


Exploring Discrete Spatial Heterogeneity Across Quantiles: A Combination Approach of Generalized Lasso and Conditional Quantile Regression

Ryo Inoue ✉ 

Graduate School of Information Sciences, Tohoku University, Sendai, Japan

Kenya Aoki ✉

Graduate School of Information Sciences, Tohoku University, Sendai, Japan

Abstract

Spatial heterogeneity has been investigated extensively. However, in addition to spatial heterogeneity, there are spatial phenomena where heterogeneity in the data generation process exists across quantiles. This study proposes a new method that combines generalized lasso (GL) and conditional quantile regression (CQR) to analyze discrete spatial heterogeneity across quantiles. GL enables the identification of spatial boundaries where the spatial data generation process varies discretely, and CQR estimates the parameters of the conditional quantile of the dependent variable. The proposed method is expressed as a linear programming problem and is simple to use. To validate its effectiveness, we applied this method to apartment rent data in Minato Ward, Tokyo. The results revealed that the neighborhoods where rent levels deviated from the overall trend in the analyzed area differed by quantiles.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases discrete spatial heterogeneity, generalized lasso, conditional quantile regression

Digital Object Identifier 10.4230/LIPIcs.COSIT.2024.12

Category Short Paper

Funding This study was funded by the Japan Society for the Promotion of Science KAKENHI, Grant Numbers JP21H01447, JP24K00997, and JP24K00175.

Acknowledgements We used “At Home Dataset” provided by At Home Co.,Ltd. via IDR Dataset Service of National Institute of Informatics.

1 Introduction

Spatial heterogeneity refers to the variation in the process of generating spatial phenomena across locations. Many researchers have analyzed this property using the abundant spatial data available in recent years. There are two analytical approaches that focus on spatial heterogeneity with different assumptions. The first approach assumes that the generative process of spatial phenomena changes continuously within the domain of analysis domain. Geographically Weighted Regression (GWR) [5] and Eigenvector Spatial Filtering-based Spatially Varying Coefficient (ESF-SVC) model [6] are examples of the analytical approach.

The second approach assumes that the generative process of spatial phenomena changes discretely at certain spatial boundaries. This study focuses on this assumption. Recent studies on the detection of discrete spatial heterogeneity (e.g., [14, 3, 7, 11]) use the generalized lasso (GL) [13], a sparse modeling technique. They preset regions that are the minimum spatial units for detecting discrete spatial heterogeneity, build a regression model with region-specific parameters, and set l_1 regularization on the region-specific parameters and the differences of neighboring region-specific parameters to search for the set of regions with the same degree of heterogeneity.



© Ryo Inoue and Kenya Aoki;
licensed under Creative Commons License CC-BY 4.0

16th International Conference on Spatial Information Theory (COSIT 2024).

Editors: Benjamin Adams, Amy Griffin, Simon Scheider, and Grant McKenzie; Article No. 12; pp. 12:1–12:8



Leibniz International Proceedings in Informatics

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Most previous studies have analyzed the spatial heterogeneity structure corresponding to the mean of the conditional distribution of the dependent variable. However, there are spatial phenomena where heterogeneity exists not only in space but also across quantiles. Therefore, we are interested in capturing the spatial heterogeneity corresponding to the different quantiles of the conditional distribution of the dependent variable.

Although limited in number, there are studies that analyze heterogeneity in space and across quantiles. As an analysis for continuous spatial heterogeneity, Chen et al. [2] proposed geographically weighted quantile regression. It combines GWR and conditional quantile regression (CQR) [8], which evaluates the influence of the explanatory variables on the dependent variable at each quantile. However, a method for analyzing discrete spatial heterogeneity has not yet been discussed.

In this study, we propose a novel method that combines GL and CQR under the assumption that spatial phenomena vary in a discrete manner. Our goal is to identify localized areas that deviate from the overall trend of the analyzed area for each quantile.

2 Methodology

2.1 Generalized Lasso (GL) in Discrete Spatial Heterogeneity Analysis

GL is applied to analyze discrete spatial heterogeneity. A regression model is used that includes common parameters that represent the relationship between explanatory and dependent variables in the entire target area, as well as region-specific parameters that indicate deviations from the aforementioned common relationship. Let y_i be the dependent variable at location i , x_{ik} be a k th explanatory variable at location i , and β_k be a k th parameter, where β_0 is an intercept. K denotes the number of parameters, excluding an intercept, R denotes the number of regions, γ_r is a region-specific parameter of region r , d_{ir} denotes a dummy variable indicating whether location i is in region r , and ε_i denotes an error term. The model is written as

$$y_i = \sum_{k=0}^K \beta_k x_{ik} + \sum_{r=1}^R \gamma_r d_{ir} + \varepsilon_i. \quad (1)$$

The purpose of the β s is to express the overall trend between the dependent and explanatory variables for the entire region under analysis. In contrast, the γ s are intended to express the deviation from the overall trend for each region. Then, l_1 regularization is introduced on the region-specific parameters and the differences of neighboring region-specific parameters. The estimation of the model is written as the following minimization problem,

$$\min_{\beta, \gamma} \left[\frac{1}{2} \sum_{i=1}^N \left(y_i - \sum_{k=0}^K \beta_k x_{ik} - \sum_{r=1}^R \gamma_r d_{ir} \right)^2 + \lambda_1 \sum_{r=1}^R |\gamma_r| + \lambda_2 \sum_{(m,n) \in A} |\gamma_m - \gamma_n| \right], \quad (2)$$

where λ_1 and λ_2 denote positive tuning hyperparameters and A denotes the set of combinations of adjacent regions.

The first regularization term is designed to identify region-specific parameters, γ s, that should be non-zero, only when there are spatial heterogeneity and the relationship between dependent and explanatory variables cannot be adequately described by the common parameters, β s. Furthermore, the second regularization term is designed to identify whether neighboring regions exhibit similar degrees of spatial heterogeneity. Thus, the method enables the detection of sets of regions exhibiting discrete spatial heterogeneity.

2.2 Conditional Quantile Regression (CQR)

While ordinary regression analysis estimates parameters for the conditional mean of the dependent variable over the explanatory variables, CQR estimates parameters for the conditional quantiles of the dependent variables. Here, the conditional quantile function $Q_{y_i|\mathbf{x}_i}(\tau)$ satisfying $P[y_i < Q_{y_i|\mathbf{x}_i}(\tau)] = \tau$ is defined, where $\tau \in [0, 1]$ is a quantile and \mathbf{x}_i is a $K \times 1$ vector of the explanatory variables at location i . CQR describes $Q_{Y|\mathbf{X}}(\tau)$ through a linear combination of the explanatory variables as

$$Q_{y_i|\mathbf{x}_i}(\tau) = \sum_{k=0}^K \beta_k^\tau x_{ik}, \quad (3)$$

where β_k^τ is a k th parameter at the τ th quantile. The estimation of β_k^τ is formulated by the following minimization problem using the loss function $\rho_\tau(\cdot)$,

$$\min_{\beta^\tau} \sum_{i=1}^N \rho_\tau \left(y_i - \sum_{k=0}^K \beta_k^\tau x_{ik} \right), \quad \rho_\tau(u) = \begin{cases} \tau u & \text{if } u \geq 0, \\ (\tau - 1)u & \text{if } u < 0. \end{cases} \quad (4)$$

Equation (4) can be reformulated as a linear programming problem. Let \mathbf{y} denotes an $N \times 1$ vector of the dependent variable, \mathbf{X} denotes an $N \times K$ matrix of the explanatory variables, and $\mathbf{1}_N$ and $\mathbf{0}_N$ denote vectors of N ones and zeros, respectively. Then,

$$\begin{aligned} \min_{\beta^+, \beta^-} \quad & \tau \mathbf{1}'_N \mathbf{u} + (1 - \tau) \mathbf{1}'_N \mathbf{v}, \\ \text{s.t.} \quad & \mathbf{y} - \mathbf{X}(\beta^+ - \beta^-) = \mathbf{u} - \mathbf{v} \\ & \beta^+, \beta^- \geq \mathbf{0}_K, \mathbf{u}, \mathbf{v} \geq \mathbf{0}_N \end{aligned} \quad (5)$$

where β^+ and $-\beta^-$ are $K \times 1$ parameter vectors with positive and negative elements, respectively, and \mathbf{u} and $-\mathbf{v}$ are $N \times 1$ error vectors with positive and negative elements, respectively.

3 Discrete Spatial Heterogeneity across Quantiles

This study proposes a method that combines GL and CQR to analyze discrete spatial heterogeneity for each quantile. This approach is similar to the model proposed by [12], which combines fused adaptive lasso [1] and CQR. The previous model establishes a quantile regression model with location-specific parameters and penalizes the difference between the location-specific parameters at a certain location and the average of the parameters at neighboring locations. The model allows for the detection of locations exhibiting quantiles that diverge from the smoothed quantiles of neighboring locations, rendering it an effective tool for hotspot analysis. However, the conventional method fails to account for the discrepancies between adjacent regional parameters, rendering it incapable of identifying boundaries where the generation process of spatial phenomena undergoes a change.

Here, let K be the number of attributes including the intercept and R be the number of regions for the discrete spatial heterogeneity detection. β_k^τ is a parameter at the τ th quantile of the attribute k , γ_r^τ is a parameter at the τ th quantile of region r , and ε_i^τ is the error term at the τ th quantile at location i . The linear model at τ th quantile is expressed by

$$y_i = \sum_{k=0}^K \beta_k^\tau x_{ik} + \sum_{r=1}^R \gamma_r^\tau d_{ir} + \varepsilon_i^\tau. \quad (6)$$

12:4 Exploring Discrete Spatial Heterogeneity Across Quantiles

Then, the GL-based l_1 regularization and the CQR-based loss function are used to estimate the parameters,

$$\min_{\beta^\tau, \gamma^\tau} \left[\sum_{i=1}^N \rho_\tau \left(y_i - \sum_{k=0}^K \beta_k^\tau x_{ik} - \sum_{r=1}^R \gamma_r^\tau d_{ir} \right) + \lambda_1 \sum_{r=1}^R |\gamma_r^\tau| + \lambda_2 \sum_{(m,n) \in A} |\gamma_m^\tau - \gamma_n^\tau| \right]. \quad (7)$$

Equation (7) can also be formulated as the following linear programming problem,

$$\begin{aligned} \min_{\beta^+, \beta^-, \gamma^+, \gamma^-} \quad & \tau \mathbf{1}'_N \mathbf{u} + (1 - \tau) \mathbf{1}'_N \mathbf{v}, \\ \text{s.t.} \quad & \mathbf{y} - \mathbf{X}(\beta^+ - \beta^-) - \mathbf{D}(\gamma^+ - \gamma^-) = \mathbf{u} - \mathbf{v} \\ & \mathbf{1}'_R(\gamma^+ + \gamma^-) \leq s \\ & \mathbf{A}(\gamma^+ - \gamma^-) = \boldsymbol{\theta}^+ - \boldsymbol{\theta}^- \\ & \mathbf{1}'_E(\boldsymbol{\theta}^+ + \boldsymbol{\theta}^-) \leq t \\ & \beta^+, \beta^- \geq \mathbf{0}_K, \gamma^+, \gamma^- \geq \mathbf{0}_R, \boldsymbol{\theta}^+, \boldsymbol{\theta}^- \geq \mathbf{0}_E, \mathbf{u}, \mathbf{v} \geq \mathbf{0}_N \end{aligned} \quad (8)$$

where \mathbf{A} is a $E \times R$ matrix representing adjacency of region-specific parameters, $\boldsymbol{\theta}^+, \boldsymbol{\theta}^-$ are vectors of positive and negative differences between adjacent parameters, respectively.

The tuning hyperparameters, λ_1 and λ_2 in Equation (7) and s and t in Equation (8), must be determined manually. To choose the setting of hyperparameters, the model is evaluated by the Bayesian Information Criterion (BIC) for CQR [9],

$$BIC(\tau) = \log \left(\sum_{i=1}^N \rho_\tau \left(y_i - \sum_{k=0}^K \beta_k^\tau x_{ik} - \sum_{r=1}^R \gamma_r^\tau d_{ir} \right) \right) + m \frac{\log N}{2N}, \quad (9)$$

where m is a number of non-zero parameters.

The hyperparameters s and t in Equation (8) must be optimized using an empirical search procedure such as grid search. It has been pointed out that the lasso estimates have a bias towards zero, which makes the interpretation based on the estimates inappropriate [4]. To overcome this property, the proposed method is first applied by varying the tuning hyperparameters within a certain range to select the candidate models. Then, CQR without l_1 regularization is applied to each of the selected models and the best model that minimizes the BIC at each quantile is selected.

4 Application: Analysis of Apartment Rent in Minato Ward, Tokyo

4.1 Data and Model

The proposed method is applied to apartment rent data to check its effectiveness. The apartment rent of 10,930 properties in Minato Ward, Tokyo in 2017, collected by At Home Co., Ltd. [10] is used. Table 1 summarizes the dependent and explanatory variables. All explanatory variables are log transformed and standardized with mean 0 and variance 1 in the rent model. Since the building age and the number of floors have values of zero, one is added to each before the logarithmic transformation.

Figure 1 shows 112 neighborhoods (*cho* and *chome* in Japanese) that are pre-specified as the spatial units for detecting discrete spatial heterogeneity. The reference neighborhood is selected by the smallest mean absolute value of the residuals from the CQR estimation of the model without neighborhood-specific parameters under the 50th percentile condition.

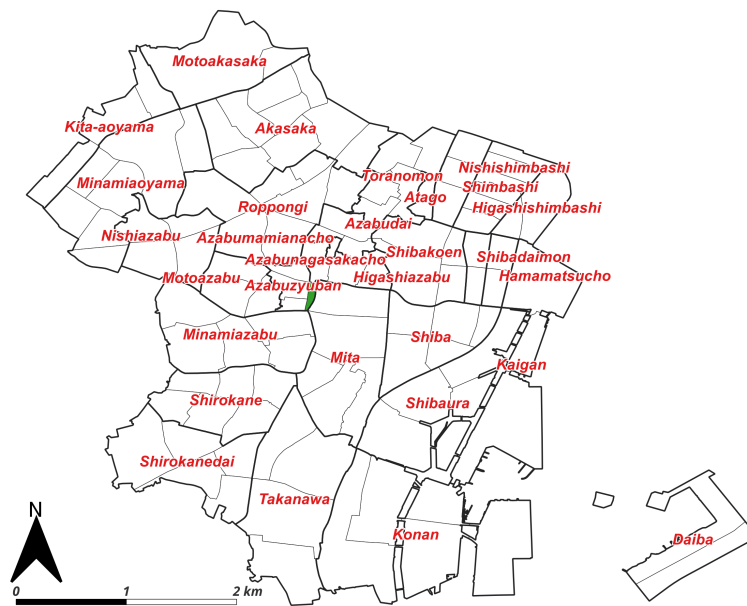
To identify differences in the effects of the explanatory variables by quantile, we analyze at five quantiles ($\tau = [0.1, 0.3, 0.5, 0.7, 0.9]$).

■ **Table 1** Summary of variables.

Variables	Minimum	Maximum	Mean	Standard deviation
Rent per square meter (JPY/m ²)	1686.77	39582.32	4317.75	1086.02
Walking time to the nearest station (min)	0	25	6.08	3.35
Building age (year)	0	35	14.48	7.16
Floor number	1	55	7.91	7.62
Area of property (m ²)	11.9	366.73	47.33	33.60

We set the linear model (Equation (6)) and estimate the parameters by solving the linear programming problem in Equation (8). The estimated neighborhood-specific parameters $\hat{\gamma}_r^\tau$ represent discrete spatial heterogeneity in the rent.

In order to estimate the model, it was necessary to determine the grid search space for the optimal hyperparameters. Therefore, we first set 1, 10, and 100, and then narrowed the interval between these values. Finally, we varied the hyperparameters by 1 for each quantile in the range of 30 to 50, and identified the combination of hyperparameters that resulted in the lowest BIC.



■ **Figure 1** 112 neighborhoods in Minato Ward, Tokyo and reference neighborhood (shown in green).

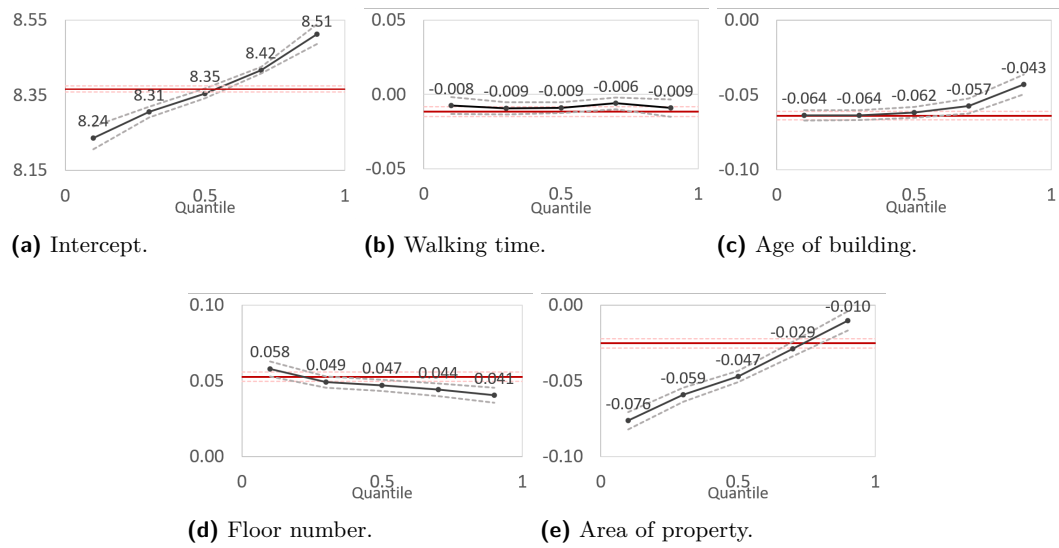
4.2 Results and Discussion

4.2.1 Estimated Parameters for Apartment Attributes

Figure 2 indicates the estimated parameters $\hat{\beta}_k^\tau$ s. The quantile τ and the estimated value are shown on the horizontal and vertical scales, respectively. The estimates at the five different quantiles ranging from 0.1 to 0.9 are shown as black solid lines with filled black dots, and the two dashed gray lines show their 95 percent confidence intervals. The red line in each

figure shows the ordinary least squares (OLS) estimate of the conditional mean effect, and the two dashed red lines represent its 95 percent confidence intervals.

Figure 2 shows that the valuation of housing attributes on rent per square meter varies by quantile. In particular, the proposed method estimates a different parameter for the area of the properties compared to OLS. The estimated parameter is negative for low quantiles, but almost zero for high quantiles (Figure 2e). This suggests that the rent per square meter decreases as the area increases for properties with lower rents, but not for those with higher rents. However, the valuation of walking time to the nearest station does not change across different rent ranges (Figure 2b).



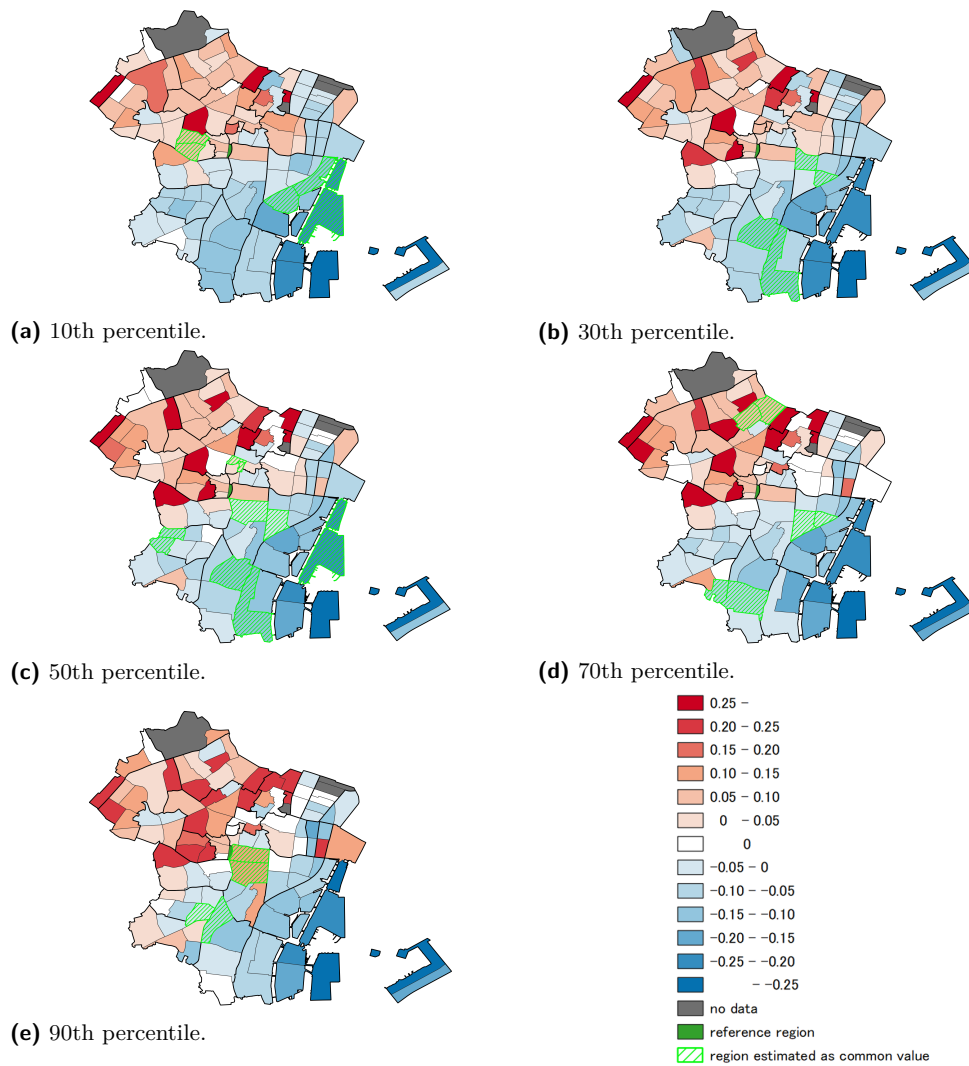
■ **Figure 2** Estimated parameters for apartment attributes.

4.2.2 Estimated Neighborhood-specific Parameters

Figure 3 shows the spatial distribution of estimated neighborhood-specific parameters $\hat{\gamma}_r^T$. Neighborhoods painted in red and blue depict higher and lower rent levels than the reference neighborhood, respectively. Shaded adjacent neighborhoods represent neighborhoods estimated to have the same value.

Figure 3 shows that rents are higher in the northwestern neighborhoods and lower in the southeastern neighborhoods. It seems natural that the northwestern neighborhoods, *Aoyama*, are known for its upscale residential area close to the central business districts, such as *Akasaka* and *Roppongi*. On the other hand, the southeastern neighborhoods, such as *Konan* and *Kaigan*, are less popular due to their proximity to the warehouse districts near the port.

We focus on the eastern neighborhoods where different quantiles yielded different estimation results. The parameters of *Kaigan 1-chome* and *Hamamatsucho 2-chome* are estimated negatively at the lower quantile, but positively at the higher quantile. Although these neighborhoods are in close proximity to the warehouse districts, they also hold many high-rise condominiums with high rents due to their convenient access to transportation and good views of the harbor. The properties of in these neighborhoods exhibit different rent trends compared to those in the surrounding neighborhoods.



■ **Figure 3** Spatial distribution of estimates of neighborhood-specific parameters.

5 Conclusion

We proposed an analysis method for discrete spatial heterogeneity across quantiles by combining GL and CQR. The introduction of GL-based l_1 regularization to the loss function of CQR enabled us to estimate region-specific parameters where some are zero and some adjacent parameters share a common value for each quantile. The proposed method was applied to analyze rent in Minato Ward, Tokyo, and the results confirmed its ability to detect discrete spatial heterogeneity for each quantile.

However, there is an issue that requires attention. It is important to note that CQR does not differentiate between high-end or low-end properties in the entire target area. This implies that the proposed method analyzes the impact of explanatory variables on overvalued or undervalued dependent variables (i.e., dependent variables are higher or lower than they should be) at each location using CQR. In the rent analysis example, we examine the higher quantile estimates to assess common factors contributing to price formation and regional influences on properties that are more expensive in high-end and low-end residential areas.

However, it is reasonable to assume that the attributes and quality of properties may vary between high-end and low-end residential areas, even within the same quantile. Additionally, the evaluation of attributes may vary depending on the location. Therefore, it is necessary to consider a method that can extract differences in attribute evaluation for different areas, even within the same price level.

References

- 1 Rinaldo. A. Properties and refinements of the fused lasso. *The Annals of Statistics*, 37(5B):2922–2952, 2009. doi:10.1214/08-AOS665.
- 2 V. Y.-J. Chen, W.-S. Deng, T.-C. Yang, and S. A. Matthews. Geographically weighted quantile regression (GWQR): An application to U.S. mortality data. *Geographical Analysis*, 44(2):134–150, 2012. doi:10.1111/j.1538-4632.2012.00841.x.
- 3 H. Choi, E. Song, S. Hwang, and W. Lee. A modified generalized lasso algorithm to detect local spatial clusters for count data. *ASIA Advances in Statistical Analysis*, 102:537–563, 2018. doi:10.1007/s10182-018-0318-7.
- 4 J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001. doi:10.1198/016214501753382273.
- 5 A. S. Fotheringham, M. E. Charlton, and C. Brunson. Geographically weighted regression: A natural evolution of the expansion method for spatial data analysis. *Environment and Planning A: Economy and space*, 30(11):1905–1927, 1998. doi:10.1068/a301905.
- 6 D. A. Griffith. Spatial-filtering-based contributions to a critique of geographically weighted regression (GWR). *Environment and Planning A: Economy and space*, 40:2751–2769, 2008. doi:10.1068/a38218.
- 7 R. Inoue, R. Ishiyama, and A. Sugiura. Identifying local differences with fused-MCP: An apartment rental market case study on geographical segmentation detection. *Japanese Journal of Statistics and Data Science*, 3:183–214, 2020. doi:10.1007/s42081-019-00070-y.
- 8 R. Koenker and K. F. Hallock. Quantile regression. *Journal of Economic Perspectives*, 15(4):143–156, 2001. doi:10.1257/jep.15.4.143.
- 9 E. R. Lee, H. Noh, and B. U. Park. Model selection via Bayesian information criterion for quantile regression models. *Journal of the American Statistical Association*, 109(505):216–229, 2014. doi:10.1080/01621459.2013.836975.
- 10 At Home Co. Ltd. At home dataset. *Informatics Research Data Repository, National Institute of Informatics.(dataset)*, 2023. doi:10.32130/idr.13.1.
- 11 R. Masuda and R. Inoue. Point event cluster detection via the Bayesian generalized fused lasso. *ISPRS International Journal of Geo-Information*, 11(3):187, 2022. doi:10.3390/ijgi11030187.
- 12 Y. Sun, H. J. Wang, and M. Fuentes. Fused adaptive lasso for spatial and temporal quantile function estimation. *Technometrics*, 58(1):127–137, 2016. doi:10.1080/00401706.2015.1017115.
- 13 R. Tibshirani and J. Taylor. The solution path of the generalized lasso. *The Annals of Statistics*, 39(3):1335–1371, 2011. doi:10.1214/11-AOS878.
- 14 H. Wang and A. Rodríguez. Identifying pediatric cancer clusters in florida using loglinear models and generalized lasso penalties. *Statistics and Public Policy*, 1(1):86–96, 2014. doi:10.1080/2330443X.2014.960120.