ORIGINAL PAPER

Spatial statistics

Identifying local differences with fused-MCP: an apartment rental market case study on geographical segmentation detection

Ryo Inoue[1] · Rihoko Ishiyama[1] · Ayako Sugiura[2,3]

Abstract

As the variety and quality of spatial data increase in recent times, the potential to analyze local characteristics based on spatial data is getting stronger. Previous spatial analysis methods structuralize the spatial autocorrelation of data by the distances between data observation points and the contiguity of the data-observed regions. It is significant for the estimation of global characteristics of spatial data. However, these approaches are not suitable for identifying local differences from the data since they assume a smooth spatial autocorrelation structure. Generalized fused lasso, which can detect local differences in spatial data, has been proposed in machine learning studies. Its limitation is that the estimated parameters are biased toward zero; however, methods that overcome the limitation have also been proposed. Fused-MCP is one of those methods and is expected to be useful in spatial analyses. This study applies fused-MCP to spatial analyses. As an example of spatial analyses based on fused-MCP, this study analyzes the structure of geographical segmentation of the real estate market in central Tokyo. Fused-MCP is utilized to extract areas where the valuation standard is the same. The results reveal that the geographical segmentation displays hierarchal patterns. Specifically, the market is divided by municipalities, railway lines and stations, and neighborhoods. The case study confirmed the applicability of fused-MCP to spatial analyses.

Keywords Local difference · Fused-MCP · Fused lasso · Geographical segmentation of the real estate market

✉ Ryo Inoue
  rinoue@tohoku.ac.jp

1   Graduate School of Information Sciences, Tohoku University, 6-6-06 Aramaki-Aoba, Aoba, Sendai, Miyagi 980-8579, Japan

2   The Tokyo Association of Real Estate Appraisers, 3-12-1, Toranomon, Minato, Tokyo 105-0001, Japan

3   Phronesis Inc, 1-14-9, Nishi-Shimbashi, Minato, Tokyo 105-0003, Japan

1 Introduction

The recent development of information communication technologies and the progress of open data policies have increased the quality and variety of spatial data rich. Thus, the potential to analyze local characteristics based on spatial data is improving, and the methods for spatial analysis are in the spotlight.

Various methods have been proposed for spatial analysis; however, we limit the discussion to regression-based analyses. It is necessary to consider the spatial autocorrelation of data in the regression analyses. Therefore, three representative approaches—spatial regression models, spatial process models, and geographically weighted regression—have been developed. Spatial regression models (Anselin 1988), which include spatial error models, spatial lag models, and moving average models, aim to estimate common parameters of regression models in the target area from region-based independent and dependent variables, representing the structure of spatial autocorrelation by spatial weight matrices. The elements of a spatial weight matrix are set based on the proximity of regions, such as the contiguity of regions and the distance between regions. Spatial process models (Cressie 1991), which include ordinary kriging, universal kriging, and cokriging, aim to interpolate values at unobserved locations in the target area, representing the spatial autocorrelation structure based on the assumption of intrinsic stationarity or second-order stationarity and estimating the common parameters of regression models. Geographically weighted regression (Fotheringham et al. 2003) aims to estimate the heterogeneous parameters by locations, assuming that the location-based parameters are continuous in space. It is a weighted regression in which the weights are set by the functions of the distance between the location where parameters are estimated and the locations where data are observed.

The above-mentioned methods of spatial data analysis structuralize the spatial autocorrelation of data by the distances between data observation points and the contiguity of data observed regions. The modeling of spatial autocorrelation structure is significant for the estimation of global characteristics of spatial data. However, these approaches are not suitable for finding the local areas where the parameters are discontinuously different from their neighborhood areas since these approaches assume a smooth spatial autocorrelation structure.

Fused lasso is proposed by Tibshirani et al. (2005) as a method to detect change points in time-series data. It is based on the least absolute shrinkage and selection operator (lasso) (Tibshirani 1996). The lasso imposes l_1 regularization on an optimum parameter estimation problem, and it functions as a variable selection method. Fused lasso is an expansion of lasso; it penalizes the parameters themselves and the differences between consecutive parameters. Tibshirani and Taylor (2011) propose a generalized fused lasso, which expands the regularization of differences between consecutive parameters to the regularization of differences between adjacent parameters; the expansion enables one to consider the adjacency of regions in spatial analyses to detect the borders where adjacent parameters are not the same. The generalized fused lasso is applied to spatial

analyses in Wang and Rodríguez (2014) and Inoue et al. (2018); it is confirmed to be a powerful method for detecting local differences.

However, Fan and Li (2001) point out that lasso estimators are biased, and several regularizations whose estimators are not biased have been proposed. One type of regularization is the adaptive lasso proposed by Zou (2006), which sets a different weight on each parameter in the l_1 penalty. The setting of adaptive weights requires initial parameter estimators to allocate small weights on the penalties of non-zero parameters and the large weights on the penalties of zero-parameters. Another type of regularization is the concave penalty function; the smoothly clipped absolute deviation (SCAD) penalty proposed by Fan and Li (2001) and the minimax concave penalty (MCP) proposed by Zhang (2010). However, the minimization of nonconvex penalized function is computationally challenging. The solution's uniqueness and the existence of local minimizers are not discussed for the SCAD penalties. Conversely, the MCP is defined to minimize the maximum concavity and is proven to have a unique path of the solution graph from the origin that all estimators are zero to an optimal fit, although there may be additional branches. Jing et al. (2018) expand the MCP to the fused-MCP, which penalizes the parameters themselves and the differences between adjacent parameters. Fused-MCP is expected to be useful in spatial analyses that aim to extract local differences.

Therefore, this study employs the fused-MCP in spatial analyses that aim to detect local differences in a target area. As an example of spatial analyses based on fused-MCP, this study introduces an analysis to identify the structure of geographical segmentation of an apartment rental market in central Tokyo and confirms the applicability of fused-MCP to the analysis.

2 Variable selection and parameter estimation with penalty functions

Suppose a linear regression model in which an $n \times 1$ dependent variable vector $\mathbf{y}$ depends on p explanatory variables of $n \times 1$ vectors $\mathbf{x}_j$ ($j = 1, \ldots, p$), and β_j is a coefficient for explanatory variable $\mathbf{x}_j$, then

$$\mathbf{y} = \mathbf{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}, \tag{1}$$

where n is the number of observations, $\mathbf{X}_1 \equiv (\mathbf{x}_1, \ldots, \mathbf{x}_p)$, $\boldsymbol{\beta}_1 \equiv (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$, and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.

However, there are P ($> p$) candidates for explanatory variables $\mathbf{x}_k$ ($k = p + 1, \ldots, P$) that does not correlate with $\mathbf{y}$. If there is no prior information on which regressors should be used in the model of Eq. (1), then the model is

$$\mathbf{y} = \mathbf{X}_1 \boldsymbol{\beta}_1 + \mathbf{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2}$$

where $\mathbf{X}_2 \equiv (\mathbf{x}_{p+1}, \ldots, \mathbf{x}_P)$, $\boldsymbol{\beta}_2 \equiv (\beta_{p+1}, \ldots, \beta_P)^{\mathrm{T}} = \mathbf{0}$, $\mathbf{X} \equiv (\mathbf{X}_1, \mathbf{X}_2)$, and $\boldsymbol{\beta} \equiv (\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}}$.

This model presents a variable selection problem of finding out p explanatory variables with non-zero coefficients from P candidates. When the number of variables

P is small, it is possible to find a good set of explanatory variables based on the criteria, such as AIC (Akaike's Information Criterion) (Akaike 1973) and BIC (Bayesian Information Criterion) (Schwarz 1978). However, when the number of candidate explanatory variables, P, is large, the number of combinations of variables becomes enormous. Thus, the variable selection is not computationally feasible. Hence, to avoid the situation, the optimization methods with the regularization by continuous penalty functions are utilized. This section introduces lasso and its derivative, MCP.

2.1 Lasso and fused Lasso

2.1.1 Lasso

The lasso introduces l_1 regularization to the model (Tibshirani 1996). Here, we estimate the linear regression model of Eq. (2) by the lasso. Let $\| \cdot \|_2$ denote the Euclidean norm (l_2-norm) and $\| \cdot \|_1$ denote the l_1-norm. Then, the parameter estimation by lasso is formulated as

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \text{ subject to } \lambda\|\boldsymbol{\beta}\|_1 \leq t, \tag{3}$$

where t is the positive lasso regularization parameter. The lasso promotes the sparsity of the model; it estimates that the parameters where the less substantial explanatory variables are zero. The solution of a constrained optimization problem of Eq. (3) selects the substantial explanatory variables from among many candidates. Equation (3) is equivalent to

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1, \tag{4}$$

where λ is a Lagrange multiplier that corresponds to t. When λ is large (and t is small), many parameters are estimated as zero.

2.1.2 Fused lasso

Tibshirani et al. (2005) expand the lasso to a fused lasso. It is a method to investigate the presence or absence of a difference between consecutive parameters by imposing a new regularization on the differences between consecutive parameters. Tibshirani and Taylor (2011) propose a generalized fused lasso and a path algorithm to estimate parameters. Sparse fused lasso is a type of generalized fused lasso that has the fused lasso penalty, the l_1-norm on the difference between adjacent coefficients, and the l_1-norm of the coefficients themselves. It is formulated as

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2 \sum_{(m,n)\in C} |\beta_m - \beta_n|, \tag{5}$$

where C is the set of pairs of parameters that are defined to have an adjacency. The l_1-norm on the difference among adjacent parameters tends to estimate common parameters. Tibshirani and Taylor (2011) state that the sparse fused lasso gives

an unbiased estimate of the number of nonzero fused groups. Hereafter, the sparse fused lasso of Eq. (5) is denoted as fused lasso in this study.

Since the fused lasso has flexibility in setting the adjacency of parameters, it is easily applicable to spatial analyses by setting adjacency based on location. Tibshirani and Taylor (2011) show an example of applying fused lasso to spatial analyses. A linear regression model with a parameter of the constant term for each US state is estimated from the flu case data. Inoue et al. (2018) apply the fused lasso to an estimation of the linear regression model that represents real estate valuation to analyze the geographical segmentation of the real estate market. In addition to the linear regression models, the fused lasso is utilized in many types of models. Wang and Rodríguez (2014) utilize the fused lasso in a Poisson regression model without covariates to identify the spatial clusters of point events, and Choi et al. (2018) propose to utilize the fused lasso in a Poisson regression model with covariates and an estimation algorithm based on the Majorize–Minimization algorithm (MM algorithm) (Hunter and Li 2005). Parker et al. (2016) utilize it to estimate the spatial covariance function that is not spatially stationary. Sun et al. (2016) introduce a method to estimate spatial and temporal quantile function with a fused adaptive lasso by penalizing the difference between neighboring quantiles.

2.2 MCP and fused-MCP

As described above, the lasso-based variable selection has already been utilized in spatial analyses that aim to detect local differences and indicate its applicability. However, Fan and Li (2001) point out that the estimators by lasso are biased. Therefore, several penalty functions that not only have an ability in variable selection but also can estimate unbiased parameters have been proposed.

This section introduces one of the penalty functions, MCP (Zhang 2010), and its expansion, the fused-MCP (Jing et al. 2018).

2.2.1 MCP

Zhang (2010) proposes the MCP and a penalized linear unbiased selection (PLUS) algorithm that estimates parameters in the minimization of the loss function with MCPs. The MCP is defined as

$$\rho(t; \lambda, \gamma) = \lambda \int_0^t \left(1 - \frac{x}{\gamma \lambda}\right)_+ dx, \tag{6}$$

with regularization hyperparameters $\gamma > 0$ and $\lambda > 0$. The MCP is obtained to minimize the maximum concavity defined by

$$\kappa(\rho) \equiv \kappa(\rho; \lambda, \gamma) \equiv \sup_{0 < t_1 < t_2} \left\{ \dot{\rho}(t_1; \lambda, \gamma) - \dot{\rho}(t_2; \lambda, \gamma) \right\} \big/ (t_2 - t_1), \tag{7}$$

which is subject to

$$\dot{\rho}(t;\lambda,\gamma) = 0 \quad \forall t \geq \gamma\lambda, \tag{8}$$

and

$$\dot{\rho}(0 +;\lambda,\gamma) = \lambda, \tag{9}$$

where $\dot{\rho}(t;\lambda,\gamma) = \frac{d\rho(t;\lambda,\gamma)}{dt}$. Equation (8) presents the condition that an MCP is not effective if t exceeds a threshold of $\gamma\lambda$, and this leads to estimate unbiased parameters when $t \geq \gamma\lambda$. Equation (9) presents the condition that an MCP imposes the same penalty to the l_1 regularization term of $\lambda\|t\|_1$; thus, the MCP works as a variable selection method.

The parameter estimation of the model in Eq. (2) by regularization using the MCP is

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \sum_{i=1}^{P} \rho\bigl(|\beta_i|;\lambda,\gamma\bigr). \tag{10}$$

Equation (10) is a non-convex optimization since the MCP is a concave function. However, since the minimum concavity of Eq. (7) is minimized to $\kappa(\rho) = 1/\gamma$, the MCP provides sparse convexity to the broadest range.

A larger γ value leads to less unbiasedness and more concavity, and the MCP is equal to the l_1 penalty at $\gamma = \infty$ and is similar to the l_0 penalty as $\gamma \to 0+$. Zhang (2010) proves that the MCP has selection consistency at the universal penalty level $\lambda_{\text{univ}} \equiv \sigma\sqrt{(2/n)\log p}$.

2.2.2 Fused-MCP

Jing et al. (2018) propose a fused-MCP, which penalizes both the parameters themselves and their adjacent difference by the MCP. This study focuses on the recovery of 1D and 2D-grid signal with noise and proposes models for each data type, as well as an algorithm to estimate the parameters.

The model for denoising 1D signal is

$$\mathbf{y} = \boldsymbol{\theta} + \boldsymbol{\varepsilon}, \tag{11}$$

where $\mathbf{y}$ is an $n \times 1$ vector of observed signal, $\boldsymbol{\theta}$ is an $n \times 1$ vector of true signal, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of noise. Thus, the recovery of the signal, which is the estimation of $\boldsymbol{\theta}$, is defined as

$$\hat{\boldsymbol{\theta}} = \text{aug}\min_{\boldsymbol{\theta}} \frac{1}{2}\|\mathbf{y} - \boldsymbol{\theta}\|_2^2 + \sum_{i=1}^{n} \rho\bigl(|\theta_i|;\lambda_1,\gamma_1\bigr) + \sum_{i=2}^{n} \rho\bigl(|\theta_i - \theta_{i-1}|;\lambda_2,\gamma_2\bigr). \tag{12}$$

If $\lambda_1 = 0$, Eq. (12) is transformable to the minimization problem that is the same as that of the regularization of parameters themselves by the MCP. Thus, the PLUS algorithm can be applied to estimate the parameters. However, if $\lambda_1 > 0$, it is not possible to apply the PLUS algorithm. Jing et al. (2018) propose

to solve the nonconvex optimization problem by the MM algorithm (Hunter and Lange [2004](); Hunter and Li [2005]())

The MM algorithm is an iterative procedure that consists of majorization and minimization. Consider the problem that minimizes the concave function $f(\boldsymbol{\theta})$. At the m step, the current value of $\boldsymbol{\theta}$, $\boldsymbol{\theta}^{(m)}$, is given. Then, find a surrogate function $g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)})$ that is a convex function and majorizes $f(\boldsymbol{\theta})$:

$$g\big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}\big) \geq f(\boldsymbol{\theta}) \,\forall\boldsymbol{\theta}$$
$$g\big(\boldsymbol{\theta}^{(m)}|\boldsymbol{\theta}^{(m)}\big) = f\big(\boldsymbol{\theta}^{(m)}\big). \tag{13}$$

Then, minimize the surrogate function $g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)})$ to find $\boldsymbol{\theta}^{(m+1)}$. Since the following relationship,

$$f\big(\boldsymbol{\theta}^{(m+1)}\big) \leq g\big(\boldsymbol{\theta}^{(m+1)}|\boldsymbol{\theta}^{(m)}\big) \leq g\big(\boldsymbol{\theta}^{(m)}|\boldsymbol{\theta}^{(m)}\big) = f\big(\boldsymbol{\theta}^{(m)}\big), \tag{14}$$

holds, the estimator approaches the local minima. If a surrogate function is set appropriately, the MM algorithm works efficiently. Jing et al. ([2018]()) proposes a surrogate function of the MCP $\rho(t; \lambda)$ that is a quadratic function with the minimum value at $t = 0$,

$$g_\varepsilon\big(t|t^{(m)}\big) = \rho\big(t^{(m)}\big) + \frac{\big(t^2 - t^{(m)^2}\big)\rho'\big(|t^{(m)}|\big)}{2\big(|t^{(m)}| + \varepsilon\big)}. \tag{15}$$

The model for denoising the 2D-grid signal used in Jing et al. ([2018]()) is

$$\mathbf{Y} = \boldsymbol{\Theta} + \boldsymbol{\varepsilon}, \tag{16}$$

where $\mathbf{Y}$ is an $n \times m$ matrix of an observed 2D-grid type signal (such as an image with noise), $\boldsymbol{\Theta}$ is the $n \times m$ matrix of true signal to estimate the recovered image, and $\boldsymbol{\varepsilon}$ is the $n \times m$ matrix of disturbances at every pixel. Thus, the signal recovery based on MCP is written as

$$\hat{\Theta} = \text{aug}\min_{\boldsymbol{\Theta}} \frac{1}{2}\|\mathbf{Y} - \boldsymbol{\Theta}\|_F^2 + \sum_{i=1}^{n}\sum_{j=2}^{m}\rho\Big(\big|\theta_{i,j} - \theta_{i,j-1}\big|; \lambda, \gamma\Big) + \sum_{i=2}^{n}\sum_{j=1}^{m}\rho\Big(\big|\theta_{i,j} - \theta_{i-1,j}\big|; \lambda, \gamma\Big), \tag{17}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. The model does not penalize parameters but their adjacent difference. The PLUS algorithm is also not applicable to this model; hence, the solution using the MM algorithm is proposed.

2.3 Estimation of a linear regression model by fused-MCP

This study applies the fused-MCP to the estimation of a linear regression model. The parameter estimation is formulated as

$$\hat{\boldsymbol{\beta}} = \text{aug} \min_{\boldsymbol{\beta}} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \sum_{i=1}^{P} \rho\big(|\beta_i|; \lambda_1, \gamma_1\big) + \sum_{(m,n)\in C} \rho\big(|\beta_m - \beta_n|; \lambda_2, \gamma_2\big).$$

(18)

This study implemented the estimation by the MM algorithm using the surrogate function of Eq. (15), following Jing et al. (2018).

3 Case study: identification of geographical segmentation of the apartment rental market in central Tokyo

As an example of the identification of local differences, this study introduces an analysis of geographical segmentation of the real estate market; it aims to detect the areas where the pricing is different from their neighbors.

The real estate market is segmented by many aspects, including consumer types, property types, and environmental factors. It also forms submarkets in which the valuation bases differ from those in others. It is well-known that location plays a major role in the submarket formation. For example, people who value the accessibility to services of urban life prefer to live in city centers, and people who value commodious property and proximity to the natural environment prefer to live in the suburbs. The difference in consumer types between the city center and the suburbs affects property supply and valuation in both submarkets. Families who care about the education of their children choose school districts and municipalities where the level of education service is high. However, couples without a child do not consider educational service for their residential selection.

In this manner, not only the living environment but also the types of consumers and supplied residential properties are related to location. As a result, the real estate market is segmented geographically such that the value of attributes of real estate properties is different for each area. The valuation of attributes of real estate properties differs by neighborhoods, school districts, municipalities, and city center or suburbs. Therefore, the geographical segmentation in the real estate market has a hierarchical structure.

3.1 Previous studies on geographical segmentation of the real estate market

The geographical segmentation of the real estate market has attracted much research interest, and several approaches have been attempted to understand its structure.

One approach is the classification of the market based on the attributes of real estate properties. Bourassa et al. (1999), Wu and Sharma (2012), and Keskin and Watkins (2017) classify the real estate properties based on the principal component analysis of their attributes and analyze the relationship between the classified results and the locational characteristics.

Other approaches are based on the hedonic analysis derived from Rosen (1974). Kestens et al. (2006) utilized geographically weighted regression to estimate the location-dependent parameters of the real estate price model and discuss the spatial

distribution of estimated parameters. However, most hedonic-based studies, such as Watkins (2001), Bourassa et al. (2003), Goodman and Thibodeau (1998, 2003, 2007), and Islam and Asami (2011), preset several types of geographical division units, estimated parameters for each unit in the settings of different geographical divisions, and selected the model that best fit the data to identify the structure of geographical segmentation. Their analyses are limited to the preset division structures, such as school districts, postal districts, and census tracts; they cannot identify the segmentation that is partially different from the preset division, and they do not consider the hierarchal division structure of real estate market.

Kuroda et al. (2014), also based on the hedonic analysis, proposed to estimate the real estate price model and searched the division of submarkets by connecting the adjacent preset geographical units simultaneously by the simulated annealing. It is possible to find the division structure that is not fixed to the preset division. However, there are several limitations. The number of submarkets has to be given when executing the simulated annealing. Thus, if the number of geographical units for the preset division is large and there are many choices for the number of submarkets, it is quite difficult to find the best model that fits the data. Moreover, it is not possible to find the hierarchical division structure.

Thus, to overcome the limitation of previous researches, Inoue et al. (2018) applied the fused lasso for the estimation of a real estate price model. Similar to Kuroda et al. (2014), the structure of the geographical market segmentation is searched by connecting adjacent spatial units of the preset division. The difference is that it can analyze real estate price models with multiple levels of division structure and estimate parameters and division structure simultaneously by solving the optimization without relying on a randomized algorithm. A real estate price model with many regional explanatory variables that depend on different spatial resolutions is constructed, and areas where the valuation standard is the same are extracted by the implementation of the fused lasso. The proposed method was applied to rent data of apartments in the Tokyo metropolitan area, and the result reveals that the valuation standard is different from most of the municipalities. However, there are sets of municipalities where the valuation for the attributes of apartments are the same. The results also detect the neighborhood-level local disparity in the valuation; it, then, confirms that there is a hierarchical structure in the apartment rental market in the target area. However, Inoue et al. (2018) did not consider the lack of unbiasedness in the lasso estimation.

This study utilizes the fused-MCP for the estimation of the apartment rent model used in Inoue et al. (2018), although the target area is reduced due to the limitation in the computation power to solve the optimization problem with MCP.
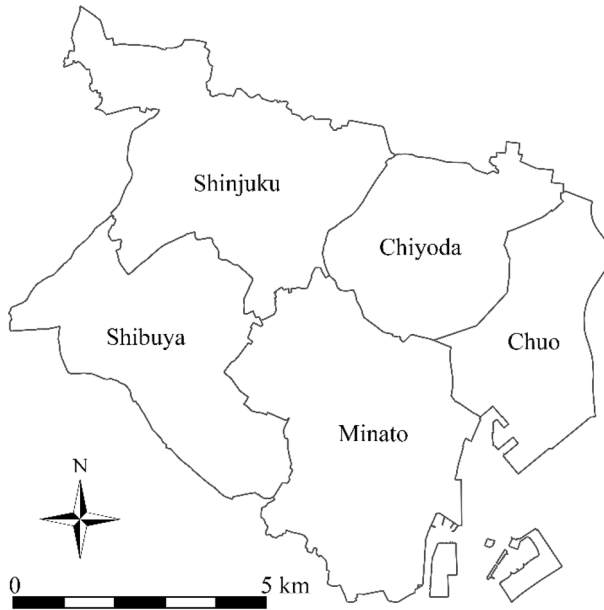
(a) Municipalities (wards, "*ku*" in Japanese)
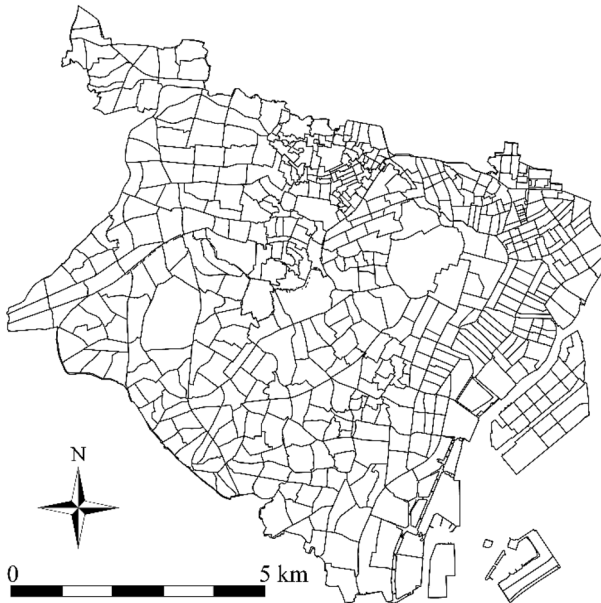
(b) Neighborhoods ("*cho*" and "*chome*" in Japanese)

Fig. 1 Spatial units

3.2 Data and model of apartment rent

This study utilized apartment rent data from 2015 to 2016 collected by At Home Co., Ltd. The target area was central Tokyo, which consists of five municipalities (wards, "*ku*" in Japanese): Chiyoda, Chuo, Minato, Shinjuku, and Shibuya (Fig. 1a).

High-rise condominiums whose number of floors exceeded 15 were excluded as their rents have a different pricing structure from the pricing of other apartments. If the rent data of apartments in the same building occupy the majority of data in the neighborhood, the estimated results of valuation in that neighborhood would be affected by the building. Thus, to avoid such issues, an apartment was randomly selected on the same floor in the same building each year. Consequently, the total number of records was 50,355. Figure 2 shows the spatial distribution of the apartment rent data.

The apartment rent model in this case study used the natural logarithm of monthly rent per square meter as the dependent variable. Five attributes of apartments, apartment age, area of property, walking time to the nearest station, floor number, and number of rooms were the explanatory variables to estimate the municipality-level parameters; in total, 25 parameters were set. The summary of attributes of apartments is shown in Table 1.

Three different levels of location factors that might affect the apartment rental market were considered: railway lines, the nearest railway stations, and neighborhoods ("*cho*" and "*chome*" in Japanese) (Fig. 1b). Dummy variables in this model
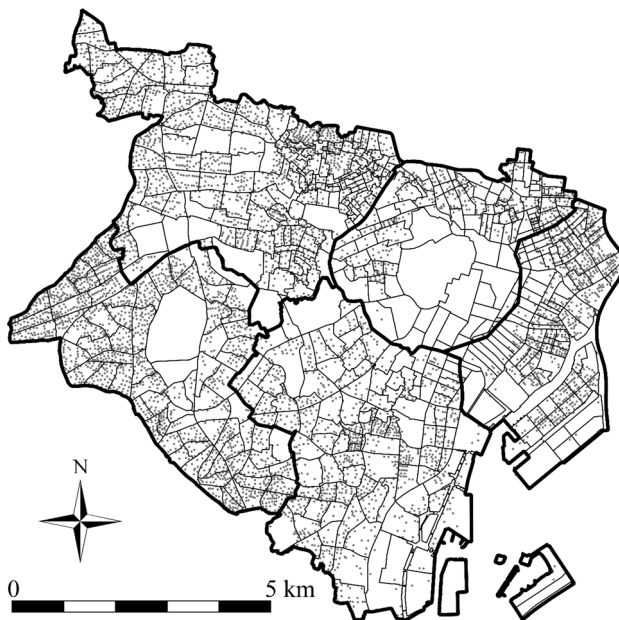

Fig. 2 Spatial distribution of apartment rent data

Table 1 Summary of variables

Variable name	Mean	Standard deviation	Maximum	Minimum
Monthly rent per square meter (JPY)	3854.88	743.46	6998.51	1314.29
Apartment age (year)	20.75	13.02	66.17	0.00
Area of property (m²)	37.20	24.22	440.00	10.00
Walking time to the nearest station (min)	5.39	3.04	41.00	1.00
Floor number	4.75	3.05	15.00	1.00
Number of rooms	1.25	0.55	7.00	1.00

represented these location factors. The detailed descriptions and the numbers of dummy variables are shown in Table 2. Note that a station and a neighborhood whose average rent per square meter were the medians were selected as the reference, and dummy variables were not set for that station and neighborhood.

The following is the explanation of the hedonic apartment rent model settings in this study. Let $y_{i(\text{in } m)}$ denote a dependent variable that is the natural logarithm of rent of apartment i located in municipality m. Furthermore, β_0 denotes the intercept of the regression; β_{jm} denotes the municipality-level regression coefficient of the j-th explanatory variables x_{ijm} of apartment i in municipality m; M, L, S, and N denote sets of municipalities, railway lines, railway stations, and neighborhoods, respectively; x_{il}^{Line}, x_{is}^{Stn}, and x_{in}^{Nbrhd} denote the dummy variables for railway line l that stops at the nearest station, the nearest railway station s, and the neighborhood n of apartment i, respectively; and β_l^{Line}, β_s^{Stn}, and β_n^{Nbrhd} denote the parameters of railway line l, railway station s, and neighborhood n, respectively. Then, the apartment rent model was

$$y_{i(\text{in } m)} = \beta_0 + \sum_{m \in M} \sum_{j \in p_m} \beta_{jm} x_{ijm} + \sum_{l \in L} \beta_l^{Line} x_{il}^{Line} + \sum_{s \in S} \beta_s^{Stn} x_{is}^{Stn} + \sum_{n \in N} \beta_n^{Nbrhd} x_{in}^{Nbrhd} + \varepsilon_i,$$

(19)

where $\varepsilon_i \sim N(0, \sigma^2)$. The total number of parameters was 651.

The following explanations are the settings of the regularization term. This study penalized all the parameters and the differences of municipality-level parameters between adjacent municipalities and between parameters of dummy variables of adjacent neighborhoods. However, the differences between parameters of connected railway lines and adjacent railway stations were not penalized since the parameters of railway line dummies and station dummies in this model were expected to represent the evaluation of accessibility as well as the environmental factors surrounding the station. It is natural to consider that the parameters of connected railway lines and parameters of neighbor railway stations are not necessarily the same.

The adjacency of municipalities was set if two municipalities shared borders. The adjacency of neighborhoods was set if two neighborhoods shared borders and belonged to the same municipalities and the same system of railway stations. Note that whether a certain neighborhood belongs to a certain system of railway

Table 2 Summary of dummy variables

Dummy variable name	Description	Number of variables
Railway line	All railway lines are included	28
	Note that express and local lines are not distinguished, and railway line dummies that are the same to the nearest station dummies are excluded	
Nearest railway station	All railway stations that appear in the nearest station attribute of apartments are included. The railway stations whose names are the same but are operated by different railway companies are aggregated	106
	Reference: Sendagaya station	
Neighborhood (*Cho* in Japanese)	All neighborhoods where apartments appear in the data are included	491
	Note that neighborhood dummies that are the same as some nearest station dummies are excluded	
	Reference: Samon *cho*, Shinjuku ward	

Table 3 Numbers of pairs of adjacent spatial units

Spatial units	Number of pairs
Municipality[a]	7
Neighborhood	594

[a]Five types of municipality-level explanatory variables are used in the model. Therefore, the number of regularization terms is 35 in total

stations was decided based on the highest frequency of appearances in the nearest station attribute of the apartment data in the neighborhood. Since the apartments tend to be allocated to the major stations in their neighborhood, the shapes of systems of stations are different from those of Voronoi polygons whose generators are the stations. This setting constructs a hierarchical structure in the dummy variable settings and enables the analysis of the hierarchical structure of the apartment rental market from municipality to neighborhood levels. The numbers of pairs of adjacency municipalities and neighborhoods are shown in Table 3.

Let $\mathbf{y}$ denote the vector of dependent variables. Furthermore, $\mathbf{X}$ denotes the matrix of explanatory variables; P denotes the set of all parameters; bs denotes the estimator of βs; P_m denotes the set of types of municipality-level parameters; N_m and N_n denote sets of pairs of adjacent municipalities and neighborhoods, respectively; λ_1, λ_2, γ_1, and γ_2 are the hyper-parameters. Equations (20) and (21) represent the estimation of parameters by the fused lasso and the fused-MCP, respectively.

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \left[\frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \sum_{(m_a,m_b)\in N_m} \sum_{j\in P_m} \left| \beta_j^{m_a} - \beta_j^{m_b} \right| + \lambda_2 \sum_{(n_p,n_q)\in N_n} \left| \beta_{n_p}^N - \beta_{n_q}^N \right| \right],$$
(20)

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \left[\frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \sum_{i\in P} \rho\left(|\beta_i|; \lambda_1, \gamma_1 \right) + \sum_{(m_a,m_b)\in N_m} \sum_{j\in P_m} \rho\left(\left| \beta_j^{m_a} - \beta_j^{m_b} \right|; \lambda_2, \gamma_2 \right) \right.$$
$$\left. + \sum_{(n_p,n_q)\in N_n} \rho\left(\left| \beta_{n_p}^N - \beta_{n_q}^N \right|; \lambda_2, \gamma_2 \right) \right].$$
(21)

The numeric explanatory variables in Eqs. (20) and (21) were scaled to have zero mean and unit variance. It is intended to adjust the scale of parameters; if the scale of parameters is different (the effect of penalties is different for the parameters), then it would lead to biased estimators.

When estimating parameters, many hyperparameter settings were tested. We selected the estimation results with the minimum AIC value, assuming that the disturbances of the apartment rent model of Eq. (19) were independently, identically, and normally distributed. Since the number of apartments in the dataset (50,355), were far larger than the number of parameters in the model (651), the parameter estimation by Eqs. (20) and (21) would be almost equivalent to the estimation by

maximizing the following penalized likelihood function for certain hyperparameters: $\mu_1, \mu_2, \eta_1,$ and $\eta_2,$

$$\left(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2\right) = \arg\max_{\boldsymbol{\beta},\sigma^2} \left[l\left(\boldsymbol{\beta}, \sigma^2; \mathbf{X}, \mathbf{y}\right) - \mu_1 \|\boldsymbol{\beta}\|_1 - \mu_2 \sum_{(m_a,m_b)\in N_m} \sum_{j\in P_m} \left| \beta_j^{m_a} - \beta_j^{m_b} \right| - \mu_2 \right.$$
$$\left. \sum_{(n_p,n_q)\in N_n} \left| \beta_{n_p}^N - \beta_{n_q}^N \right| \right],$$

$$(22)$$

$$\left(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2\right) = \arg\max_{\boldsymbol{\beta},\sigma^2} \left[l\left(\boldsymbol{\beta}, \sigma^2; \mathbf{X}, \mathbf{y}\right) - \sum_{i\in P} \rho\left(|\beta_i|; \mu_1, \eta_1\right) - \sum_{(m_a,m_b)\in N_m} \sum_{j\in P_m} \rho\left(\left| \beta_j^{m_a} - \beta_j^{m_b} \right|; \mu_2, \eta_2 \right) \right.$$
$$\left. - \sum_{(n_p,n_q)\in N_n} \rho\left(\left| \beta_{n_p}^N - \beta_{n_q}^N \right|; \mu_2, \eta_2 \right) \right],$$

$$(23)$$

where l is the log-likelihood function. Then, the model selection by AIC would be fair, although the parameters are biased toward zero, especially in the estimation by the fused lasso.

3.3 Estimation results by two methods and comparison

This section introduces the estimation results by the fused lasso and the fused-MCP.

Table 4 Summary of estimated parameters by fused lasso when $\lambda_1 = 0.023$ and $\lambda_2 = 0.23$

Variable name	Number of parameters				
	Setting	Nonzero	Common	Unique	Zero
Constant term	1	1	0	1	0
Apartment age (year)	5	5	0	5	0
Area of property (m²)	5	5	0	5	0
Walking time to the nearest station (min)	5	5	0	5	0
Floor number	5	5	0	5	0
Number of rooms	5	5	0	5	0
Dummy variable of railway lines	28	18		18	10
Dummy variable of nearest railway stations	106	82		82	24
Dummy variable of neighborhoods	491	348	5	343	143
Total	651	474	5	469	177

Table 5 Estimated results of municipality-level parameters (Parameter values shown in this table are based on the estimation using the standardized explanatory variables) by fused lasso when $\lambda_1 = 0.023$ and $\lambda_2 = 0.23$

Municipalities	Variables				
	Apartment age (year)	Area of property (m^2)	Walking time to the nearest station (min)	Floor number	Number of rooms
Chiyoda	− 0.09184[a]	− 0.02874	− 0.00325	0.02303	0.00448
Chuo	− 0.09307	− 0.14606	− 0.00671	0.02155	0.01692
Minato	− 0.10565	− 0.02004	− 0.01258	0.02387	− 0.00626
Shinjuku	− 0.10097	− 0.12084	− 0.00575	0.03399	0.00672
Shibuya	− 0.09184[a]	− 0.02041	− 0.01115	0.03013	− 0.01896

[a]The two parameters were estimated to the same value; however, since the two municipalities were not contiguous, no penalty between the two was imposed

3.3.1 Estimation by fused lasso

In the estimation by the fused lasso, three settings of $\lambda_2/\lambda_1 = 0.1$, 1, and 10 were tested. The calculation was done by the R package 'genlasso' that outputs a solution path for each setting. When $\lambda_1 = 0.023$ and $\lambda_2 = 0.23$ ($\lambda_2/\lambda_1 = 10$), the model with 469 unique parameters out of 651 initially set parameters was adopted. Table 4 summarizes the numbers of parameters, and Table 5 shows the estimated values of municipality-level parameters. In this estimation, no municipality-level coefficients were estimated to be the same to the adjacent municipalities. Note that the estimated parameters of apartment age in Chiyoda and Shibuya were the same, but they are not contiguous. For reference, the adjusted coefficient of determination was 0.583 and the AIC was -64647, although it underestimated the model since the estimated parameters were biased toward zero.

3.3.2 Estimation by fused-MCP

The estimation of parameters by the fused-MCP was achieved by solving Eq. (21) by the MM algorithm, thereby, setting the estimated parameters of fused lasso as initial values.

The model has four hyperparameters: λ_1, λ_2, γ_1, and γ_2; λ_1 and λ_2 mainly act as the weights on the penalties, and their relative size effects the balance between the penalties on parameters and the penalties on the difference of adjacent parameters. Furthermore, γ_1 and γ_2 mainly act as the shape of the MCP. Since there are no positive reasons for using the penalty functions with different shapes for the parameters and the differences of adjacent parameters, this study fixed γ_1 and γ_2 to the same value γ. Then, it took the following two-step procedure to decide the hyperparameters. At the first step, by fixing λ_1 and λ_2 to the same values λ, the balance between λ and γ was searched by the grid search. At the second step, by fixing γ, the balance between λ_1 and λ_2 was tested.

Table 6 shows the results of the grid search of the hyperparameters λ and γ that minimize the AIC. After the broad grid search with the search range of λ from 0.1 to 100 and γ from 0.0001 to 0.1 was tested (Table 6a), the narrow grid search with the search range of λ from 1 to 10 at the interval of 1 and γ from 0.001 to 0.01 at the interval of 0.001 was executed (Table 6b). Consequently, the AIC was minimized when $\lambda \ (= \lambda_1 = \lambda_2) = 2$ and $\gamma \ (= \gamma_1 = \gamma_2) = 0.007$.

Table 7 shows the results of the grid search of the hyperparameters λ_1 and λ_2 by fixing $\gamma_1 = \gamma_2 = 0.007$. Both hyperparameters were tested from 1.0 to 3.0 at the interval of 0.5, and it was confirmed that the hyperparameter setting of $\lambda_1 = \lambda_2 = 2.0$ minimized the AIC. For reference, the adjusted coefficient of determination was 0.585. Hereafter, the estimated results in this setting are introduced.

Tables 8 and 9 show the summary of the estimated parameters and the estimation results of municipality-level parameters, respectively. The estimated parameters for dummy variables are shown in Figs. 3, 4 and 5. Figure 3 indicates the estimation results of railway related dummy variables (railway lines and stations), and Figs. 4 and 5 indicate the estimation results of neighborhood dummy variables.

Table 6 AIC values in grid search by fused-MCP when $\lambda = \lambda_1 = \lambda_2$ and $\gamma = \gamma_1 = \gamma_2$

(a) Results of grid search of settings $0.1 \leq \lambda \leq 100$ and $0.0001 \leq \gamma \leq 0.1$

λ	γ						
	0.0001	0.0005	0.001	0.005	0.01	0.05	0.1
0.1	− 64,709	− 64,711	− 64,711	− 64,713	− 64,723	− 64,738	− 64,741
0.5	− 64,716	− 64,726	− 64,730	− 64,761	− 64,793	− 64,884	− 64,823
1	− 64,722	− 64,732	− 64,748	− 64,809	− 64,898	− 64,751	− 64,393
5	− 64,728	− 64,775	− 64,828	− 64,840	− 64,273	− 61,854	− 61,198
10	− 64,753	− 64,828	− 64,918	− 63,895	− 62,531	− 60,258	− 60,141
50	− 64,799	− 64,157	− 62,716	− 55,001	− 54,060	− 53,688	− 53,644
100	− 64,626	− 61,821	− 58,221	− 48,892	− 48,722	− 48,591	− 48,587

(b) Results of grid search of settings $1 \leq \lambda \leq 10$ and $0.001 \leq \gamma \leq 0.01$

λ	γ									
	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009	0.01
1	− 64,748	− 64,759	− 64,776	− 64,791	− 64,809	− 64,834	− 64,850	− 64,869	− 64,890	− 64,898
2	− 64,765	− 64,798	− 64,836	− 64,880	− 64,916	− 64,928	− 64,935	− 64,926	− 64,917	− 64,923
3	− 64,783	− 64,847	− 64,907	− 64,933	− 64,925	− 64,923	− 64,904	− 64,876	− 64,846	− 64,790
4	− 64,803	− 64,896	− 64,934	− 64,918	− 64,909	− 64,861	− 64,815	− 64,714	− 64,650	− 64,597
5	− 64,828	− 64,930	− 64,913	− 64,893	− 64,840	− 64,729	− 64,630	− 64,547	− 64,449	− 64,273
6	− 64,855	− 64,925	− 64,911	− 64,834	− 64,689	− 64,591	− 64,473	− 64,351	− 64,122	− 63,768
7	− 64,881	− 64,900	− 64,849	− 64,760	− 64,564	− 64,435	− 64,225	− 63,813	− 63,643	− 63,481
8	− 64,902	− 64,894	− 64,806	− 64,610	− 64,459	− 64,196	− 63,741	− 63,550	− 63,368	− 63,074
9	− 64,922	− 64,883	− 64,756	− 64,495	− 64,296	− 63,709	− 63,488	− 63,307	− 62,949	− 62,743
10	− 64,918	− 64,830	− 64,610	− 64,411	− 63,895	− 63,465	− 63,264	− 62,884	− 62,687	− 62,531

3.3.3 Comparison of estimations by the fused lasso and the fused-MCP

It is confirmed that the estimation by the fused-MCP with the adjusted coefficient of determination of 0.585 fit the data better than that by the fused lasso with the adjusted coefficient of determination of 0.583, although the total number of unique parameters (Table 8) is smaller than the estimation by fused lasso (Table 4).

Comparing the estimated results of municipality-level parameters by the fused lasso (Table 5) and the fused-MCP (Table 9), it is confirmed that no parameters coincided with those of adjacent municipalities in the estimation by the fused lasso. However, parameters of three explanatory variables coincided in the estimation by the fused-MCP; the parameters for apartment age and walking time to the nearest station in Chiyoda and Chuo wards were estimated to be the same. Likewise, the parameters for the area of the property in Shinjuku and Shibuya wards were estimated to be the same. The results indicated that the same submarkets might be established across the borders of the municipalities.

It is worth mentioning that most of the absolute values of the estimators by the fused-MCP are larger than those by the fused lasso, especially when the absolute values of estimators are large. This result was caused by the characteristics of the MCP, which had no gradient when parameter values were larger than $\lambda\gamma$, 0.014 in this estimation. It is expected that the estimated results that exceed this value (parameters of apartment ages, area of the property, and floor number) were unbiased. However, some results for parameters of walking time to the nearest station and the number of rooms did not exceed $\lambda\gamma$; thus, the estimated parameters might be biased toward zero.

Figure 4 presents the differences in the estimation results of neighborhood dummy variables by fused-MCP and fused lasso. It is evident that the number of non-zero parameters by fused-MCP is smaller than that by fused lasso. The estimation result by fused-MCP reveals the neighborhoods where apartment rents are distinguishable from their neighbors. The number of parameters that are common to the adjacent neighborhoods by fused-MCP is larger than that by fused lasso; the estimation by fused-MCP clearly indicates the structure of market segmentation by detecting the areas where the valuation of apartments is the same. The absolute values of estimated non-zero parameters tend to be larger by fused-MCP than those by fused lasso; this result corresponds to the discussion and results of Jing et al.'s (2018) simulation study.

From the comparison of estimated results by the two methods, the estimation by the fused-MCP has higher interpretability with smaller biases. It enables the selection of areas with local differences strictly by the regularization of parameters and the differences of adjacent parameters by the MCP.

3.3.4 Comparison of results by SCAD with fused conditions

This section briefly introduces SCAD (Fan and Li 2001), one of the other regularization methods that can output the non-biased estimates, applies it to the analysis of the apartment rent dataset, and compares the estimation result with the fused-MCP.

The gradient of penalty function of SCAD, $p(|\beta|)$, is defined by

Table 7 AIC values in grid search by fused−MCP when $\gamma_1 = \gamma_2 = 0.007$

		λ_2				
		1.0	1.5	2.0	2.5	3.0
λ_1	1.0	− 64,850	− 64,892	− 64,915	− 64,906	− 64,901
	1.5	− 64,898	− 64,915	− 64,932	− 64,927	− 64,917
	2.0	− 64,914	− 64,928	− *64,935*	− 64,923	− 64,915
	2.5	− 64,911	− 64,922	− 64,919	− 64,924	− 64,921
	3.0	− 64,897	− 64,900	− 64,902	− 64,911	− 64,904

$$p'(|\beta|;\lambda,a) = \lambda \left\{ I(|\beta| \le \lambda) + \frac{(a\lambda - |\beta|)_+}{(a-1)\lambda} I(|\beta| > \lambda) \right\} \text{ for some } a > 2 \text{ and } |\beta| > 0.$$

(24)

When the parameter is zero, the gradient is λ, which is the same as that of the l_1 penalty with the weight parameter λ. When the absolute value of the parameter is larger than $a\lambda$, the gradient is zero; this means that no penalty is imposed when the parameter is large. This characteristic of the penalty function leads to the unbiased estimation for parameters that are not close to zero.

Although the SCAD is not applied to the fused conditions to the best of the author's knowledge, this section applies it to the apartment rent analysis. The estimation by SCAD is based on the MM algorithm by Hunter and Li (2005), which we utilize for the fused-MCP estimation. This study sets the hyperparameter a to 3.7 according to the discussion by Fan and Li (2001). Equation (25) represents the estimation of parameters by the SCAD,

**Table 8** Summary of estimated parameters by fused-MCP

Variable name	Number of parameters				
	Set	Non-zero	Common	Unique	Zero
Constant term	1	1	0	1	0
Apartment age (year)	5	5	1	4	0
Area of property (m^2)	5	5	1	4	0
Walking time to the nearest station (min)	5	5	1	4	0
Floor number	5	5	0	5	0
Number of rooms	5	5	0	5	0
Dummy variable of railway lines	28	14		14	14
Dummy variable of nearest railway stations	106	69		69	37
Dummy variable of neighborhoods	491	242	25	210	249
Total	651	351	28	316	300

Table 9 Estimated results of municipality-level parameters (Note that parameter values shown in this table are based on the estimation using the standardized explanatory variables) by fused-MCP

Municipalities	Explanatory variables				
	Apartment age (year)	Area of property (m²)	Walking time to the nearest station (min)	Floor number	Number of rooms
Chiyoda	− 0.09314	− 0.02830	− 0.00563	0.02272	0.00387
Chuo	− 0.09314	− 0.14613	− 0.00563	0.02223	0.01697
Minato	− 0.10574	− 0.02070	− 0.01095	0.02377	− 0.00559
Shinjuku	− 0.10097	− 0.12108	− 0.00568	0.03370	0.00658
Shibuya	− 0.09196	− 0.02070	− 0.01059	0.02991	− 0.01872
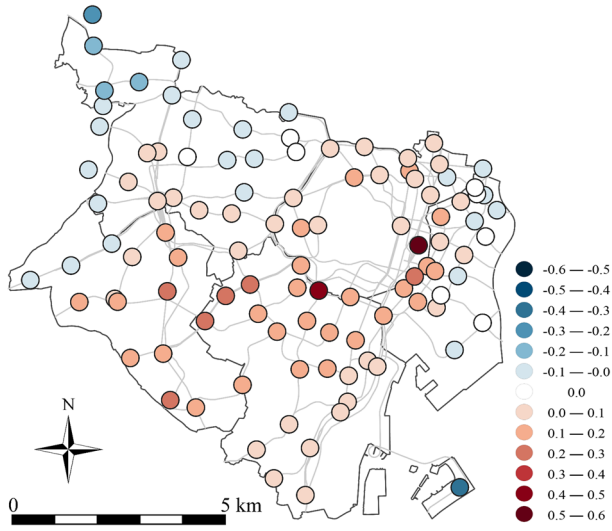
(a) Evaluations of railway stations that are the sum of
evaluations of railway lines and stations
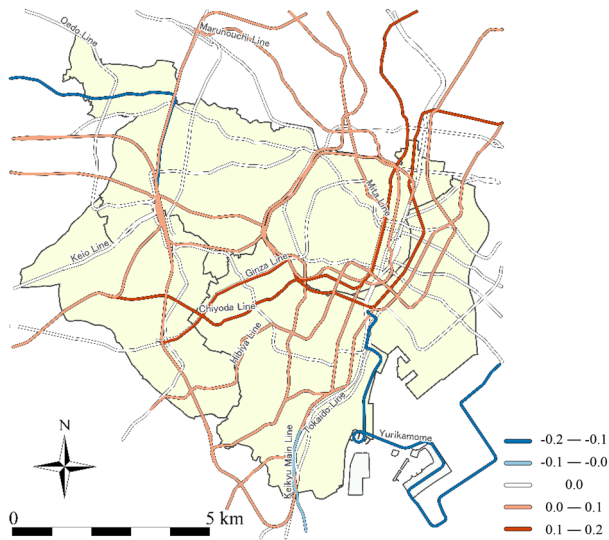
(b) Estimated parameters of railway line dummy variables

Fig. 3 Estimated parameters of railway related dummy variables by fused-MCP
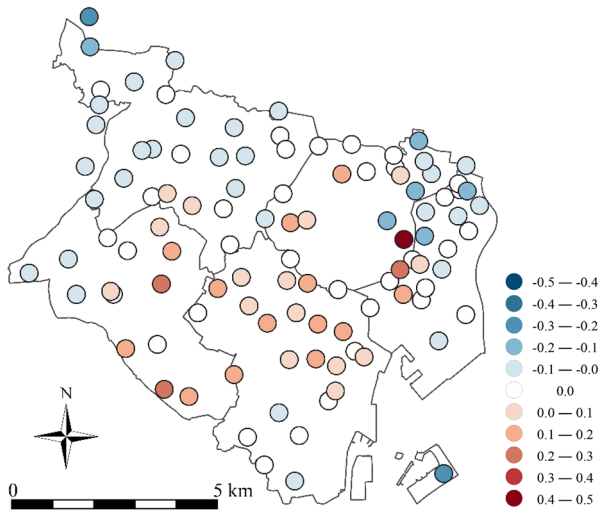
(c) Estimated parameters of railway station dummy variables

Fig. 3 (continued)

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left[\frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \sum_{i \in P} p\left(|\beta_i|; \lambda_1\right) + \sum_{(m_a, m_b) \in N_m} \sum_{j \in P_m} p\left(\left|\beta_j^{m_a} - \beta_j^{m_b}\right|; \lambda_2\right) \right.$$

$$\left. + \sum_{(n_p, n_q) \in N_n} p\left(\left|\beta_{n_p}^N - \beta_{n_q}^N\right|; \lambda_2\right) \right],$$

$$(25)$$

and the hyperparameter λ is decided by AIC, as is similar to the estimation of fused-MCP and fused lasso.

Similar to the fused lasso, the model has two hyperparameters, λ_1 and λ_2, and their relative size effects the balance between the penalties on parameters and the penalties on the difference of adjacent parameters. The balance between λ_1 and λ_2 was tested to decide the hyperparameters.

Table 10 presents the results of the grid search of the hyperparameters λ_1 and λ_2. The hyperparameter λ_1 was tested from 0.05 to 0.4 by the power of two, and λ_2 was tested from 0.0005 to 0.004 by the power of two, and it was confirmed that the hyperparameter setting of $\lambda_1 = 0.1$ and $\lambda_2 = 0.001$ minimized the AIC. For reference, the adjusted coefficient of determination was 0.587. Hereafter, the estimated results in this setting are introduced.

Tables 11 and 12 show the summary of the estimated parameters and the estimation results of municipality-level parameters by SCAD, respectively. The estimated parameters for the dummy variables are shown in Fig. 6. Figure 6 indicates the estimation results of the neighborhood dummy variables by SCAD.
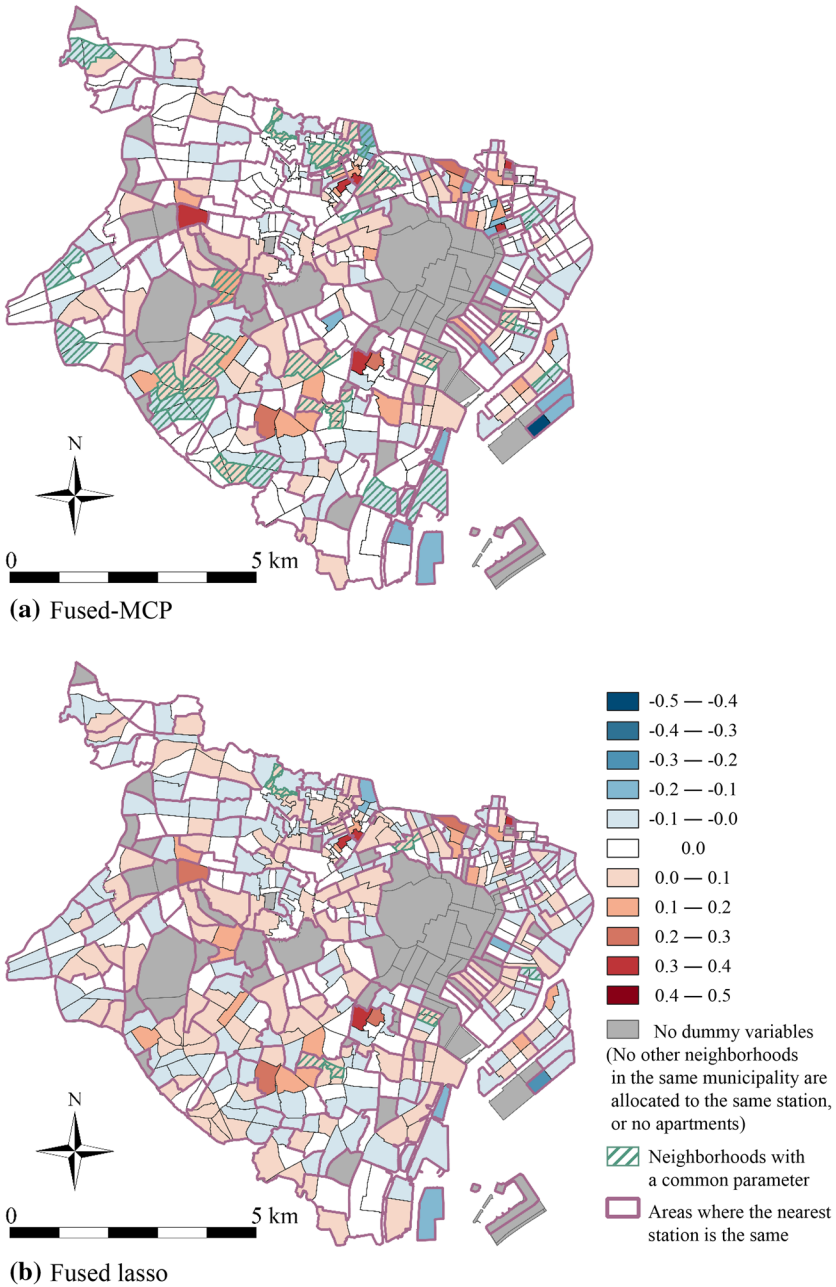
(a) Fused-MCP

(b) Fused lasso

Legend:
- -0.5 — -0.4
- -0.4 — -0.3
- -0.3 — -0.2
- -0.2 — -0.1
- -0.1 — -0.0
- 0.0
- 0.0 — 0.1
- 0.1 — 0.2
- 0.2 — 0.3
- 0.3 — 0.4
- 0.4 — 0.5
- No dummy variables (No other neighborhoods in the same municipality are allocated to the same station, or no apartments)
- Neighborhoods with a common parameter
- Areas where the nearest station is the same

Fig. 4 Estimated parameters of neighborhood dummy variables by fused-MCP and fused lasso

(a) Azabu-juban station

(b) Shibuya station
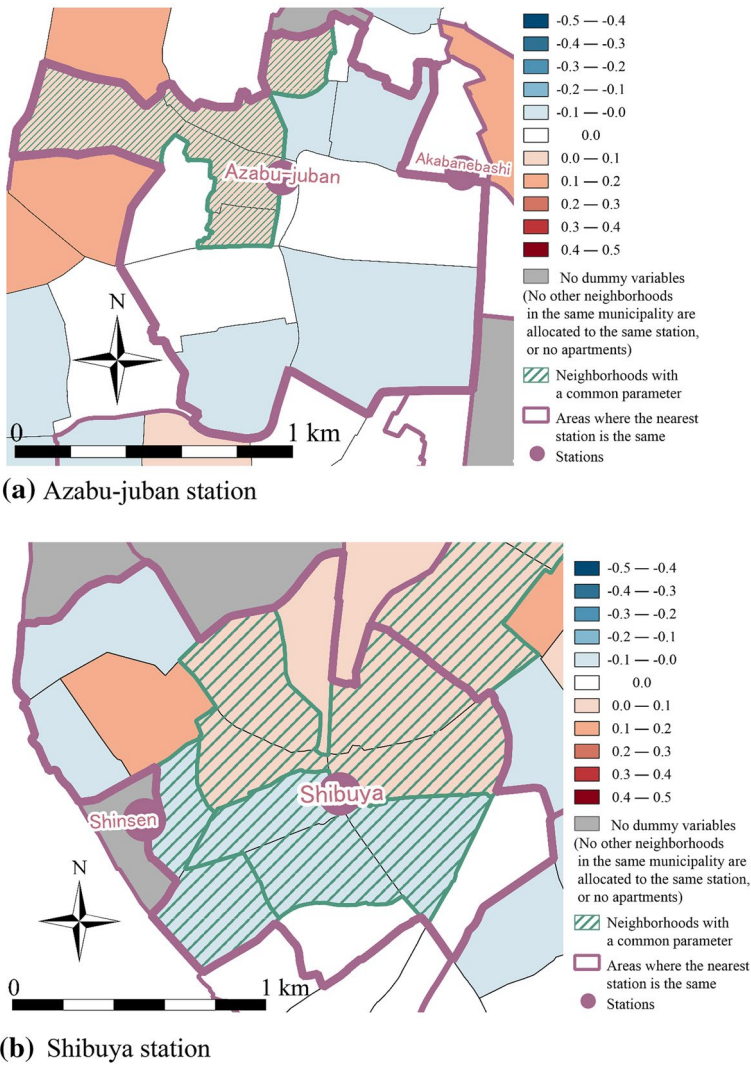
Fig. 5 Estimated parameters of neighborhood dummy variables by fused-MCP

Table 10 AIC values in grid search by SCAD with fused conditions

		λ_2			
		0.0005	0.001	0.002	0.004
λ_1	0.05	− 64,425.903	− 64,414.886	− 64,404.557	− 64,376.870
	0.1	− 64,631.448	− *64,633.822*	− 64,633.815	− 64,606.606
	0.2	− 64,633.290	− 64,625.121	− 64,629.256	− 64,607.433
	0.4	− 64,613.379	− 64,618.199	− 64,615.244	− 64,589.037

The result by the SCAD with fused conditions shows that none of adjacent parameters are estimated to be the same. The AIC by the SCAD is not improved from that by fused lasso in this estimation.

The apartment rent model in this study has many parameters and pairs of adjacency parameters that penalties are imposed on, and the minimization problem with non-convex penalties must have many local minima. Since the prevention of the problem caused by the non-convexity of the penalty function is not considered when proposing the SCAD, the possibility that the estimated result by the SCAD is one of local minima that is larger than the case by the fused-MCP. The difference in the estimated results by these two regularization methods may be what can output the non-biased estimates.

3.4 Discussions on results by fused-MCP

This section focuses on the parameters of location-related dummy variables by the fused-MCP estimation and discusses the revealed local differences in the apartment rental market in central Tokyo.

Figure 3a shows the sum of the parameters of the station and railway lines that operate at the station; it presents the evaluation of accessibility and the neighborhood environment of the stations. It is observed that the stations located near the border of Chiyoda and Chuo and in the northern area of Minato are evaluated as high. The former stations are located close to the major central business districts (CBD) of Marunouchi and Ginza, and the latter stations are located close to the other CBD of Akasaka and the urban residential area of Aoyama. However, the stations in the eastern half of Chuo and the northern area of Shinjuku are evaluated as low. The former is an area populated with small buildings and is located in an environment not conducive to live in. The latter is the area densely populated with small detached houses and is located near the Kanda river that used to overflow often; it is

Table 11 Summary of estimated parameters by SCAD with fused conditions when $\lambda_1 = 0.1$ and $\lambda_2 = 0.001$

Variable name	Number of parameters				
	Setting	Nonzero	Common	Unique	Zero
Constant term	1	1	0	1	0
Apartment age (years)	5	5	0	5	0
Area of property (m²)	5	5	0	5	0
Walking time to the nearest station (min)	5	5	0	5	0
Floor number	5	5	0	5	0
Number of rooms	5	5	0	5	10
Dummy variable of railway lines	28	18		18	24
Dummy variable of nearest railway stations	106	82		82	146
Dummy variable of neighborhoods	491	345	0	345	0
Total	650	470	0	470	180

Table 12 Estimated results of municipality-level parameters (Note that parameter values shown in this table are based on the estimation using the standardized explanatory variables) by SCAD with fused conditions when $\lambda_1 = 0.1$ and $\lambda_2 = 0.001$

Municipalities	Variables				
	Apartment age (year)	Area of property (m²)	Walking time to the nearest station (min)	Floor number	Number of rooms
Chiyoda	− 0.09217	− 0.02933	− 0.00349	0.02307	0.00478
Chuo	− 0.09265	− 0.14523	− 0.00619	0.02151	0.01680
Minato	− 0.10565	− 0.01984	− 0.01261	0.02377	− 0.00637
Shinjuku	− 0.10086	− 0.12088	− 0.00565	0.03395	0.00668
Shibuya	− 0.09179	− 0.02035	− 0.01078	0.03018	− 0.01904
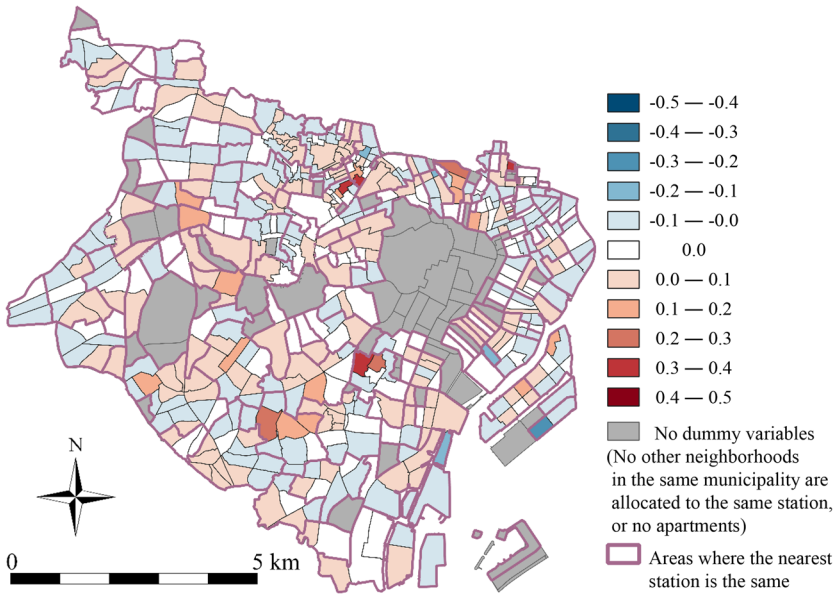
Fig. 6 Estimated parameters of neighborhood dummy variables by SCAD with fused conditions

far from the CBD and has a less conducive environment to live in within the target area. It seems that the estimated results demonstrated in Fig. 3a is quite reasonable.

Figure 3b indicates the estimated parameters of railway line dummy variables. It features the extraction of railway lines in the target area and enables the analysis of whether railway lines might affect the popularity in the apartment market. The railway lines that run through the CBD area of Chiyoda and Chuo and the northern part of Minato, Ginza, and Chiyoda subway lines are evaluated as high, and Seibu Shinjuku line, which runs the northern part of Shinjuku, and Yurikamome line, which runs harbor districts, are evaluated as low. Figure 4c is the estimated parameters of railway station dummy variables. Since this is the rest of the evaluation of railway stations without the evaluation of railway lines, it is hard to interpret.

Figure 4 indicates the estimated parameters of neighborhood dummy variables. A neighborhood is the minimum spatial unit in the analysis. The broad gray lines are the borders of the system of railway stations, and the neighborhood parameter represents the local differences in each system. Almost half of the parameters are estimated to be zero, and local differences are judged to be non-existent (Table 8). The analysis succeeds in selecting neighborhoods with substantial parameters from many among the candidates and reveals that apartment rents are locally homogeneous in most areas. However, many non-zero parameters are estimated, especially in Minato and Shibuya, the southwestern region of the target area. The results confirm that neighborhood-level local geographical segmentation occurs in these areas. It is confirmed that the fused-MCP can extract the local differences.

The shaded areas in the figure indicate the neighborhoods that have the same parameter values to the adjacent neighborhoods. By the regularization of the

difference of parameters of adjacent neighborhoods, the combinations of neighborhoods where the valuation is the same is extracted by the fused-MCP estimation. It is also confirmed that the fused-MCP is capable of extracting areas with local differences without relying excessively on the preset spatial divisions by combining adjacent spatial units.

Figure 5 is the enlarged view of Fig. 4. Figure 5a shows the area around Azabujuban station in Minato. It consists of 14 neighborhoods. Five neighborhoods in the northwestern part have the same positive value; five neighborhoods, mainly in the middle, do not have non-zero parameters; and the rest of the four neighborhoods have different negative values. The large redevelopment site of Roppingi Hills, where offices, shopping mall, hotels, and high-rise residential towers were provided, was northwest of the system of stations; the redevelopment has correspondingly affected neighboring apartment rents along the path from the station. Figure 5b shows the area around Shibuya station in Shibuya. It consists of 15 neighborhoods. Four neighborhoods on the north side of station form a group with positive parameter values, and three and two neighborhoods on the south side form two groups with negative parameter values. This figure visualizes the fact that the rental apartment market near Shibuya station is divided by north and south. The neighborhood with the highest parameter values of neighborhood dummy in this area is located the northwest of the station. It is the famous residential area of Shoto. These results confirmed that the local differences of apartment rent are clearly extracted in this analysis.

3.5 Summary

As mentioned above, the case study extracted local differences in the apartment market in central Tokyo. Although many local parameters were set in the model, the regularization of the parameters enables the detection of local areas where the pricing is different. Moreover, although the spatial division for the analysis is preset, the regularization of the differences of adjacent parameters enables the extraction of areas where the pricing is the same. In this case study, the model with hierarchical spatial divisions of municipalities, system of railway stations, and neighborhoods is estimated. The fused-MCP based analysis can set the model with flexibility. Thus, it is expected to be useful in other types of spatial analyses.

4 Conclusion

This study focused on the applicability of fused-MCP to the spatial analyses that detect local differences from spatial data. After introducing the variable selection methods (lasso and MCP), this study presented a case study to identify the geographical segmentation of the apartment rental market in central Tokyo.

In the case study, the rent model with many regional parameters to represent the difference in price formation by the municipality, the system of railway stations, and

the neighborhood is constructed. Parameter estimation was performed by the fused-MCP to extract substantial parameters and to search for common parameters in adjacent regions. It detected the hierarchical structure of the apartment rental market in which pricing is different by several levels of spatial divisions. The case study confirmed the applicability of the fused-MCP to spatial analysis that aims to search local differences.

There are several challenges to be solved to utilize fused-MCP. The most important problem is the absence of an efficient estimation algorithm. The PLUS algorithm can output a solution path of the MCP. However, it does not apply to the fused-MCP. This case study utilized the MM algorithm, but it only outputs a solution for each hyperparameter setting. Since the model used in the case study had four hyperparameters, it was impossible to set the optimum values. Another challenge is the criterion of model selection. This case study utilized AIC under the assumption that disturbances are independently and identically normally distributed, considering that the number of observations is much larger than that of parameters. However, it would be inappropriate when the number of observations is small. Information criterion, such as EBIC (Chen and Chen 2008), should be tested. From the perspective of a spatial analyst, it would be better if the more flexible setting for the adjacent parameters is applied. If the spatial units are not contiguous but are located nearby, the parameters might be equal. The question of how to set the adjacency of parameters is difficult to answer.

Despite the challenges, sparse modeling, which includes lasso and MCP, is a promising analysis method for spatial data. As is introduced earlier, the fused-lasso is being utilized in spatial analysis, especially in the analyses to detect spatial clusters of data on point events that focus on local differences. Sparse modeling is also useful in the regression analysis that focuses on local differences; however, few studies have utilized it in spatial analysis. We hope that this study would be helpful in the introduction of the methods.

Compliance with ethical standards

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & E. Caski (Eds.), *2nd international symposium on information theory* (pp. 267–281). Budapest: Akademiai Kiado. **(Reproduced in: Parzen E., Tanabe K., Kitagawa G. (eds) Selected Papers of Hirotugu Akaike. Springer Series in Statistics (Perspectives in Statistics). Springer, New York, 1998)**.

Anselin, L. (1988). *Spatial econometrics: Methods and models*. Netherlands: Springer.

Bourassa, S. C., Hamelink, F., Hoesli, M., & MacGregor, B. D. (1999). Defining housing submarkets. *Journal of Housing Economics, 8*(2), 160–183.

Bourassa, S. C., Hoesli, M., & Peng, V. S. (2003). Do housing submarkets really matter? *Journal of Housing Economics, 12*(1), 12–28.

Chen, J., & Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika, 95*(3), 759–771. https://doi.org/10.1093/biomet/asn034.

Choi, H., Song, E., Hwang, S., & Lee, W. (2018). A modified generalized lasso algorithm to detect local spatial clusters for count data. *AStA Advances in Statistical Analysis, 102*(4), 537–563. https://doi.org/10.1007/s10182-018-0318-7.

Cressie, N. (1991). *Statistics for spatial data*. New York: Wiley.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 95*(456), 1348–1360.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2003). *Geographically weighted regression*. New York: Wiley.

Goodman, A. C., & Thibodeau, T. G. (1998). Housing market segmentation. *Journal of Housing Economics, 7*(2), 121–143.

Goodman, A. C., & Thibodeau, T. G. (2003). Housing market segmentation and hedonic prediction accuracy. *Journal of Housing Economics, 12*(3), 181–201.

Goodman, A. C., & Thibodeau, T. G. (2007). The spatial proximity of metropolitan area housing submarket. *Real Estate Economics, 35*(2), 209–232.

Hunter, D. R., & Lange, K. A. (2004). Tutorial on MM algorithms. *The American Statistician, 58,* 30–37.

Hunter, D. R., & Li, R. (2005). Variable selection using MM algorithms. *The Annals of Statistics, 33*(4), 1617–1642. https://doi.org/10.1214/009053605000000200.

Inoue, R., Ishiyama, R., & Sugiura, A. (2018). Identification of geographical segmentation of the rental apartment market in the Tokyo Metropolitan Area. In *Proceedings of 10th International Conference on Geographic Information Science*, vol. 32, pp. 1–6. https://doi.org/10.4230/LIPIcs.GISCIENCE.2018.32.

Islam, K. S., & Asami, Y. (2011). Addressing structural instability in the housing market segmentation of the used houses of Tokyo, Japan. *Procedia Social and Behavioral Sciences, 21,* 33–42.

Jing, B., Yang, G., Yu, X., & Zhang, C. (2018). Fused-MCP with application to signal processing. *Journal of Computational and Graphical Statistics, 27*(4), 872–886.

Keskin, B., & Watkins, C. (2017). Defining spatial housing submarkets: Exploring the case for expert delineated boundaries. *Urban Studies, 54*(6), 1446–1462.

Kestens, Y., Thériault, M., & Des Rosiers, F. (2006). Heterogeneity in hedonic modeling of house prices: Looking at buyer's household profiles. *Journal of Geographical Systems, 8*(1), 61–96.

Kuroda, S., Morito, T., & Naraoka, K. (2014). Delineating geographical property market segmentation based on the spatial heterogeneity of price structures. *Journal of Real Estate Financial Engineering, 7,* 1–18. **(in Japanese)**.

Parker, R. J., Reich, B. J., & Eidsvik, J. (2016). A fused lasso approach to nonstationary spatial covariance estimation. *Journal of Agricultural, Biological, and Environmental Statistics, 21*(3), 569–587. https://doi.org/10.1007/s13253-016-0251-8.

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *The Journal of Political Economy, 82*(1), 34–55.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464.

Sun, Y., Wang, H. J., & Fuentes, M. (2016). Fused adaptive lasso for spatial and temporal quantile function estimation. *Technometrics, 58,* 127–137. https://doi.org/10.1080/00401706.2015.1017115.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Methodological), 58*(1), 267–288. https://doi.org/10.1111/j.1467-9868.2011.00771.x.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society, Series B (Methodological), 67*(1), 91–108. https://doi.org/10.1111/j.1467-9868.2005.00490.x.

Tibshirani, R., & Taylor, J. (2011). The solution path of the generalized lasso. *The Annals of Statistics, 39*(3), 1335–1371.

Wang, H., & Rodríguez, A. (2014). Identifying pediatric cancer clusters in Florida using loglinear models and generalized lasso penalties. *Statistics and Public Policy, 1*(1), 86–96. https://doi.org/10.1080/2330443X.2014.960120.

Watkins, C. A. (2001). The definition and identification of housing submarkets. *Environment and Planning A: Economy and Space, 33*(12), 2235–2253.

Wu, C., & Sharma, R. (2012). Housing submarket classification: The role of spatial contiguity. *Applied Geography, 32*(2), 746–756.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics, 38,* 894–942.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association, 101*(476), 1418–1429.