TR-No. 21-01, Hiroshima Statistical Research Group, 1–11

Optimizations for Categorizations of Explanatory Variables in Linear Regression via Generalized Fused Lasso

Mineaki Ohishi[1*], Kensuke Okamura[2], Yoshimichi Itoh[2] and Hirokazu Yanagihara[2,3]

[1]Education and Research Center for Artificial Intelligence and Data Innovation,
Hiroshima University, 1-1-89 Higashi-Senda-machi, Naka-ku, Hiroshima 730-0053, Japan
[2]Tokyo Kantei Co., Ltd.
Meguro Nishiguchi Bldg., 8F, 2-24-15 Kami-Osaki, Shinagawa 141-0021, Japan
[3]Graduate School of Advanced Science and Engineering, Hiroshima University
1-3-1 Kagamiyama, Higashi-Hiroshima 739-8526, Japan

Abstract

In a linear regression, a non-linear structure can be naturally considered by transforming quantitative explanatory variables to categorical variables. Moreover, smaller categories make estimation more flexible. However, a trade-off between flexibility of estimation and estimation accuracy occurs because the number of parameters increases for smaller categorizations. We propose an estimation method wherein parameters for categories with equal effects are equally estimated via generalized fused Lasso. By such a method, it can be expected that the degrees of freedom for the model decreases, flexibility of estimation and estimation accuracy are maintained, and categories of explanatory variables are optimized. We apply the proposed method to modeling of apartment rents in Tokyo's 23 wards.

(Last Modified: January 8, 2021)

Key words: Coordinate descent algorithm, Generalized fused Lasso, Linear model, Real estate data analysis

*Corresponding author
E-mail address: mineaki-ohishi@hiroshima-u.ac.jp

1. Introduction

For a given n-dimensional vector $\boldsymbol{y}$ of a response variable and a given $n \times k$ matrix $\boldsymbol{A}$

of explanatory variables, a linear regression model simply describes their relationship as follows:

$$\boldsymbol{y} = \mu \mathbf{1}_n + \boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \tag{1}$$

where μ is a location parameter, $\mathbf{1}_n$ is an n-dimensional vector of ones, $\boldsymbol{\theta}$ is a k-dimensional vector of regression coefficients, and $\boldsymbol{\varepsilon}$ is an n-dimensional vector of an error variable with mean vector $\mathbf{0}_k$ and covariance matrix $\sigma^2 \boldsymbol{I}_n$. Here, $\mathbf{0}_k$ is a k-dimensional vector of zeros. Although a linear regression model is a simple statistical model, many researchers study estimation methods for unknown model parameters μ and $\boldsymbol{\theta}$. If a response variable and each explanatory variable have a strong linear relationship, the linear regression model is appropriate. If not, the model does not fit well. To improve fitting in such a situation, we transform quantitative variables in explanatory variables to categorical variables by splitting into ranges. Through this transformation, a non-linear structure can be considered in a framework of a linear regression model. Although it is desirable to use smaller categories to allow more flexible estimation, the estimation accuracy declines as the number of parameters increases. Then, to maintain flexibility of estimation and estimation accuracy, we propose an estimation method via a generalized fused Lasso (GFL; e.g., Xin *et al.*, 2014; Ohishi *et al.*, 2019).

As the name suggests, the GFL is a generalized version of fused Lasso (FL) proposed by Tibshirani *et al.* (2005). It estimates model parameters by a penalized estimation method with the following penalty term:

$$\sum_{j=1}^{k} \sum_{\ell \in D_j} |\beta_j - \beta_\ell|,$$

where $D_j \subseteq \{1, \ldots, k\} \backslash \{j\}$ is an index set. When $D_j = \{j+1\}$ $(j = 1, \ldots, k-1)$ and $D_k = \emptyset$, the GFL coincides with the original FL. The GFL shrinks the difference of two parameters and the two parameters are equally estimated when the difference is zero. That is, the GFL reduces the degrees of freedom for the model. We apply the GFL to parameter estimation for categorical variables with 3 or more categories. The GFL can maintain flexibility of estimation and estimation accuracy. Furthermore, when several parameters are equally estimated, it means that corresponding categories are considered as the same category. Hence, we can approach the optimization of categories for categorical variables via the GFL. We describe a model and an estimation method via the GFL, and apply the method to modeling apartment rent data covering Tokyo's 23 wards.

The remainder of the paper is organized as follows. In Section 2, we describe a model and its estimation method. Section 3 shows a real data example.

2. Model & Estimation

2.1. Model

First, we rewrite model (1) by transforming quantitative variables in $\boldsymbol{A}$. Each quantitative variable is transformed to a categorical variable with 3 or more categories by splitting into small ranges. Then, we define an $n \times p$ matrix $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m)$ and an $n \times q$ matrix $\boldsymbol{Z}$. The $\boldsymbol{X}_i$ is an $n \times p_i$ matrix ($p_i \geq 3$) of a categorical variable with p_i categories that is obtained from a quantitative variable or that is originally a qualitative categorical variable, where $p = \sum_{i=1}^{m} p_i$ and each $\boldsymbol{X}_i$ includes a baseline. Note that an element of $\boldsymbol{X}_i$ takes the value of 1 or 0, and $\boldsymbol{X}_i \boldsymbol{1}_{p_i} = \boldsymbol{1}_n$ holds. The $\boldsymbol{Z}$ consists of remainder variables. Then, we consider the following linear regression model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \tag{2}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are p- and q-dimensional vectors of regression coefficients, respectively, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_m')'$, corresponding to the split of $\boldsymbol{X}$. Moreover, an intercept is not included in model (2) since each $\boldsymbol{X}_i$ includes a baseline. Note that $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m$ are vectors of regression coefficients for categorical variables with 3 or more categories. Hence, they are estimated by the GFL. Accordingly, we estimate $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ based on minimizing the following penalized residual sum of squares (PRSS):

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{\gamma}\|^2 + \lambda \sum_{i=1}^{m} \sum_{j=1}^{p_i} \sum_{\ell \in D_{i,j}} w_{i,j\ell} |\beta_{i,j} - \beta_{i,\ell}|, \tag{3}$$

where λ is a non-negative tuning parameter, $D_{i,j} \subseteq \{1, \ldots, p_i\} \backslash \{j\}$ is an index set, $w_{i,j\ell}$ is a positive penalty weight, and $\beta_{i,j}$ is the jth element of $\boldsymbol{\beta}_i$. Actually, a tuning parameter is required for each penalty term. That is, m tuning parameters are needed for the PRSS (3). However, it is complex to optimize multiple tuning parameters. Hence, we seek unification of tuning parameters by using penalty weights.

The $D_{i,j}$ is an important set to decide pairs to shrink differences with $\beta_{i,j}$, and must be appropriately defined from data. For example, when $\boldsymbol{X}_i$ is a matrix of a categorical variable obtained from a quantitative variable, the indexes $1, \ldots, p_i$ are naturally ordered, and hence $D_{i,j}$ is defined as $D_{i,j} = \{j+1\}$ ($j = 1, \ldots, p_i - 1$) and $D_{i,p_i} = \emptyset$. This means that parameters for a categorical variable obtained from a quantitative variable

are estimated by the FL. In this paper, such $D_{i,j}$ is called the FL-index. On the other hand, the penalty weight $w_{i,j\ell}$ is used based on the idea of adaptive-Lasso proposed by Zou (2006), and a general penalty weight is the inverse of the estimate corresponding to a form of the penalty term. It is reasonable to calculate the least squares estimator (LSE) $\tilde{\beta}_{i,j}$ of $\beta_{i,j}$ and to use the following weight:

$$w_{i,j\ell} = \frac{1}{|\tilde{\beta}_{i,j} - \tilde{\beta}_{i,\ell}|}. \tag{4}$$

However, since each $\boldsymbol{X}_i$ is a dummy variable matrix including a baseline, they are rank deficient and the LSEs cannot be calculated. To solve this problem, we calculate LSEs for the following model wherein baselines are removed from each $\boldsymbol{X}_i$ and an intercept is added:

$$\boldsymbol{y} = \mu \mathbf{1}_n + \sum_{i=1}^{m} \boldsymbol{X}_i^{(-)} \boldsymbol{\beta}_i^{(-)} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \tag{5}$$

where $^{(-)}$ denotes that a column vector or an element for a baseline is removed from the original matrix or vector. If the first column is a baseline, $\boldsymbol{X}_i^{(-)}$ is the $n \times (p_i - 1)$ matrix obtained by removing the first column from $\boldsymbol{X}_i$ and $\boldsymbol{\beta}_i^{(-)}$ is the $(p_i - 1)$-dimensional vector obtained by removing the first element from $\boldsymbol{\beta}_i$, i.e., $\boldsymbol{\beta}_i^{(-)} = (\beta_{i,2}, \ldots, \beta_{i,p_i})'$. By removing the baselines, we can calculate LSEs for model (5). Then, let $\tilde{\mu}$ and $\tilde{\beta}_{i,j}$ $(j = 2, \ldots, p_i)$ be LSEs of μ and $\beta_{i,j}$, and we define $\tilde{\beta}_{i,1} = \tilde{\mu}$ and calculate the penalty weight (4).

2.2. Estimation

We minimize the PRSS (3) to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ via a coordinate descent algorithm (CDA). This algorithm can obtain the optimal solution by repeating minimization along the coordinate direction. Broadly speaking, the CDA for (3) consists of two steps: a minimization step with respect to $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m$ and a minimization step with respect to $\boldsymbol{\gamma}$. It is also important how to optimize the GFL and, fortunately, there are algorithms available for this purpose (e.g., Tibshirani & Taylor, 2011; Xin *et al.*, 2014; Ohishi *et al.*, 2019). In this paper, we solve the optimization problem by using the CDA for the GFL (GFL-CDA) proposed by Ohishi *et al.* (2019). An algorithm to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ is summarized as follows:

Algorithm 1.

input*: Initial vectors for* $\boldsymbol{\beta}$ *and* $\boldsymbol{\gamma}$*, and* λ*.*

Step 1. For all $i \in \{1, \ldots, m\}$*, fix* $\boldsymbol{\beta}_j$ $(j \neq i)$ *and* $\boldsymbol{\gamma}$*, and calculate the GFL estimates of* $\boldsymbol{\beta}_i$ *via the GFL-CDA.*

Step 2. Fix $\boldsymbol{\beta}$, and calculate the LSE of $\boldsymbol{\gamma}$.

Step 3. If all parameters converge, the algorithm terminates. If not, return to Step 1.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\gamma}')' = (\theta_1, \ldots, \theta_r)'$ $(r = p + q)$. The following criterion is used for convergence judgment in the algorithm:

$$\frac{\max_{j\in\{1,\ldots,r\}}(\theta_j^{\text{new}} - \theta_j^{\text{old}})^2}{\max_{j\in\{1,\ldots,r\}}(\theta_j^{\text{old}})^2} \leq \frac{1}{100000}.$$

Since the tuning parameter adjusts the strength of the penalties, the selection of this parameter is key to obtain better estimates. Following Ohishi *et al.* (2019), we calculate $\lambda_{i,\max}$ when the estimate of $\boldsymbol{\beta}_i$ is given by $\hat{\boldsymbol{\beta}}_i = \hat{\beta}_{i,\max}\mathbf{1}_{p_i}$ and select the optimal tuning parameter in 100 points given by $\lambda_{\max}(3/4)^{j-1}$ $(j = 1, \ldots, 100)$, where $\lambda_{\max} = \max_{i\in\{1,\ldots,m\}}\{\lambda_{i,\max}\}$. By executing Algorithm 1 for each λ, the optimal tuning parameter is selected based on minimizing the EGCV criterion (Ohishi *et al.*, 2020) with the strength of the penalty being $\log n$. To calculate each $\lambda_{i,\max}$, $\hat{\beta}_{i,\max}$ satisfying $\hat{\beta}_{i,1} = \cdots = \hat{\beta}_{i,p_i} = \hat{\beta}_{i,\max}$ is required. When $\boldsymbol{\beta}_i = \beta_i\mathbf{1}_{p_i}$ $(i = 1, \ldots, m)$, model (2) is rewritten as

$$\boldsymbol{y} = \sum_{i=1}^{m}\beta_i\boldsymbol{X}_i\mathbf{1}_{p_i} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} = \sum_{i=1}^{m}\beta_i\mathbf{1}_n + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}. \tag{6}$$

Although $\hat{\beta}_{i,\max}$ is given as the LSE of β_i for this model, such a solution cannot be obtained in closed form. Then, we use a CDA to search for the solution. That is, using an algorithm like Algorithm 1, we minimize the (non-penalized) residual sum of squares for model (6).

3. Application to Modeling Apartment Rents

3.1. Data & Model

In this section, we apply the method described in the previous section to real data covering studio apartment rents in Tokyo's 23 wards. The data were collected by Tokyo Kantei Co., Ltd. between April 2014 and April 2015, and consist of rent data for n = 61,913 apartments, with 12 items for each case. Table 1 shows data items. A1 to A5 are quantitative variables and are transformed to categorical variables when modeling. B1 to B4 are dummy variables that take the value of 1 or 0, and C1 and C2 are qualitative categorical variables with 3 or more categories. Moreover, this dataset specifies the location of each apartment in terms of the 852 areas demarcated in figure 1. Figure 2 is a bar graph of monthly rents for each area, and we can find that there

Table 1. Data items

Y	Monthly apartment rent (yen)
A1	Land area of apartment (m^2)
A2	Building age (years)
A3	Top floor
A4	Room floor
A5	Walking time (min) to the nearest station
B1	Whether the apartment has a parking lot
B2	Whether the apartment is a condominium
B3	Whether the apartment is a corner apartment
B4	Whether the apartment is a fixed-term tenancy agreement
C1	Facing direction (one of the following 8 categories) N; NE; E; SE; S; SW; W; NW
C2	Building structure (one of the following 10 categories) Wooden; Light-SF; SF; RF-C; SF-RF-C; ALC; SF-PC; PC; RF-Block; other

Y and A1 to A5 are quantitative variables. B1 to B4 are dummy variables that take the value of 1 or 0. C1 and C2 are dummy variables with multiple categories. Regarding C2, SF is steel frame, RF is reinforced, C is concrete, PC is precast C.

Figure 1. The 852 areas in Tokyo's 23 wards

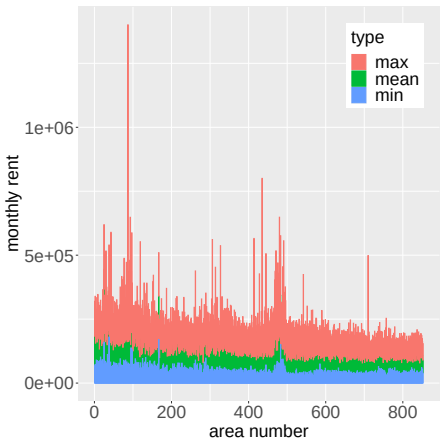


Figure 2. Apartment rents for each area

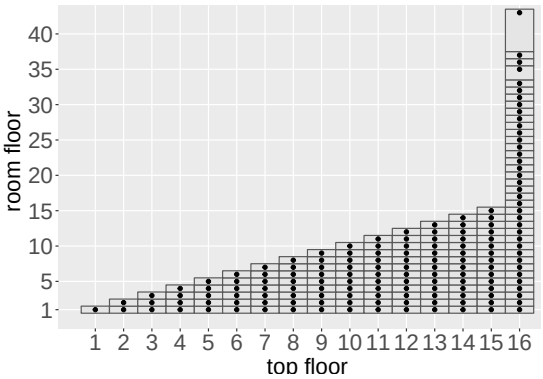
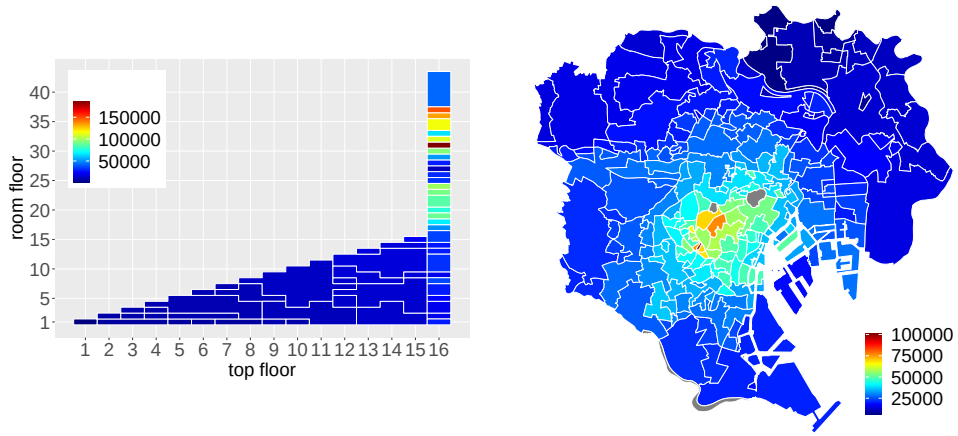


Figure 3. Floor type categories

are regional differences in the rents. In this application, let the response variable be monthly rent with the remainder set as explanatory variables. First, we describe the transformations of quantitative variables. Land area is logarithm-transformed and divided into 100 ranges by using each 1% quantile point. Note that 15% and 16% quantile points and 27% to 29% quantile points are equal. Hence, land area is transformed to a categorical variable with 97 categories. Next, since building age is a discrete quantitative variable that ranges from 0 years to 50 years, we regard it as a categorical variable with 51 categories. Similarly, we regard walking time as a categorical variable with 25 categories because the range is 1 minute to 25 minutes. Finally, top floor and room floor are dealt with as a combined variable named floor type. Top floor is a discrete quantitative variable and data are sparse beyond the 16th floor. Hence, we conflate data for 16 or more floors into the same category and regard top floor as a categorical variable with 16 categories. Room floor is also a discrete quantitative variable and 34 and 38 to 42 are missing. Hence, we regard 34 and 35, and 38 to 43 as the same categories and regard room floor as a categorical variable with 37 categories. Then, by plotting top floor and room floor on a scatter plot, we can find 157 categories in figure 3. Hence, we regard floor type (which is a combined variable of top floor and room floor) as a categorical variable with 157 categories.

The above data are formulated as follows. Let $\boldsymbol{X}_1$, $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ be the $n \times p_1$ matrix and the n-dimensional vectors, for land area, respectively, where $p_1 = 96$, $\boldsymbol{X}_1$ expresses the dummy variables for the first 96 categories, $\boldsymbol{z}_1$ expresses the dummy variable for the last category, and $\boldsymbol{z}_2$ expresses logarithm-transformed land area for the last category. That is, land area is evaluated by constants for the first 96 categories and by a linear function for the last category. Let $\boldsymbol{X}_i$ $(i = 2, \ldots, 5)$ be $n \times p_i$ matrices of dummy



(a) Floor effects (b) Regional effects

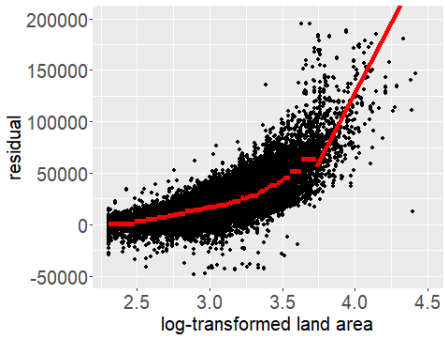
Figure 4. Estimation results for the two effects

variables for building age, walking time, facing direction, and building structure, where $p_2 = 51$, $p_3 = 25$, $p_4 = 8$, and $p_5 = 10$. Let $\boldsymbol{X}_6$ be an $n \times p_6$ matrix of dummy variables for floor type, where $p_6 = 157$. Moreover, since monthly rent depends on location, we consider regional effects according to Ohishi *et al.* (2019). Let $\boldsymbol{X}_7$ be an $n \times p_7$ matrix of dummy variables expressing the location of each apartment, where $p_7 = 852$. Note that all $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_7$ include baselines. Furthermore, let $\boldsymbol{z}_j$ $(j = 3, \ldots, 6)$ be the n-dimensional vectors of dummy variables for B1 to B4. Then, $p = 1205$, $q = 6$, and $m = 7$.

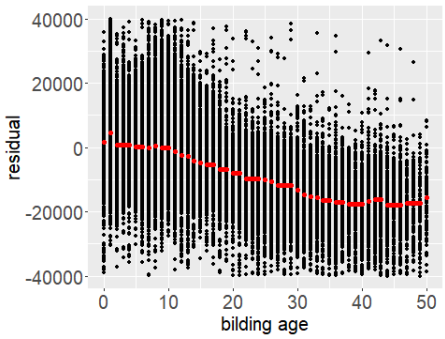
Index sets $D_{i,j}$ $(i = 1, \ldots, m)$ are defined as follows. For $i = 1, \ldots, 4$, indexes $1, \ldots, p_i$ are naturally ordered. Since land area, building age, and walking time are quantitative variables, $D_{i,j}$ $(i = 1, 2, 3)$ is defined by the FL-index. On the other hand, the facing direction has the following order: N $\to$ NE $\to$ E $\to \cdots \to$ NW $\to$ N. Hence, $D_{4,j}$ is defined by the FL-index with $D_{4,p_4} = \{1\}$. In contrast, since building structure has no order, $D_{5,j}$ is defined by $D_{5,j} = \{1, \ldots, p_5\} \backslash \{j\}$ $(j = 1, \ldots, p_5)$. This means that differences of parameters for all pairs are shrunken. Floor type and areas have adjacent relationships; see figures 1 and 3. Hence, $D_{6,j}$ and $D_{7,j}$ are defined according to the adjacent relationships.

3.2. Results

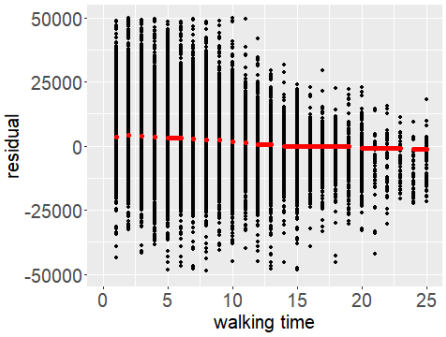
Figure 4 shows estimation results for floor effects and regional effects by choropleth



(a) Log-transformed land area

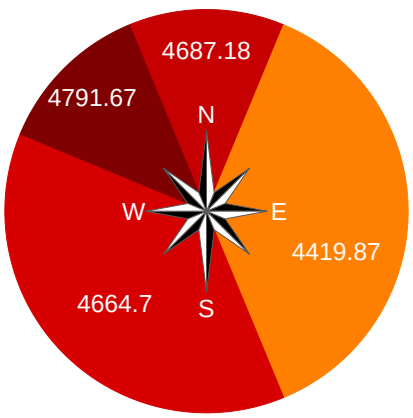


(b) Building age

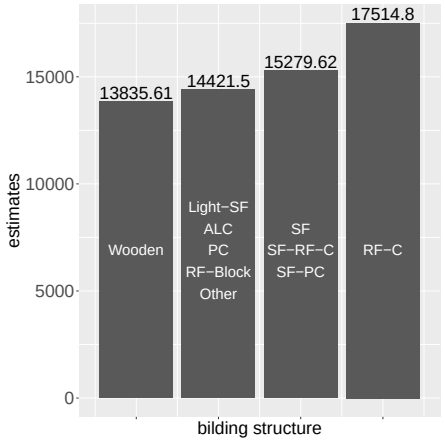


(c) Walking time

Figure 5. Residual plots for quantitative variables



(a) Facing direction



(b) Building structure

Figure 6. Estimation results for qualitative variables

Table 2. Regression coefficients for B1 to B4

B1	B2	B3	B4
25145.49	-2046.31	293.21	-786.10

Table 3. Model fit and run time

R^2	MER (%)	run time (min)
0.861	6.064	4.96

maps. Figure 4a shows that floor type and floor effect tend to increase as top floor or room floor increases. Moreover, floor types with 157 categories are clustered to 55 types. Figure 4b displays the results for regional effects and we can find that these effects are higher for central areas compared to peripheral areas. Moreover, 852 areas are clustered to 188 areas.

Figure 5 shows residual plots for quantitative variables: land area, building age, and walking time. The residual plots are unproblematic. Since there are non-linear structures for all variables, it can be considered that transforming the quantitative variables was beneficial. Land area has 97 categories and the first 96 categories are clustered to 48 categories as per figure 5a. Building age has 51 categories and they are clustered to 31 categories as per figure 5b. Walking time has 25 categories and they are clustered to 14 categories as per figure 5c.

Figure 6 shows estimation results for the qualitative variables with 3 or more categories: facing direction and building structure. Both of these variables are clustered to 4 categories; see figures 6a and 6b, respectively.

Table 2 summarizes estimates for dummy variables that take the value of 1 or 0. Finally, table 3 summarizes model fit and run time, and reveals that the results obtained are reasonable.

Acknowledgment The first author's research was partially supported by JSPS KAKENHI Grant Number JP20H04151. The last author's research was partially supported by JSPS KAKENHI Grant Numbers JP16H03606, JP18K03415, and JP20H04151.

References

Ohishi, M., Fukui, K., Okamura, K., Itoh, Y. & Yanagihara, H. (2019). Estimation for spatial effects by using the fused Lasso. Technical Report TR-No. 19-07. Hiroshima Statistical Research Group. Hiroshima.

Ohishi, M., Yanagihara, H. & Fujikoshi, Y. (2020). A fast algorithm for optimizing ridge parameters in a generalized ridge regression by minimizing a model selection

criterion. *J. Statist. Plann. Inference*, **204**, 187–205.

Tibshirani, R., Saunders, M. & Rosset, S. (2005). Sparsity and smoothness via the fused Lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **67**, 91–108.

Tibshirani, R. & Taylor, J. (2011). The solution path of the generalized Lasso. *Ann. Statist*, **39**, 1335–1371.

Xin, B., Kawahara, Y., Wang, Y. & Gao, W. (2014). Efficient generalized fused Lasso and its application to the diagnosis of Alzheimer's disease. Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence. Association for the Advancement of Artificial Intelligence. California. 2163–2169.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.*, **101**, 1418–1429.