Machine-Learning Models for Predicting the Class of Divorce Cases in Iranian Judiciary Courts

Elham Tabrizi[1]*, Mohadeseh Alsadat Farzammehr[2]

[1]Department of Mathematics, Faculty of Mathematics and Computer Science,
Kharazmi University, Tehran, Iran.

[2]Judiciary Research Institute, Tehran, Iran.

**Abstract:**

This paper introduces a machine-learning model for predicting divorce case outcomes in Iranian Judiciary Courts, leveraging various classification algorithms, including Naïve Bayes, Multinomial Logistic Regression, kNN, Decision Tree, Random Forest, GraBoost, AdaBoost, Neural Network, SGD, and SVM. It utilizes historical divorce case data and socioeconomic indicators like literacy rate, urbanization rate, and employment status. Comparative analysis reveals that the Random Forest classifier achieves the highest accuracy. Additionally, the study highlights key factors linked to divorce cases in Iran, including the population aged 15 and over, unemployment rate, urbanization rate, and participation rate. These findings offer valuable insights for crafting more effective policies and interventions to address the social and economic challenges associated with divorce in Iran.

**Keywords:** Divorce Cases, Data Mining, Machine Learning Techniques, Iran, Judiciary.

**Mathematics Subject Classification (2010):** 62M30, 62H30, 62N05.

# 1 Introduction

In recent years, Iran has experienced a substantial increase in divorce rates, growing from 8.7 per 1,000 marriages in 2006 to 20.8 per 1,000 marriages in 2020, as reported by the Statistical Center of Iran (2020). This surge has raised concerns among policymakers and the broader society due to the significant social and economic consequences of divorce on families and society at large. Accurate prediction of divorce trends within civil courts has

*Speaker: elham.tabrizi@khu.ac.ir

become essential for anticipating the demand for legal services, resource allocation, policy planning, and support program development for families navigating divorce. Predictive modeling, utilizing machine learning techniques, offers a promising approach to forecast divorce trends by considering a wide range of socioeconomic factors and identifying influential predictors. These insights are invaluable for informed policy decisions and tailored interventions to assist at-risk families (Rosili et al. , 2021).

Machine learning algorithms, grounded in statistical and computational techniques, excel in detecting patterns and relationships within extensive datasets, making predictions based on historical data. In predicting divorce trends, these algorithms leverage historical divorce rates and socioeconomic data to construct predictive models capable of foreseeing future trends. Machine learning offers distinct advantages, including heightened accuracy and the ability to handle large and intricate datasets, enabling the discovery of critical predictors not readily apparent through traditional methods (Narendran et al. , 2021; Sharma et al. , 2021).

In conclusion, the integration of machine learning into predictive modeling equips Iranian civil courts, policymakers, and practitioners with insights into future divorce case volumes, facilitating the development of more effective policies and interventions to address the intricate social and economic challenges associated with divorce. The reviewed literature underscores the multifaceted nature of divorce prediction, encompassing socioeconomic, demographic, and behavioral factors, and highlights the potential of machine learning algorithms in enhancing prediction accuracy and informing decision-making in this context.

## 2 Data and Methodology

Our study employs a comprehensive dataset comprising 49 features. By harnessing this extensive dataset, our aim is to achieve precise forecasts of divorce case volumes in Iranian Judiciary courts. Some of the features are unemployment rate, population aged 15 and over, literacy rate in the population aged 6 and over, participation rate, number of cases related to drugs, number of cases related to alcoholic beverages, number of cases related to theft that require punishment, Gini coefficient - rural areas, Gini coefficient - urban areas, consumer price index (= annual inflation rate) , gross domestic product (at market price in billion rials), total added value of 18 sectors (at market price in billion rials), average age of men's first marriage, average age of women's first marriage, share of provinces in total migration, urbanization rate, and internet penetration rate for population aged 15 to 24. The dataset for this experiment includes 217 instances gathered from the Iranian Statistics Center and the Judiciary Statistics and Information Technology Center. To provide more clarity, a new variable named 'Divorce Category' was introduced. This nominal variable, with values 'Low,' 'Medium,' and 'High,' depends on the percentage of divorce-related cases. If 'Divorce Court Cases' is below 33%, it falls into the 'Low' category; if it's between 33% and 66%, it's labeled 'Medium,' and if it's 66% or higher, it's categorized as 'High.' Careful calculations were made for all 217 instances.

In this study, data mining is applied to classify divorce case levels using a subset of high-level features to enhance classifier performance. Various machine learning techniques, such as Neural Networks, Naïve Bayes, and Decision Trees, are employed to create predictive models for divorce cases based on socioeconomic factors. Metrics like AUC, CA, F1, Precision, and Recall are used to assess model performance, with Orange software facilitating implementation. The dataset is split into training and testing sets, and k-fold cross-validation ensures model generalization. ROC curves help evaluate diagnostic test accuracy in categorizing divorce levels, with a larger AUC indicating better discriminative ability. The study's model evaluation and selection process identifies the most suitable machine learning technique for predicting divorce case volumes, complemented by experiments with four feature selection algorithms to improve robustness.

# 3 Analysis and Performance Evaluation

## 3.1 Optimal Model Choice

Table 1 unveils a comprehensive ranking of our top-performing models, meticulously evaluated against essential metrics like AUC, CA, F1, precision, and recall. It also provides insights into their performance across four distinct target classes: average over classes, Low, Medium, and High. Intriguingly, when we scrutinize the AUC scores, the standout performers on both training and test data emerge as Random Forest and Neural Network. These models showcased exceptional predictive capabilities. However, the story takes an interesting twist when we delve into CA, F1, Precision, and Recall. While AdaBoost exhibits superior accuracy on training data when compared to Random Forest and Neural Network, it's the latter two that shine on the test data, signifying their robust predictive prowess. This nuanced performance variation underscores the importance of a comprehensive evaluation. Our findings also unveil that the 'High' target class reaps the most promising results, while 'Medium' class encounters more challenges. Importantly, there are no telltale signs of overfitting, as all models deliver consistent and reliable classification outcomes. In essence, our experiments confirm that all ten prediction models demonstrate impressive performance, with AUC values consistently exceeding 0.88. This robust performance reaffirms the potential of these models for accurate and dependable predictions in divorce case classification.

Table 1: Performance metrics of the nine data mining models

| Model (Average Over Classes) | | kNN | Decision Tree | SVM | SGD | Random Forest | Neural Network | Naïve Bayes | Multinomial Logistic Regression | GraBoost | AdaBoost |
|---|---|---|---|---|---|---|---|---|---|---|---|
| AUC | Train | 0.9525 | 0.9233 | 0.9695 | 0.9274 | 0.981 | 0.9718 | 0.9683 | 0.9277 | 0.9727 | 0.9401 |
| | Test | 0.9663 | 0.9313 | 0.9571 | 0.9034 | 0.9905 | 0.9787 | 0.9652 | 0.9197 | 0.9813 | 0.8886 |
| CA | Train | 0.8355 | 0.8553 | 0.9013 | 0.9013 | 0.9145 | 0.9013 | 0.8487 | 0.8092 | 0.8947 | 0.9211 |
| | Test | 0.8462 | 0.8923 | 0.8769 | 0.8615 | 0.9385 | 0.9385 | 0.8923 | 0.8 | 0.8308 | 0.8308 |
| F1 | Train | 0.836 | 0.8562 | 0.9013 | 0.901 | 0.9143 | 0.901 | 0.8483 | 0.8093 | 0.8942 | 0.9205 |
| | Test | 0.8472 | 0.894 | 0.8798 | 0.8681 | 0.9393 | 0.9393 | 0.8925 | 0.8074 | 0.8397 | 0.8397 |
| Precision | Train | 0.8431 | 0.8584 | 0.9015 | 0.903 | 0.9142 | 0.9014 | 0.8505 | 0.8099 | 0.8944 | 0.9233 |
| | Test | 0.8488 | 0.9082 | 0.9046 | 0.8859 | 0.9429 | 0.9429 | 0.8944 | 0.8186 | 0.8903 | 0.8903 |
| Recall | Train | 0.8355 | 0.8553 | 0.9013 | 0.9013 | 0.9145 | 0.9013 | 0.8487 | 0.8092 | 0.8947 | 0.9211 |
| | Test | 0.8462 | 0.8923 | 0.8769 | 0.8615 | 0.9385 | 0.9385 | 0.8923 | 0.8 | 0.8308 | 0.8308 |
| Model (Low) | | kNN | Decision Tree | SVM | SGD | Random Forest | Neural Network | Naïve Bayes | Multinomial Logistic Regression | GraBoost | AdaBoost |
| AUC | Train | 0.9459 | 0.9327 | 0.9652 | 0.9073 | 0.981 | 0.9635 | 0.972 | 0.9262 | 0.982 | 0.928 |
| | Test | 0.9556 | 0.9038 | 0.9477 | 0.8782 | 0.9877 | 0.9773 | 0.9586 | 0.9103 | 0.9921 | 0.8462 |
| CA | Train | 0.875 | 0.9276 | 0.9276 | 0.9079 | 0.9474 | 0.9079 | 0.875 | 0.8684 | 0.9408 | 0.9539 |
| | Test | 0.8615 | 0.9231 | 0.8769 | 0.8923 | 0.9385 | 0.9385 | 0.9077 | 0.8308 | 0.8769 | 0.8769 |
| F1 | Train | 0.7765 | 0.8791 | 0.8791 | 0.8542 | 0.913 | 0.8511 | 0.7816 | 0.7778 | 0.9011 | 0.9195 |
| | Test | 0.8235 | 0.8936 | 0.8261 | 0.8571 | 0.92 | 0.92 | 0.88 | 0.7843 | 0.8182 | 0.8182 |
| Precision | Train | 0.8462 | 0.8889 | 0.8889 | 0.82 | 0.913 | 0.8333 | 0.8293 | 0.7955 | 0.9111 | 0.9756 |
| | Test | 0.84 | 1 | 0.95 | 0.913 | 0.9583 | 0.9583 | 0.9167 | 0.8 | 1 | 1 |
| Recall | Train | 0.7174 | 0.8696 | 0.8696 | 0.8913 | 0.913 | 0.8696 | 0.7391 | 0.7609 | 0.8913 | 0.8696 |
| | Test | 0.8077 | 0.8077 | 0.7308 | 0.8077 | 0.8846 | 0.8846 | 0.8462 | 0.7692 | 0.6923 | 0.6923 |
| Model (Medium) | | kNN | Decision Tree | SVM | SGD | Random Forest | Neural Network | Naïve Bayes | Multinomial Logistic Regression | GraBoost | AdaBoost |
| AUC | Train | 0.9349 | 0.8813 | 0.9573 | 0.885 | 0.9736 | 0.9593 | 0.9587 | 0.8916 | 0.9804 | 0.9183 |
| | Test | 0.9384 | 0.909 | 0.9104 | 0.8599 | 0.9811 | 0.951 | 0.9272 | 0.8221 | 0.965 | 0.8662 |
| CA | Train | 0.8355 | 0.8553 | 0.9013 | 0.9013 | 0.9145 | 0.9013 | 0.8487 | 0.8158 | 0.8947 | 0.9211 |
| | Test | 0.8615 | 0.8923 | 0.8769 | 0.8615 | 0.9385 | 0.9385 | 0.8923 | 0.8 | 0.8308 | 0.8308 |
| F1 | Train | 0.7967 | 0.8136 | 0.8696 | 0.8624 | 0.885 | 0.8649 | 0.8067 | 0.7586 | 0.8571 | 0.8947 |
| | Test | 0.6897 | 0.7742 | 0.7647 | 0.7273 | 0.8667 | 0.8667 | 0.7586 | 0.5806 | 0.7027 | 0.7027 |
| Precision | Train | 0.7424 | 0.7869 | 0.8621 | 0.9038 | 0.8929 | 0.8889 | 0.7742 | 0.7458 | 0.8727 | 0.8947 |
| | Test | 0.6667 | 0.7059 | 0.65 | 0.6316 | 0.8125 | 0.8125 | 0.7333 | 0.5294 | 0.5652 | 0.5652 |
| Recall | Train | 0.8596 | 0.8421 | 0.8772 | 0.8246 | 0.8772 | 0.8421 | 0.8421 | 0.7719 | 0.8421 | 0.8947 |
| | Test | 0.7143 | 0.8571 | 0.9286 | 0.8571 | 0.9286 | 0.9286 | 0.7857 | 0.6429 | 0.9286 | 0.9286 |
| Model (High) | | kNN | Decision Tree | SVM | SGD | Random Forest | Neural Network | Naïve Bayes | Multinomial Logistic Regression | GraBoost | AdaBoost |
| AUC | Train | 0.9859 | 0.9436 | 0.998 | 0.995 | 0.9991 | 1 | 0.9937 | 0.9599 | 1 | 0.9755 |
| | Test | 0.997 | 0.975 | 1 | 0.96 | 1 | 1 | 0.999 | 0.999 | 0.982 | 0.9475 |
| CA | Train | 0.9605 | 0.9276 | 0.9737 | 0.9934 | 0.9671 | 0.9934 | 0.9737 | 0.9342 | 0.9539 | 0.9671 |
| | Test | 0.9692 | 0.9692 | 1 | 0.9692 | 1 | 1 | 0.9846 | 0.9692 | 0.9538 | 0.9538 |
| F1 | Train | 0.9375 | 0.8842 | 0.9592 | 0.9899 | 0.9495 | 0.9899 | 0.9592 | 0.898 | 0.9307 | 0.9515 |
| | Test | 0.96 | 0.9615 | 1 | 0.9583 | 1 | 1 | 0.9804 | 0.9583 | 0.9388 | 0.9388 |
| Precision | Train | 0.9574 | 0.913 | 0.9592 | 0.98 | 0.94 | 0.98 | 0.9592 | 0.898 | 0.9038 | 0.9074 |
| | Test | 0.96 | 0.9259 | 1 | 1 | 1 | 1 | 0.9615 | 1 | 0.9583 | 0.9583 |
| Recall | Train | 0.9184 | 0.8571 | 0.9592 | 1 | 0.9592 | 1 | 0.9592 | 0.898 | 0.9592 | 1 |
| | Test | 0.96 | 1 | 1 | 0.92 | 1 | 1 | 1 | 0.92 | 0.92 | 0.92 |

In the comprehensive evaluation of all models, it becomes apparent that the Random Forest model and Neural Network consistently outperform their counterparts in accurately predicting the classification of divorce cases within the Iranian Judiciary Courts.

## 3.2 Feature Subset Selection and Model Evaluation

In the course of model application, a series of experiments were conducted to identify the most impactful feature sets. Testing encompassed a spectrum of outcomes, ranging from individual high-impact features to sets of 49 features. This comprehensive analysis leveraged five distinct feature selection algorithms, meticulously recording the ensuing classification results.

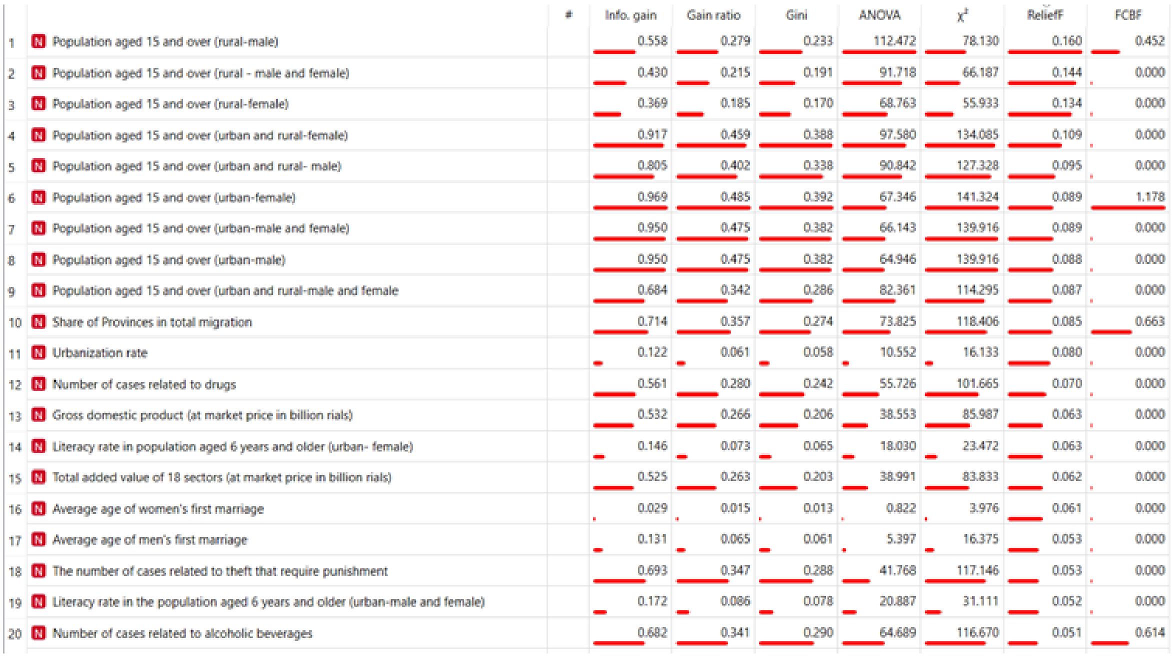
In this subsection, our objective is to rank features by assigning scores based on their correlation with the discrete target variable. We employed various internal scoring methods, including information gain, chi-square, and others, to assess each feature's significance.

The resulting sequence of influential features, obtained through diverse feature selection methods, is presented in Figure 1, along with the algorithm-assigned scores. For instance, according to the ReliefF algorithm, the foremost feature is identified as ”Population aged 15 and over (rural - male),” while the least impactful feature based on the same algorithm is ”Consumer price index.” Conversely, the Information Gini algorithm designates ”Population aged 15 and over (urban - female)” as the most potent feature, with ”Participation rate (rural - female)” having the least impact within the same algorithm.

Similarly, the Gain Ratio algorithm identifies ”Population aged 15 and over (urban - female)” as the most influential feature, while deeming ”Participation rate (rural - female)” as the least influential. Lastly, in accordance with the Chi-Square algorithm, the feature ”Population aged 15 and over (urban - female)” is ranked as the most effective, with ”Unemployment rate (urban - male and female)” being considered the least effective.

After identifying the sequences of impactful features, these attributes were subjected to classification using a variety of algorithms. As inferred from subsection 5.1, it becomes evident that both the random forest and neural network algorithms outperform the others in terms of performance.

In the initial phase, employing the chi-square feature selection algorithm, we exclusively utilized the first identified effective feature for modeling purposes, leveraging the random forest and neural network algorithms. Subsequently, we assessed accuracy based on the test data. This process was iterated for the first two features, followed by the first three features, and so forth, until all the features were incorporated. The outcomes of these evaluations are visually presented in Figure 2. It is notable that the highest success rate was achieved with the Random Forest algorithm, boasting an AUC of 1. This



| | | # | Info. gain | Gain ratio | Gini | ANOVA | $\chi^2$ | ReliefF | FCBF |
|---|---|---|---|---|---|---|---|---|---|
| 1 | N Population aged 15 and over (rural-male) | | 0.558 | 0.279 | 0.233 | 112.472 | 78.130 | 0.160 | 0.452 |
| 2 | N Population aged 15 and over (rural - male and female) | | 0.430 | 0.215 | 0.191 | 91.718 | 66.187 | 0.144 | 0.000 |
| 3 | N Population aged 15 and over (rural-female) | | 0.369 | 0.185 | 0.170 | 68.763 | 55.933 | 0.134 | 0.000 |
| 4 | N Population aged 15 and over (urban and rural-female) | | 0.917 | 0.459 | 0.388 | 97.580 | 134.085 | 0.109 | 0.000 |
| 5 | N Population aged 15 and over (urban and rural- male) | | 0.805 | 0.402 | 0.338 | 90.842 | 127.328 | 0.095 | 0.000 |
| 6 | N Population aged 15 and over (urban-female) | | 0.969 | 0.485 | 0.392 | 67.346 | 141.324 | 0.089 | 1.178 |
| 7 | N Population aged 15 and over (urban-male and female) | | 0.950 | 0.475 | 0.382 | 66.143 | 139.916 | 0.089 | 0.000 |
| 8 | N Population aged 15 and over (urban-male) | | 0.950 | 0.475 | 0.382 | 64.946 | 139.916 | 0.088 | 0.000 |
| 9 | N Population aged 15 and over (urban and rural-male and female | | 0.684 | 0.342 | 0.286 | 82.361 | 114.295 | 0.087 | 0.000 |
| 10 | N Share of Provinces in total migration | | 0.714 | 0.357 | 0.274 | 73.825 | 118.406 | 0.085 | 0.663 |
| 11 | N Urbanization rate | | 0.122 | 0.061 | 0.058 | 10.552 | 16.133 | 0.080 | 0.000 |
| 12 | N Number of cases related to drugs | | 0.561 | 0.280 | 0.242 | 55.726 | 101.665 | 0.070 | 0.000 |
| 13 | N Gross domestic product (at market price in billion rials) | | 0.532 | 0.266 | 0.206 | 38.553 | 85.987 | 0.063 | 0.000 |
| 14 | N Literacy rate in population aged 6 years and older (urban- female) | | 0.146 | 0.073 | 0.065 | 18.030 | 23.472 | 0.063 | 0.000 |
| 15 | N Total added value of 18 sectors (at market price in billion rials) | | 0.525 | 0.263 | 0.203 | 38.991 | 83.833 | 0.062 | 0.000 |
| 16 | N Average age of women's first marriage | | 0.029 | 0.015 | 0.013 | 0.822 | 3.976 | 0.061 | 0.000 |
| 17 | N Average age of men's first marriage | | 0.131 | 0.065 | 0.061 | 5.397 | 16.375 | 0.053 | 0.000 |
| 18 | N The number of cases related to theft that require punishment | | 0.693 | 0.347 | 0.288 | 41.768 | 117.146 | 0.053 | 0.000 |
| 19 | N Literacy rate in the population aged 6 years and older (urban-male and female) | | 0.172 | 0.086 | 0.078 | 20.887 | 31.111 | 0.052 | 0.000 |
| 20 | N Number of cases related to alcoholic beverages | | 0.682 | 0.341 | 0.290 | 64.689 | 116.670 | 0.051 | 0.614 |

Figure 1: Effective feature orders obtained according to different feature selection algorithms

achievement was reached by utilizing the first 46 features. It becomes evident that the chi-square algorithm's initial 46 features, classified by the Random Forest algorithm, yield the highest levels of AUC, CA, F1, Precision, and Recall.

## 3.3 Feature Importance Analysis

Feature importance is a pivotal component of data mining algorithms. It involves quantifying the unique contribution of each feature to the predictive accuracy of the model. This metric not only enhances our comprehension of the underlying data relationships but also plays a pivotal role in feature selection. It enables us to pinpoint key attributes that exert the most substantial influence on the model's overall performance. Figure 3 illustrates the outcomes of our feature importance assessment. In this analysis, our dataset forms the bedrock for evaluating the significance of individual features in relation to predictive outcomes. By doing so, we effectively sever the inherent link between the feature and the target variable, allowing us to gain insight into its genuine impact on prediction accuracy. Based on the AUC, a standout feature emerges: the percentage of the urban population aged 15 or above residing in the province. This feature exhibits a strong correlation with the divorce rate, suggesting a direct impact of urbanization on divorce rates. In essence, variations in divorce rates may be attributed to cultural disparities between urban and rural lifestyles within the country. Shifting our focus to gender-related factors,

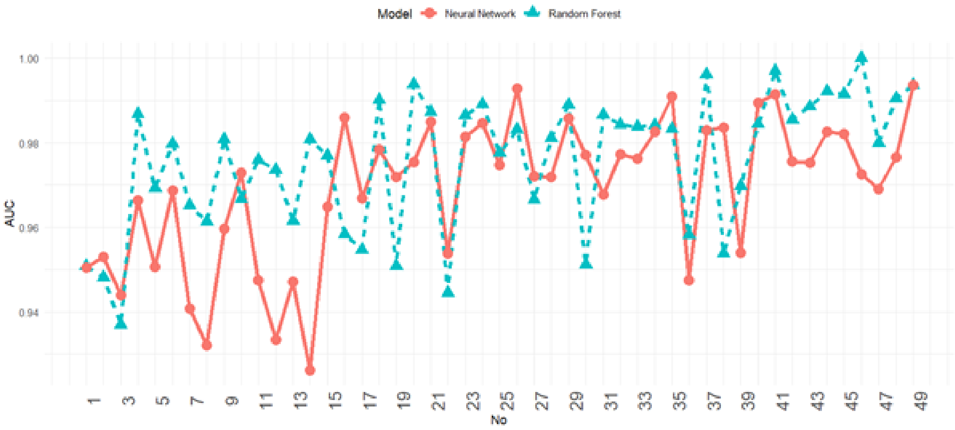


Figure 2: Feature count vs. AUC performance in Random Forest and Neural Network algorithms: unveiling the impact of feature selection

provinces with a higher percentage of female residents tend to experience elevated divorce rates. Among other influential attributes, noteworthy mentions include urban and rural unemployment rates, as well as the literacy rate among rural women.

# Conclusion

This paper conducts an extensive comparative analysis of ten classification algorithms, encompassing Neural Network, Naïve Bayes, Multinomial Logistic Regression, AdaBoost, GraBoost, Random Forest, Decision Tree, kNN, SGD, and SVM, for predicting the ’Divorce Category’ attribute with labels ’Low’, ’Medium’, and ’High’. The experimental results unambiguously reveal Random Forest and Neural Network as superior performers among the algorithms when applied to the divorce dataset. This assertion is grounded in rigorous evaluations employing 10-fold cross-validation. The implications of these findings extend to law enforcement agencies, underscoring the potential advantages of leveraging machine learning algorithms like Random Forest for effective divorce management. Notably, the utilization of feature selection algorithms demonstrates a notably positive and favorable impact compared to employing all features. This observation holds particular significance, given the widespread adoption of numerous features in prior studies on divorce case diagnosis. The challenge of identifying genuinely influential features has persisted, and this study contributes significantly to addressing this issue. Future research endeavors entail the application of spatiotemporal classification algorithms to the divorce dataset, with a specific focus on evaluating prediction performance for Iranian provinces. Additionally, exploring alternative techniques for feature selection and investigating their effects on the prediction performance of different algorithms represents promising avenues

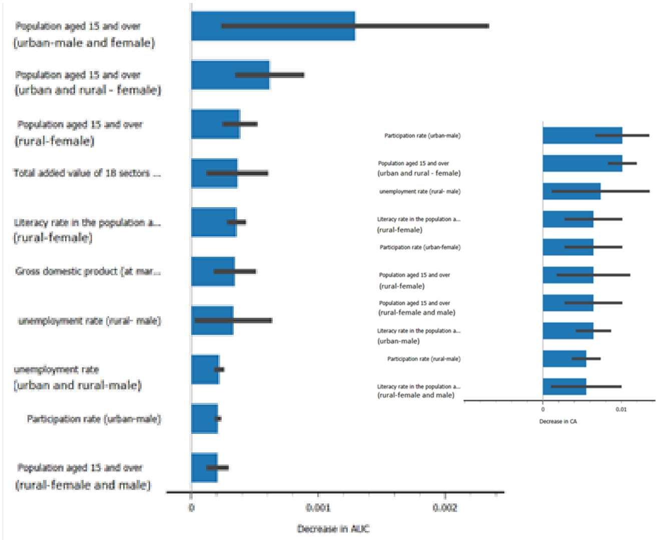


Figure 3: Features importance based on AUC and CA scores in the Random Forest algorithm.

for further exploration in this domain. While this study aimed to predict the volume of court cases related to divorce using data mining models, the limited quantity of available data may have influenced the precision of our models, particularly in predicting different levels of divorce. To enhance the accuracy of such models for categorical variables with multiple levels, future researchers should prioritize the management and collection of a larger, more diverse dataset of historical divorce cases, encompassing a sufficient number of cases for each level of the categorical variable. This approach will facilitate improved model training and validation, ultimately leading to more precise predictions.

# References

Narendran, D. J., Abilash, R., & Charulatha, B. S. (2021). Exploration of classification algorithms for divorce prediction. In *Proceedings of International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications: ICMISC 2020* (pp. 291-303). Springer Singapore.

Rosili, N. A. K., Zakaria, N. H., Hassan, R., Kasim, S., Rose, F. Z. C., & Sutikno, T. (2021). A systematic literature review of machine learning methods in predicting court decisions. *IAES International Journal of Artificial Intelligence*, **10**(4), 1091.

Sharma, A., Chudhey, A. S., & Singh, M. (2021). Divorce case prediction using machine learning algorithms. In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)* (pp. 214-219).

# Variable Selection in Spatial Regression Models Using Boosting Algorithm

Sedigheh Zamani Mehreyan*
Department of Statistics, Imam Khomeini International University, Qazvin, Iran.

**Abstract:**

Boosting algorithm is a learning method that overcomes the weaknesses of machine learners. This method is used for classification and regression. This method reduces the error by combining it in parallel or sequentially and correcting the classification. In this paper, we proposed a boosting algorithm based on the maximum likelihood function for variable selection in spatial regression models. We studied the performance of this algorithm and compared it with usual variable selection methods using simulation studies.

**Keywords:** Akaike information criterion, Boosting algorithm, Spatial regression model, Variable selection.

**Mathematics Subject Classification (2010):** 62J05, 62F07, 62R07, 62M30.

## 1 Introduction

The concept of boosting has been widely applied to various pattern classification problems in machine learning. Schapire (1990) introduced boosting method as a general method to combine multiple classifiers to improve the overall classification accuracy for almost any type of learning algorithm. Schapire (1999) formulated the adaptive boosting as a novel ensemble learning (model combination) algorithm. The first widely used boosting algorithm is adaptive boosting, which successfully solves binary classification problems. Buhlmann and Yu (2003) proposed $L_2$ boosting that builds a linear model by minimizing the $L_2$ loss. The $L_2$ boosting is computationally simple and successful if the learner is sufficiently weak. Friedman (2001), introduced the gradient boosting machine. The gradient boosting is a generalization of adaptive boosting and $L_2$ boosting. The details

*Speaker: zamani@sci.ikiu.ac.ir

of boosting for nonlinear time series models are discussed by Robinzonov et al. (2012). They discussed two methods of component-wise boosting: linear weak learner and p-spline weak learner. Audrino and Buhlmann (2016) discussed volatility estimation via functional gradient descent for high-dimensional financial time series.

Model choice and variable selection are issues of significant concern in practical regression analysis. Model-based boosting is a tool to fit a statistical model while performing variable selection simultaneously. Buhlmann and Hothorn (2007) proposed a boosting algorithm for estimation and variable selection in regression models. Wolfson (2011) obtained a modification of the standard boosting (or functional gradient descent) technique for variable selection and prediction, which can be applied in high-dimensional settings where inference for low-dimensional parameters would typically be based on estimating equations.

In this paper, a boosting algorithm based on the maximum likelihood function is proposed for variable selection in spatial regression models. The performance of this algorithm has been compared with usual variable selection methods using simulation studies. The remainder of the paper is organized as follows: In Section 2, the spatial regression model is introduced. In Section 3, the boosting algorithm for spatial regression models is given based on the model selection criteria, such as the Akaike information criterion. In Section 4, the performance of the boosting algorithm is studied using criteria such as the Akaike information criterion, the mean-squared error of the prediction and the relative frequency of variable selection.

## 2 Spatial regression model

We start with the standard linear regression model

$$y = \mathbf{x}\beta + \epsilon,$$

where $y$ is an $(n \times 1)$ vector of observations on a dependent variable taken at each of $n$ locations, $\mathbf{x}$ is an $(n \times k)$ matrix of explanatory variables, $\beta$ is an $(k \times 1)$ vector of parameters, $\epsilon$ is an $(n \times 1)$ normally distributed vector of disturbances with zero mean, fixed variance $\sigma^2$ and identity matrix $I$.

We consider $F(\mathbf{x})$ as the function of interest and minimize an objective function. If we consider maximum likelihood estimation, the objective function $\mathcal{C}$ can be defined as the log likelihood function of the regression model based on all training data as:

$$\mathcal{C}(F(\mathbf{x})) = \sum_{i=1}^{n} L(y_i; F(\mathbf{x}_i))$$

where $n$ is the number of training samples. The maximum likelihood estimator of the unknown parameter $\beta$ can be calculated as

$$\hat{\beta} = (\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{X}^{'}y.$$

We need to use a modified maximum likelihood method to get over the high-dimension problem because the matrix $(\mathbf{X}^{'}\mathbf{X})$ is not invertible, so the ordinary maximum likelihood method falls down. The idea of boosting is to use only one explanatory variable at a time. Since only one variable is used in one particular iteration, the matrix $x_j^{'}x_j$ is a scalar and thus invertible. Several alternative forms of spatial regression models use spatial weights matrices to represent spatial processes. We consider the spatial error models. In spatial error models, the spatial autocorrelation instead affects the covariance structure of the random disturbance terms. Andrews (2005) suggested a theoretical framework based on common shocks as a mechanism to motivate spatially correlated errors. Spatial error autocorrelation is a particular case of a non-spherical error covariance matrix in which the off-diagonal elements are non-zero, i.e., $E[\epsilon\epsilon^{'}] = \Sigma$. The spatial covariance structure can be obtained in several ways. One of the approaches obtains structure for the error covariance matrix by specifying a spatial process for the random disturbance. The most common choice is a spatial autoregressive process, or SAR:

$$\begin{aligned} y &= \mathbf{x}\beta + \epsilon \\ \epsilon &= \lambda W\epsilon + u, \end{aligned} \tag{2.1}$$

With a row-standardized spatial weights matrix (i.e., the weights standardized such that $\sum_j w_{ij} = 1, \forall i$), $\lambda$ is the autoregressive parameter and $u$ is an $(n \times 1)$ normally distributed vector of the error term, typically assumed to be i.i.d. So that the error variance-covariance matrix follows as follows:

$$E[\epsilon\epsilon^{'}] = \sigma^2(I - \lambda w)^{-1}(I - \lambda w^{'})^{-1}.$$

Using the standard result for a multivariate normal distribution and taking into account the Jacobian term, the log-likelihood for the spatial error autocorrelation model follows as follows:

$$L = -\frac{n}{2}\log(2\pi) - \frac{1}{2}log \mid \Sigma_\theta \mid -(y - \mathbf{X}\beta)^{'}\Sigma_\theta^{-1}(y - \mathbf{X}\beta)$$

where $\Sigma_\theta = \sigma^2(I - \lambda w)^{-1}(I - \lambda w^{'})^{-1}$ and $\theta = (\lambda, \sigma^2)$. So

$$\hat{\beta} = [\mathbf{X}^{'}(I - \hat{\lambda}W)^{'}(I - \hat{\lambda}W)\mathbf{X}]^{-1}\mathbf{X}^{'}(I - \hat{\lambda}W)^{'}(I - \hat{\lambda}W)y$$

Unlike the time series counterpart, a consistent estimate for  cannot be obtained from a simple auxiliary regression. Still, the first-order condition must be solved explicitly by numerical means.

## 3 Boosting in spatial regression model

In this section, we propose the boosting algorithm for variable selection in a spatial regression model. The model (2.1) can be represented as

$$\begin{aligned} &y = \mathbf{x}\beta + \epsilon = \sum_{k=1}^{K} c_k f_k(x) + \epsilon = F_K(x) + \epsilon, \\ &\epsilon = \lambda W \epsilon + u, \end{aligned}$$

where $K$ is the number of variables. At each stage of the boosting algorithm, a new variable is added to the previous model with $k-1$ variable to grow into a new model with $k$ variable. The general procedure of the proposed boosting algorithm is described as

- Step 1: Initialize $F_0(x)$
- Step 2: For $m = 1$ to $M$

  $\{c_m^*, f_m^*\} = argmax_{c_m,f_m}\mathcal{C}(F_m(\mathbf{x}))$  
  Continue to add the new variable?  
  Yes: $F_m(x) = F_{m-1}(x) + c_m^* f_m(x)$  
  No, Go to Step 4
- Step 3: Go to Step 2
- Step 4: Output final $F_M(x) = \sum_{m=1}^{M} c_m f_m(x)$

A new variable and its weight optimally, as in Step 2, is derived using the functional gradient method, see Kim and Pavlovic (2007). When a new variable is added, hopefully, it will increase the objective function concerning $F$ as much as possible:

$$\mathcal{C}(F_{m-1}(\mathbf{x}) + c_m f_m(x)) > \mathcal{C}(F_{m-1}(\mathbf{x}))$$

From the functional Taylor expansion of $\mathcal{C}(F_{m-1}(\mathbf{x}) + c_m f_m(x))$ around $c_m = 0$ we obtain

$$\begin{aligned} \mathcal{C}(F_{m-1}(\mathbf{x}) + c_m f_m(x)) &= \mathcal{C}(F_{m-1}(\mathbf{x})) + c_m \langle \bigtriangledown \mathcal{C}(F_{m-1}(\mathbf{x})), f_m(x) \rangle + O\left(||c_m f_m(x)||\right) \\ &\approx \mathcal{C}(F_{m-1}(\mathbf{x})) + c_m \langle \bigtriangledown \mathcal{C}(F_{m-1}(\mathbf{x})), f_m(x) \rangle \end{aligned}$$

where $\bigtriangledown\mathcal{C}(F_{m-1}(\mathbf{x})) = \frac{1}{F_{m-1}(\mathbf{x})}$ and $\langle P, Q\rangle$ is a functional inner product space based on an inner product between any two models $P$ and $Q$ using the available training sample $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ as

$$\langle P, Q\rangle = \frac{1}{n}\sum_{i=1}^{n} P(\mathbf{x}_i)Q(\mathbf{x}_i).$$

Therefore basic idea of proposed boosting is to learn each $f_m(x)$ incrementally as:

$$\begin{aligned} f_m^*(x) &= argmax_{f_m} \langle \bigtriangledown\mathcal{C}(F_{m-1}(\mathbf{x})), f_m(x)\rangle \\ &= argmax_{f_m} \sum_{i=1}^{n} \frac{f_m(x_i)}{F_{m-1}(x_i)} \end{aligned}$$

It shows that the new component $f_m(x)$ is estimated, where the objective function grows the most. Also in Step 2, $c_m^*$ can be obtained using

$$c_m^* = argmax_{c_m}\mathcal{C}(F_{m-1}(\mathbf{x}) + c_m f_m(x))$$

# 4 Simulation Analysis

In this section, we examine by simulation the relative performance of the material presented so far in the paper. In particular, we examine the performance of the proposed boosting algorithm for variable selection and estimation of spatial regression models. We consider model

$$\begin{aligned} y_i &= 2 + 0.9x_{1,i} + \epsilon_i, \quad i = 1, \cdots, n \\ \epsilon_i &= 0.6\sum_j w_{ij}\epsilon_j + u_i \end{aligned} \tag{4.1}$$

as true model, where $u_i$'s independently and identically distributed as $N(0, 1)$. The observations are generated from true models, where $x_{1,i}$'s are generated from Uniform distribution $U(0, 1)$. We will ignore the true model and consider $x_1, x_2, x_3, x_4$ and $x_5$ as explanatory variables, where $x_1, x_2, x_3, x_4$ and $x_5$, are generated from Uniform distribution, $U(0, 1)$, Normal distribution, $N(0, 2)$, Gamma distribution, $G(2, 5)$, Weibull distribution, $W(5, 3)$, and Uniform distribution, $U(-2, 2)$, respectively. The fitted model results based on the proposed boosting algorithm and maximum likelihood estimation, MLE, are presented in Table 1. Package ”spatial-reg” is used for maximum likelihood estimators. The results show that the estimators obtained from the boosting algorithm and the maximum likelihood method are the same. In other words, the boosting algorithm performs well in variable selection.

Now, we study the case where the observations are generated from a spatial regression

Table 1: The variable selection and parameter estimation for Model 4.1

| n | Method | $\hat{\lambda}$ | $\hat{B}_0$ | $\hat{B}_1$ | $\hat{\mu}$ | $\hat{\sigma}^2$ | $AIC$ |
|---|---|---|---|---|---|---|---|
| 50 | Boost | 0.6773 | 2.1688 | 0.7902 | -9.363e-18 | 0.86420 | 148.7749 |
|  | MLE | 0.6772 | 2.1688 | 0.7902 | 3.2423e-17 | 0.9807 | 148.7749 |
| 100 | Boost | 0.5260 | 1.7177 | 1.0397 | -2.4333e-16 | 0.9160 | 290.5116 |
|  | MLE | 0.5260 | 1.7177 | 1.0397 | -3.1452e-16 | 1.0054 | 290.5116 |
| 250 | Boost | 0.5294 | 1.8759 | 0.8259 | -4.6129e-16 | 0.8474 | 682.3100 |
|  | MLE | 0.5293 | 1.8758 | 0.8258 | -4.9408e-16 | 0.8736 | 682.3100 |
| 500 | Boost | 0.5462 | 1.9050 | 0.8442 | -2.7104e-16 | 0.9184 | 1441.169 |
|  | MLE | 0.5461 | 1.9053 | 0.8441 | -7.5387e-16 | 0.9323 | 1441.171 |
| 1000 | Boost | 0.5794 | 1.9957 | 0.8550 | -1.1536e-15 | 0.9820 | 2832.276 |
|  | MLE | 0.5992 | 1.9957 | 0.8550 | -1.2896 e-15 | 0.9722 | 2832.323 |

model as:

$$\begin{aligned} y_i &= 2 + 0.9x_{1,i} + 0.4 * x_{2,i} + \epsilon_i, i = 1, \cdots, n \\ \epsilon_i &= 0.6 \sum_j w_{ij}\epsilon_j + u_i \end{aligned} \tag{4.2}$$

where $u_i$'s independently and identically distributed as $N(0, 1)$. The results of the values of estimated parameters based on the proposed boosting algorithm and MLE are presented in Table 2. Table 2 shows that the proposed boosting algorithm performs well in variable selection and estimation.

Table 2: The variable selection and parameter estimation for Model 4.2

| n | Method | $\hat{\lambda}$ | $\hat{B}_0$ | $\hat{B}_1$ | $\hat{B}_2$ | $\hat{\mu}$ | $\hat{\sigma}^2$ | $AIC$ |
|---|---|---|---|---|---|---|---|---|
| 50 | Boost | 0.5674 | 2.0603 | 0.8835 | 0.3319 | 5.8831e-17 | 0.8540 | 147.2651 |
|  | MLE | 0.5665 | 2.0766 | 0.8963 | 0.3278 | -1.6815e-16 | 0.9821 | 150.2592 |
| 100 | Boost | 0.5856 | 2.0133 | 1.0154 | 0.3467 | 2.0122e-16 | 0.9023 | 288.8227 |
|  | MLE | 0.5812 | 2.0144 | 1.0159 | 0.3436 | 1.7589e-16 | 0.9998 | 292.2195 |
| 250 | Boost | 0.6238 | 2.0029 | 0.8618 | 0.3648 | 2.9725e-16 | 0.9563 | 569.9535 |
|  | MLE | 0.6462 | 2.0099 | 0.8599 | 0.3638 | 4.0032e-16 | 0.9873 | 574.3331 |
| 500 | Boost | 0.6175 | 1.9985 | 0.9245 | 0.3826 | 6.6389e-16 | 1.0090 | 1438.1950 |
|  | MLE | 0.6469 | 1.9073 | 0.9339 | 0.3761 | 5.9725e-16 | 1.0295 | 1442.7890 |
| 1000 | Boost | 0.6063 | 1.9993 | 0.8917 | 0.3968 | -2.2106e-16 | 0.9915 | 2834.548 |
|  | MLE | 0.5974 | 1.9964 | 0.8836 | 0.4006 | -2.0355e-16 | 0.9920 | 2838.6530 |

We perform $10^4$ replications. The mean values of the Akaike information criterion, $AIC$, mean-squared error of the prediction, $MSE_h$, and the relative frequency, $RF$ of variable selection, are given in Table 3. The results of Table 3 are interesting. We observe that the relative frequency of variable selection for multivariate spatial regression models has decreased compared to univariate spatial regression models.

# Discussion and Results

We proposed a boosting algorithm for variable selection in spatial regression models. In each iteration, a new variable is added to the model according to an objective function in

Table 3: The relative frequency of variable selection for Model 4.2

| Model | n | $MSE_h$ | $AIC$ | $RF$ |
|---|---|---|---|---|
| 2 | 50 | 0.0203 | 140.6722 | 0.9788 |
| 3 | 50 | 0.0740 | 139.8940 | 0.8449 |
| 2 | 100 | 0.0096 | 282.7290 | 0.9958 |
| 3 | 100 | 0.0595 | 281.8947 | 0.8736 |
| 2 | 250 | 0.0039 | 708.4117 | 0.9999 |
| 3 | 250 | 0.0041 | 707.5330 | 0.9092 |
| 2 | 500 | 0.0017 | 1416.9575 | 1.0000 |
| 3 | 500 | 0.0017 | 1416.9575 | 0.9794 |
| 2 | 1000 | 0.0016 | 2835.6664 | 1.0000 |
| 3 | 1000 | 0.0016 | 2835.6664 | 1.9920 |

order to maximize the objective function. In this paper, the likelihood function (Akaike information criterion) is considered as the objective function. In other words, in each iteration, a new variable is added to the model so that the likelihood function (Akaike information criterion) of the new model increases (decreases) compared to the likelihood function (Akaike information criterion) of the current model. The performance of the proposed algorithm of variable selection in spatial regression models was evaluated using simulated data. The simulation results show that the boosting algorithm performs well in variable selection and estimation.

# References

Andrews, D. W. (2005), Cross-section Regression with Common Shocks. *Econometrica*, **73**, 1551-1585.

Audrino, F., Buhlmann, P. (2016), Volatility Estimation with Functional Gradient Descent for Very High-dimensional Financial Time Series. *The Journal of Computational Finance*, **6**(3), 65-89.

Buhlmann, P., Hothorn, T. (2007), Boosting Algorithms: Regularization, Prediction and Model Fitting, *Statistical Science* , **22**(4), 477-505.

Buhlmann, P., Yu, B. (2003), Boosting with the $L_2$ Loss: Regression and Classification, *Journal of the American Statistical Association*, **98**(462), 324-339.

Friedman, J. H. (2001), Greedy Function Approximation: a Gradient Boosting Machine, *The Annals of Statistics*, **29**, 1189-1232.

Kim, M., Pavlovic, V. (2007). A Recursive Method for Discriminative Mixture Learning, *in Proc. ICML*, 409-416.

Robinzonov, N., Tutz, G., Hothorn, T. (2012), Boosting Techniques for Nonlinear Time Series Models. *AStA Advances in Statistical Analysis*, **96**(1), 99-122.

Schapire, R. (1990), The Strength of Weak Learnability, *Mach. Learn*, **5**, 197-227.

Schapire, R. (1999) , Theoretical Views of Boosting and Applications, *In Proc. Comput. Learn. Theory*, 1-10.

Wolfson, J. (2011). EEBoost: A General Method for Prediction and Variable Selection Based on Estimating Equations, *Journal of the American Statistical Association*, **106**(493), 296-305.

Imam Khomeini International University,
Faculty of Science, Statistics, Qazvin, Iran
Postal Code: 34148-96818
Phone: +98 (28) 3390 1547
Fax: +98 (28) 3378 6579
email: spatial5@ikiu.ac.ir