Spatial Smoothing via a Resampling Method: Estimation with Area-based Panel Data

Atsushi Yoshida *
Osaka Prefecture University
1-1, Gakuen-cho, Sakai, Osaka 599-8531

Tatsuhiro Shichijo
Osaka Prefecture University
1-1, Gakuen-cho, Sakai, Osaka 599-8531

(Preliminary)

Abstract

When we use area-based panel data to analyze economic activities in a metropolitan region, we have to treat cluster effects because some economic activities agglomerate in a group of areas adjacent each other and thus form a cluster. We propose a resampling method, namely *leave-one-out cross-validation*, to find how many clusters are there in the region and which area belongs to which cluster. We examine the effectiveness of the method with simulation studies and compare the estimates with the within-class estimates. We also apply our method to find potential demand for houses in Tokyo Metropolitan Area.

Key words: cluster-effects model, housing start, leave-one-out cross-validation, panel data, resampling method

JEL classification:

*This research was supported by a grant-in-aid from the Zengin Foundation for Studies on Economics and Finance.

1 INTRODUCTION

In this paper, we will consider statistical issues when we use area-based panel data models in order to analyze economic activities in a metropolitan region. The metropolitan region consists of officially pre-determined areas like counties or municipalities. Most of the data available for us are based on these areas.

These official borders do not necessarily constrain economic activities of a private sector in a metropolitan region. Economic infrastructure such as railways, subways, highways, roads, canals, ports and so on, which are called the *second nature* by Krugman (1993, 1996), lays across the borders and possibly affects the activities. We often find manufacturing factories agglomerate along a canal, while software firms agglomerate around a university in the metropolitan region. Their activities are across the pre-determined official areas but are concentrated in areas adjacent each other. We will call the group of areas where some economic activities agglomerate a cluster.

Although detecting the clusters may not be difficult when the economic activities of concern are observable, it is a statistical issue to do so when they are unobservable. We will consider the case where area-based panel data are available and the clusters are represented by cluster parameters in a linear regression model, which are the same within a cluster but different between clusters. Thus the model is regarded as a panel data model that has cluster-effects as fixed effects.

If we are not concerned with the cluster-effects but concerned only with the parameters of observable explanatory variables, we can adopt an area-effects model that has area-specific-effect parameters for each area and obtain the within-class estimates of parameters of concern by applying analysis of variance (see Hsiao (1986) for details). Even in this case, the within-class estimates are possibly less efficient than the estimates of the cluster-effects model, as long as we can detect the structure of clusters among the metropolitan areas.

In section 2, we will consider a statistical method how to detect which area belongs to which cluster. From statistical viewpoint, this issue is regarded as a sort of model selection problems. We adopt a resampling method, namely *leave-one-out cross-validation*, since the method is robust to distributional assumptions and the calculation is easily implemented in a linear model. The method is introduced by Stone (1974) and Geisser (1975) and its application for broad model selection problems is discussed in Davison and Hinkley (1997), though it has not been applied to detect clusters with area-based data. The selection procedure with the method is also explained. Section 3 shows results of simulation studies how well the *leave-one-out cross-validation* works to detect the clusters. We also compare the within-class estimate with the cluster-effects estimates. We find the cluster-effects estimates are more efficient than the within-class estimates. The method is also applied to estimate a housing demand function in Tokyo Metropolitan Area. Housing demand in an area depends upon basically income per household, amenity of an area and the disutility caused by congestion. It also depends on unobservable utility-improving environmental factors that are not capitalized in the land price of an area. Potential housing demand is affected by these factors in an area, which are represented by the cluster-effects in a statistical model and are to be estimated. Section 4 concludes and discusses remaining issues.

2 STATISTICAL MODEL WITH CLUSTERS

Let us consider a model with area-based panel data. Assume that we have observations of m areas for T periods. The area-effects model is expressed as follows:

$$y_{it} = \mu_i + x_{it}\beta + v_{it}, \quad i = 1, \ldots, m; \; t = 1, \ldots, T \tag{1}$$

The $\{y_{it}\}$ and $\{x_{it}\}$, which does not include a constant term, are dependent and independent variables that represent socio-economic properties of the ith area, respectively. The β $(K \times 1)$ represents the relationship between them which is of concern for researchers. The μ_i, one of area-effects, represents unobservable socio-economic characteristics in the ith area. The v_{it} is a error term that is independent and identically distributed for all i and t.

Let us assume there are q $(q \ll m)$ clusters, which are unobservable and thus we have to decide q statistically. The area-effects, $\mu_i, i = 1, \ldots, m$, should be classified into q classes, say

$\mu_1, \dots, \mu_q$. Then the cluster-effects model is as follows:

$$y_{it} = \mu_q + x_{it}\beta + v_{it}, \quad i = 1, \dots, m; \; t = 1, \dots, T; \quad \text{if } i \in q\text{th cluster} \tag{2}$$

The vector form of eq.(2) is

$$y_t = D_0\mu_0 + X_t\beta + v_t, \quad t = 1, \dots, T. \tag{3}$$

The $\mu_0 = (\mu_1, \dots, \mu_q)$ is a parameter vector to be estimated. D_0 is an $m \times q$ matrix of dummies that indicates which area belong to which cluster. For example, if sth area and lth area belong to the same cth cluster, then the cth element of the sth and lth rows of D_0 are the same, namely 1, and the other elements of the rows are 0s.

We will consider how we can estimate the rank of D_0 (namely q), μ_0 and β, and identify the structure of D_0 based on the model eq.(3). There are two points to be considered for the estimation. Firstly, we have to find how many clusters are there, which area belongs to which cluster and to estimate the parameters of concern at the same time. Without the classification of area-effects into cluster-effects, we cannot obtain the consistent estimates of μ_0. Secondly, let A be an adjacent matrix of areas in the region, which is an $m \times m$ symmetric dummy matrix indicating which areas are neighbors of an area. For example, if the i, jth element is 1, then i, jth areas are adjacent. The diagonal elements are 1 by the definition. Since D_0 indicates the structure of clusters, it should be a transformed matrix of the adjacent matrix A by using information of which areas are combined together to a cluster.

From the statistical point of view, detecting the rank and structure of D_0 is regarded as a model selection problem. In this case, the largest model is the case where $\mu_1, \dots, \mu_m$ have different values, that is, they are not classified into fewer classes at all, namely the area-effects model. On the other hand, the smallest model is the case where $\mu_1, \dots, \mu_m$ have the same value, that is, they all are classified into one class. There are a lot of possibilities of classification between the largest and smallest models.

In a statistical model selection context, there are two major methods, one is using Kullback-Leibler-information-based selection criteria, namely AIC, BIC and SBIC (see Lütkepohl (1991), for example), and the other using a resampling-method-based selection criterion. The former criteria are easily calculated but they heavily depend upon the assumptions of distributions. On the other hand, the latter criterion needs huge computation time, though they are robust to them. The asymptotic equivalence of *cross-validation* and AIC is proved by Stone (1979).

The model-selection criterion with the resampling methods is *aggregate prediction error*. In a linear regression model, it is defined as

$$\Delta = \frac{1}{n}\sum_{i=1}^{m} E((Y_{+j} - \eta(X_j, \hat{F}))^2 | \hat{F})$$

where Y_{+j} is one of possible realizations at X_j, $\eta(X_j, \hat{F})$ being an estimate of mean response function and $\hat{F}$ is an empirical distribution of Y and X that represents data. One of the estimate of the *aggregate prediction error* is obtained by *leave-one-out cross-validation*, which is defined as

$$\hat{\Delta}_{CV} = \frac{1}{n}\sum_{i=1}^{m}(y_j - \eta(x_j, \hat{F}_{-j}))^2$$

where $\hat{F}_{-j}$ represents the $n-1$ observations $\{(x_k, y_k), k \neq j\}$. In a linear regression model, we have $\eta(x_j, \hat{F}_{-j}) = x_j\hat{\beta}_{-j}$ where $\hat{\beta}_{-j}$ is the estimate using just the data of Y and X excluding the jth sample. To select a model among possible combination of explanatory variables, we calculate $\hat{\Delta}_{CV}$ for all combinations and select the combination which attains the minimum value in principle. It is, however, almost impossible because there are too many possibilities to try.

In general, forward, backward, or stepwise methods are often used for selecting combinations of variables when trying all combinations is impossible. In our model-selection problem, we select the backward method, that is, starting with the largest model, we combine two adjacent areas for all possible cases, selecting the combination that attains the smallest *APE* and regard a newly integrated area as a cluster. Then the number of the areas decreases by one in every step of the procedure. By repeating this process, we can find the clusters where the *APE* is the smallest.

Let us explain the procedure stated above more precisely. The matrix form of eq.(3) is

$$y = (\mathbf{1}_T \otimes D_0)\mu_0 + X\beta + v, \tag{4}$$

where $y = (y_1', \ldots, y_T')'$, $X = (X_1', \ldots, X_T')'$ and $v = (v_1', \ldots, v_T')'$. The purpose is to find the structure of clusters which is expressed in D_0 and estimate μ_0 and β. Since we use the backward method to combine the adjacent areas, we need an $m \times m$ adjacent matrix, $A(m)$, as an initial areal condition, while an initial matrix for D is an $m \times m$ identity matrix. First, we calculate *APE*s for all possible combinations of two areas adjacent each other. For example, assuming the kth area and lth area are adjacent, we calculate *APE* with an $m \times (m-1)$ matrix $D(m-1; k=l)$, which created by integrating the kth and lth column vectors. Among *APE*s for all possible combinations, we can select the minimum-attained combination. We define the value of *APE(m-1)* as $APE^*(m-1)$ and new $(m-1) \times (m-1)$ adjacent matrix as $A(m-1)$. In the next step, we use the adjacent matrix $A(m-1)$ for searching possible combinations of areas. Let us redefine y_j and x_j as the jth element and row vector of y and X for all $j = 1, \dots, mT$, respectively. Then *APE(k)* is calculated as

$$APE(k) = \frac{1}{mT}\sum_{j=1}^{mT}(y_j - d_j(k)\hat{\mu}_{-j} - x_j\hat{\beta}_{-j})^2$$

where $d_j(k)$ is the jth column of $\mathbf{1} \otimes D(k)$, $\hat{\mu}_{-j}$ and $\hat{\beta}_{-j}$ is the estimates using the data excluding $y_j, d_j(k)$ and x_j. The optimal *APE(k)* is

$$APE^*(k) = \min APE(k)$$

Repeating this procedure from $k = m$ to $k = 2$ and selecting k^* as $k^* = \min_k \mathrm{APE}^*(k)$. Then k^* is the optimal rank of D_0 and its corresponding $D(k^*)$ is the structure of clusters.

3 SIMULATIONS AND EMPIRICAL EXAPMPLE

In this section, we will examine if the method proposed in the previous section would work well and apply it to analyze municipality-based data of housing start in Tokyo Metropolitan Area. In the simulations, we will compare the estimates of cluster-effects model with that of the area-effects model, namely within-class estimates. Note that we are not able to find the structure of clusters or obtain consistent estimates of μ with the area-effects model. The within-class estimate of β is defined as follows:

$$\hat{\beta}_W = (X'QX)^{-1}X'Qy$$

where $Q = I_{mT} - Z(Z'Z)^{-1}Z'$, $Z = I_m \otimes \mathbf{1}_T$.

3.1 SIMULATIONS

In the simulations, we generate the necessary data based on eq.(2) for 3 years (T=3). We use a 6×6 lattice for a total region to be examined, where there are 36 areas ($m = 36$ in eq.(2)). We have to define which are neighbors of an area at first. We assume that left, right, upper and lower adjacent areas of an area are its neighbors. Note that there are just 2 neighbors for 4 corner-areas of this region and 3 neighbors for edge areas. We make sequential numbers for the areas in an order so that the i, jth cell of the lattice should be $6 \times (i-1) + j$ (see Fig. 1). We set three clusters, the upper-left cluster consisting of 9 areas (1,2,3,7,8,9,13,14,15), the upper-right cluster consisting 9 areas (4,5,6,10,11,12,16,17,18), and the rest consisting of 18 areas. The cluster effects are set as $\mu = (2, 5, 10)$ for upper-left, upper-right and the rest clusters, respectively.

The explanatory variable and the errors , x_{it} and v_{it}, are independently drawn from $N(3, 9)$ and $N(0, 4)$, respectively. The parameter β is set to be 2. We conducted 1000-times simulations.

In these simulations, we also consider the other case, where the cluster-effects are random and spatially correlated. The model is specified as

$$y_{it} = x_{it}\beta + u_i + v_{it}, \quad i = 1, \ldots, m; \; t = 1, \ldots, T,$$

where u_i represents random cluster-effects. We specify the conditional density function of the ith variable, u_i, as follows:

$$f(u_i | \{u_j, j \in \mathcal{N}_i\}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-(u_i - m(u_j; j \in \mathcal{N}_i))^2/\sigma^2\right], \tag{5}$$

where $\mathcal{N}_i$ is a set of neighbors of the ith area and the conditional mean is defined as

$$m(u_j; j \in \mathcal{N}_i) = \mu_i + \sum_{j \in \mathcal{N}_i} \lambda c_{ij}(u_j - \mu_j)$$

where $c_{ij} = c_{ji}$, $c_{ii} = 0$ and $c_{ik} = 1$ if there is pairwise dependence between area i and area k, otherwise it is 0. The joint distribution of $u = \{u_1, \ldots, u_m\}$ is obtained as follows:

$$u \sim N(\mu, (I - \lambda C)^{-1} M), \tag{6}$$

where $\mu = (\mu_1, \ldots, \mu_m)$, C a $m \times m$ matrix with its i, jth element being c_{ij} and M is a $m \times m$ diagonal matrix with its ith diagonal element being σ^2. The detailed explanation of its properties is discussed in Cressie (1993). The random cluster-effects, u_i, $i = 1, \ldots, m$, are generated from normal distribution of eq.(6), where $\mu = 0$, $\lambda = 1/4$, $M = 3 \times I_m$ and the adjacent matrix is defined above.

Firstly, we evaluate how correctly we can select the number of clusters, the structure of clusters and how efficiently we can estimate the values of μ with the method. For the first point, we examine the distribution of the selected number of clusters in the simulations. We also examine the expectation how many estimated clusters lay across the true clusters. For the second point, we evaluate the efficiency by the mean squared error for each 36 areas, comparing three mean squared errors of the estimates, namely estimates with true clusters, those with selected clusters and the within-class estimates.

In table 1-1 and 1-2, we can see descriptive statistics of the distribution of the estimated number of clusters and the number of estimated clusters laying across the true clusters for the cases in fixed- and random-effects models. The mean and median of estimated number of clusters obtained with 1000 times simulations are 9.58 and 9, respectively, in the fixed-effects model and 14.16 and 14, respectively, in the random-effects model. Its standard deviation is 1.80 and about 80 % of the estimates is in the region from 7 to 12 in the former case. With this simulation, we can see the method tend to select larger number of clusters. Even if the estimated number of the clusters is larger than the true clusters, it does not cause bias of the estimates of the parameters as long as the column vectors of true cluster matrix, D_0 of eq.(3) are expressed as linear combinations of the estimated cluster matrix $\hat{D}$, though it affects the efficiency of them. The expectation of the number of estimated clusters that lay across the true clusters is less than 0.5, its median being 0 and the 90 percentile is 1 in the fixed-effects model. Thus the probability of the estimated clusters lying across the true clusters is extremely small. On the other hand, in the random-effects model, the expectation is more than 1 and the median is 1 so that there is a little possibility that the estimates of parameters are biased.

In table 2-1 and 2-2, we can see means of $\hat{\mu}$ and mean squared errors (MSEs) of the OLS estimates in the estimated cluster model, true cluster model and non-clustered model, namely the within-class estimates. In both cases of fixed- and random-effects models, the estimates of the cluster-effects in each area are almost unbiased. In the fixed-effects model, the mean squared error of the clustered model is uniformly smaller than the non-clustered model, though they are larger than the true model. Note that the mean squared error of the clustered model consists of three parts, that is, the squared bias, variance of the estimate and the bias caused by misclustering. The third factor of the MSE is negligibly small from table 1-1. Even in the random-effects model, the MSEs of the estimates in clustered model are uniformly smaller than the non-clustered model, though the values are not so different from the non-clustered model.

Secondly, we compare the estimates of β in the estimated clustering structure with the within-class estimate. In table 3, we can see means, standard deviations and MSEs in OLS estimates with estimated clustering structure, in the within-class estimate and in OLS estimates with true clustering structure. In the case of fixed-effects model, both the standard deviation and MSE of the estimates of the clustered model are superior to the within-class estimate, though OLS estimates of the true cluster model is the most efficient among these estimates. In the random-effects model, since the within-class estimate is obtained by eliminating the effects, it is the most efficient estimate among them. Even in this case, the estimate of the clustered model is nearly as efficient as the within-class estimate. The OLS estimate with true clustering structure is the worst.

From the results of the simulations, we are able to conclude as follows: Firstly, the *leave-one-out cross-validation* tends to select larger model than the true model but the estimated clusters seldom lay across the true clusters so that the estimates of the cluster-effects are almost unbiased and also efficient than the within-class estimates. Secondly, in the case where the cluster-effects are fixed, the parameter-estimates of the explanatory variables are more efficient than the within-class estimates. Even in the cluster-effects being random, they are nearly as efficient as the within-class estimates. Thus, the estimates proposed in this paper are more preferable than the within-class estimates when the cluster-effects exist.

3.2 EMPIRICAL EXAMPLE

In this subsection, we apply our method to examine the determinants of the number of housing start per household of the municipalities in Tokyo Metropolitan Area (TMA) from 1996 to 1998, which is defined as a collection of areas that locate within 60-minute-distant from the Tokyo station. There are 87 municipalities in the region.

The explanatory variables are logarithm of income per household, logarithm of average price of residential land and logarithm of population density in an area. The explained variable is also taken logarithm. We expect housing start in an area with higher income per household will be larger than other areas. The average price of residential land in an area represents amenity of the area since amenity is capitalized in the land price. Thus in the area with higher land price demand for houses is larger than the other areas. The population density represents disutility caused by congestion. These three variables are the kernel of the determinants of housing start. At the same time, we are concerned with unobservable factor that affects housing start except for the kernel. The unobservable factor represents potential demand for houses in an area where, for example, housing stock per household is below the standard. Or it may be resistant or improving factor to build new houses by some official restrictions or area-development policies.

In table 4, we can see the estimates with clustered and non-clustered models. The within-class estimates are somewhat different from the clustered model. In both models, the income factor is insignificant but land price and population density are significant, though the values of the estimates are different. The coefficient of the land price is positive because it represents amenity of an area. Population density affects the housing start negatively because of the disutility of congestion. In figure 2, we can see which areas belong to the same cluster. Clusters are found along railways and river. East areas of TMA along Sumida River; Southern areas of Tokyo and western areas of Yokohama along Odakyu line; Western areas of Tokyo along Seibu-Shinjuku line, areas in Chiba along JR Sobu-line and so on. In figure 3, we can see the potential demand for houses represented by the cluster-effects. Potential strong demand for houses is found in northern part of Tokyo-23-districts and a southern area of Yokohama. On the other hand, the potential demands in the center and the border areas of TMA are weak.

Let us compare figure 3 and 4 that is a crude map of logarithm of housing start per household. These two maps give us different impression. In figure 4, the border areas and center of TMA have strong demand for houses per household. After adjusting the data with income, amenity and disutility by congestion, the potential demand is found in the areas where the demand seems to be weak in figure 3.

4 DISCUSSION

We propose a method of deciding how many clusters are there and which area belongs to which cluster and show by simulations that it works well and the estimated parameters of concern are more efficient than the within-class estimates. We also apply our method to examine what are the determinants of the number of housing start in Tokyo metropolitan area and spatial distribution of the unobservable potential demand for houses.

We adopt an *aggregate prediction error* as a model-selection criterion, which is estimated by a resampling method, namely *leave-one-out cross-validation*. It is possible to estimate the criterion with other resampling methods, say bootstrap or a hybrid type of them, *leave-one out bootstrap*, that may be work better than *leave-one-out cross-validation*.

The cluster-detecting procedure proposed in this paper does not search for all possibilities of clusters, because the calculation cost is huge. There may be, however, another efficient procedure to find the optimum among the possibilities.

What if we cannot use a panel data set? One possible solution is to assume μ is unknown function of location, which is often called an intensity function, and estimate it with a nonparametric method. Though the method in this paper is regarded as spatial smoothing by decreasing parameters related to clusters, the method employing the intensity function is regarded as a nonparametric smoothing method.

References

[1] Cressie, N. A. C. (1993): *Statistics for Spatial Data*, New York: John Wiley and Sons.

[2] Davison, A. C., and D. V. Hinkley (1997): *Bootstrap Methods and Their Application*, Cambridge: Cambridge University Press.

[3] Geisser, S. (1975): "The Predictive Sample Reuse Method with Application", *Journal of the American Statistical Association*, 70, 320-328.

[4] Hsiao, C. (1986): *Analysis of Panel Data* , Cambridge: Cambridge University Press.

[5] Krugman, P. R. (1993):"First Nature, Second Nature, and Metropolitan Location," *Journal of Regional Science*, 33, 293-298.

[6] Krugman, P. R. (1996):*The Self-organizing Economy*, Cambridge, MA: Blackwell.

[7] Lütkepohl, H. (1991): *Introduction to Multiple Time Series Analysis*, 2nd edition, Berlin: Springer-verlag. "Cross-validatory Choice and Assessment of Statistical Predictions," *Journal Of the Royal Statistical Society*, Ser. B, 36, 111-147.

[8] Stone, M (1974): "Cross-validatory Choice and Assessment of Statistical Predictions," *Journal Of the Royal Statistical Society*, Ser. B, 36, 111-147.

[9] Stone, M (1977): "An Asymptotic Equivalence of Choice of Model by Cross-validation and Akaike's Criterion," *Journal Of the Royal Statistical Society*, Ser. B, 39, 44-47.

Table 1-1: Descriptive Statistics of Selecting of Clusters in Fixed-effects Model

	Number of Clusters	Number of Estimated Clusters Laying across True Clusters
Mean	9.58	0.46
Standard Deviation	1.80	0.58
5 percentile	7	0
10 percentile	7	0
Median	9	0
90 percentile	12	1
95 percentile	13	1

Table 1-2: Destrictive Statistics of Selecting of Clusters in Radom-effects Model

	Number of Clusters	Number of Estimated Clusters Laying across True Clusters
Mean	14.16	1.047
Standard Deviation	2.50	0.81
5 percentile	10	0
10 percentile	11	0
Median	14	1
90 percentile	17	2
95 percentile	18	2

Table 2-1: Estimates of the Cluster Effects and Mean Squared Errors in Fixed-effects Model

Cluster	Area	Mean	mean squared error in clustered model	mean squared error in true model	mean squared error in non-clustered model
UPPER-LEFT CLUSTER	1	2.01	1.00	0.16	1.29
	2	2.00	1.00	0.16	1.36
	3	2.06	1.33	0.16	1.41
	7	2.02	0.96	0.16	1.34
	8	2.05	1.07	0.16	1.36
	9	2.07	1.26	0.16	1.41
	13	1.99	1.02	0.16	1.34
	14	1.99	1.04	0.16	1.38
	15	2.10	1.25	0.16	1.33
UPPER-RIGHT CLUSTER	4	4.92	1.32	0.23	1.36
	5	4.99	1.16	0.23	1.57
	6	5.02	1.00	0.23	1.32
	10	4.92	1.26	0.23	1.45
	11	5.00	1.12	0.23	1.49
	12	5.03	0.99	0.23	1.36
	16	4.91	1.45	0.23	1.48
	17	5.05	1.43	0.23	1.71
	18	5.00	1.12	0.23	1.32

Cluster	Area	Mean	mean squared error in clustered model	mean squared error in true model	mean squared error in non-clustered model
LOWER CLUSTER	19	10.02	1.02	0.10	1.32
	20	10.04	1.06	0.10	1.48
	21	10.01	1.08	0.10	1.43
	22	9.96	1.09	0.10	1.33
	23	9.99	1.05	0.10	1.38
	24	10.02	1.09	0.10	1.33
	25	10.05	1.00	0.10	1.36
	26	9.96	1.02	0.10	1.44
	27	10.02	0.99	0.10	1.34
	28	10.04	0.93	0.10	1.31
	29	10.04	0.90	0.10	1.26
	30	10.00	0.97	0.10	1.32
	31	9.97	1.13	0.10	1.38
	32	10.02	1.02	0.10	1.37
	33	10.00	1.03	0.10	1.41
	34	10.05	1.02	0.10	1.35
	35	10.03	1.11	0.10	1.55
	36	10.00	1.01	0.10	1.32

Table 2-2: Estimates of the Cluster Effects and Mean Squared Errors in Random-effects Model

Cluster	Area	Mean	mean squared error in clustered model	mean squared error in true model	mean squared error in non-clustered model	Cluster	Area	Mean	mean squared error in clustered model	mean squared error in true model	mean squared error in non-clustered model
UPPER-LEFT CLUSTER	1	2.05	4.61	1.59	4.92	LOWER CLUSTER	19	10.03	5.31	1.20	5.44
	2	2.03	4.71	1.59	5.22		20	9.99	6.14	1.20	6.48
	3	2.08	5.19	1.59	5.40		21	9.93	6.58	1.20	6.92
	7	1.97	4.81	1.59	5.31		22	9.72	7.15	1.20	7.37
	8	1.92	5.48	1.59	5.90		23	9.75	6.34	1.20	6.63
	9	2.00	6.10	1.59	6.47		24	9.79	5.34	1.20	5.37
	13	2.04	5.40	1.59	5.58		25	10.02	4.83	1.20	5.25
	14	1.96	6.26	1.59	6.55		26	9.97	5.54	1.20	5.97
	15	2.01	6.45	1.59	6.70		27	9.95	6.30	1.20	6.91
UPPER-RIGHT CLUSTER	4	4.88	5.65	1.72	5.76		28	10.00	6.35	1.20	6.88
	5	5.05	5.19	1.72	5.50		29	9.91	6.19	1.20	6.70
	6	5.00	4.69	1.72	4.97		30	9.88	5.30	1.20	5.67
	10	4.97	5.79	1.72	5.91		31	9.98	4.75	1.20	5.05
	11	4.96	5.22	1.72	5.69		32	9.88	5.13	1.20	5.43
	12	5.06	5.26	1.72	5.65		33	9.97	5.26	1.20	5.66
	16	4.92	6.91	1.72	7.07		34	10.03	5.64	1.20	5.98
	17	4.87	6.44	1.72	6.80		35	9.88	5.25	1.20	5.69
	18	4.91	5.87	1.72	6.07		36	9.98	4.44	1.20	4.81

Table 3: Comaprion of the OLS Estiamte with the Within-class Estimate

	OLS estimate with estimated clustering structure	Within-class estiamte	OLS estimate with true clustering structure
		Fixed-effects Model	
mean	1.997	1.997	1.996
standard deviation	0.076	0.079	0.067
mean squared error	0.0058	0.0062	0.0045
		Random-effects Model	
mean	1.998	2.000	2.002
standard deviation	0.076	0.075	0.084
mean squared error	0.0058	0.0056	0.0071

Table 4: Estimates of the Parameters with Estimated Clustering Structure and Within-class Estimate

	The Number of Areas within the cluster	Estimate with estimated clustering structure	t-value		The Number of Areas within the cluster	Estimate with estimated clustering structure	t-value		The Number of Areas within the cluster	Estimate with estimated clustering structure	t-value		The Number of Areas within the cluster	Estimate with estimated clustering structure	t-value
Cluster 1	5	−8.07	−6.28	Cluster 16	2	−8.46	−6.62	Cluster 31	1	−9.00	−7.20	Cluster 46	1	−8.12	−6.26
Cluster 2	2	−7.39	−5.34	Cluster 17	2	−8.22	−6.40	Cluster 32	1	−9.89	−7.95	Cluster 47	1	−7.88	−5.96
Cluster 3	2	−8.17	−6.38	Cluster 18	1	−10.70	−8.67	Cluster 33	1	−8.60	−6.86	Cluster 48	1	−8.71	−6.78
Cluster 4	2	−7.33	−5.68	Cluster 19	1	−7.68	−6.05	Cluster 34	1	−8.35	−6.37	Cluster 49	1	−8.28	−6.39
Cluster 5	4	−7.28	−5.56	Cluster 20	1	−7.12	−5.47	Cluster 35	1	−12.04	−8.10	Cluster 50	1	−9.25	−7.03
Cluster 6	5	−7.66	−5.79	Cluster 21	1	−7.45	−5.88	Cluster 36	1	−7.11	−5.14	Cluster 51	1	−6.20	−4.64
Cluster 7	2	−8.73	−7.01	Cluster 22	1	−8.80	−6.98	Cluster 37	1	−7.86	−5.81	Cluster 52	1	−7.89	−6.10
Cluster 8	4	−7.17	−5.36	Cluster 23	1	−6.68	−5.04	Cluster 38	1	−6.89	−4.93	Cluster 53	1	−6.96	−5.34
Cluster 9	3	−7.31	−5.49	Cluster 24	1	−8.83	−6.86	Cluster 39	1	−8.06	−5.94	Cluster 54	1	−8.43	−6.62
Cluster 10	2	−7.64	−5.84	Cluster 25	1	−9.99	−8.04	Cluster 40	1	−6.28	−4.56	Cluster 55	1	−9.59	−7.29
Cluster 11	2	−8.47	−6.62	Cluster 26	1	−9.50	−7.61	Cluster 41	1	−6.88	−5.06	Cluster 56	1	−7.52	−5.79
Cluster 12	3	−6.91	−5.05	Cluster 27	1	−10.64	−8.53	Cluster 42	1	−6.47	−4.68	Cluster 57	1	−8.52	−6.61
Cluster 13	2	−6.88	−5.16	Cluster 28	1	−11.56	−9.02	Cluster 43	1	−6.73	−4.97	Cluster 58	1	−8.67	−6.81
Cluster 14	2	−8.77	−6.36	Cluster 29	1	−7.78	−6.13	Cluster 44	1	−7.00	−5.21				
Cluster 15	2	−6.45	−4.81	Cluster 30	1	−7.98	−6.29	Cluster 45	1	−8.34	−6.46				

Clustered Model		t-value
Income per Household	-0.302	-0.868
Land Price	1.384	8.257
Population Density	-2.888	-8.990
Adjusted R square	0.644	

Within-Class Estimate		t-value
Income per Household	0.797	0.685
Land Price	1.545	5.040
Population Density	-4.942	-3.904
Adjusted R square	0.654	

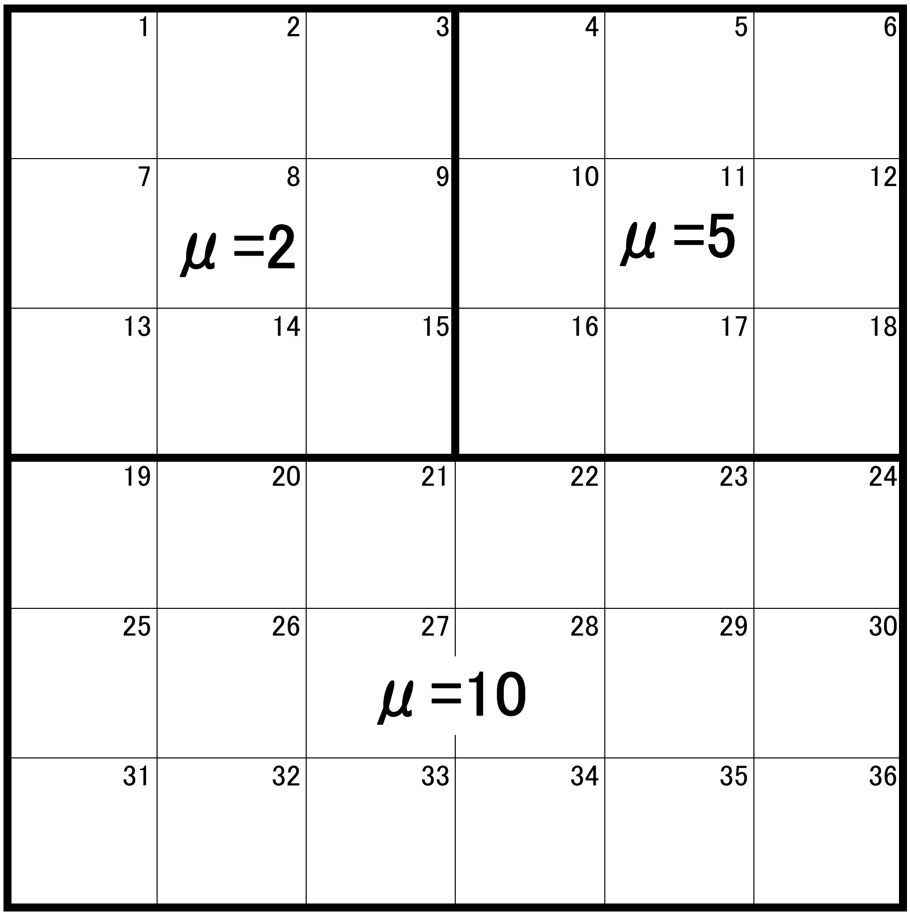

Fig.1 True Structure of Clusters

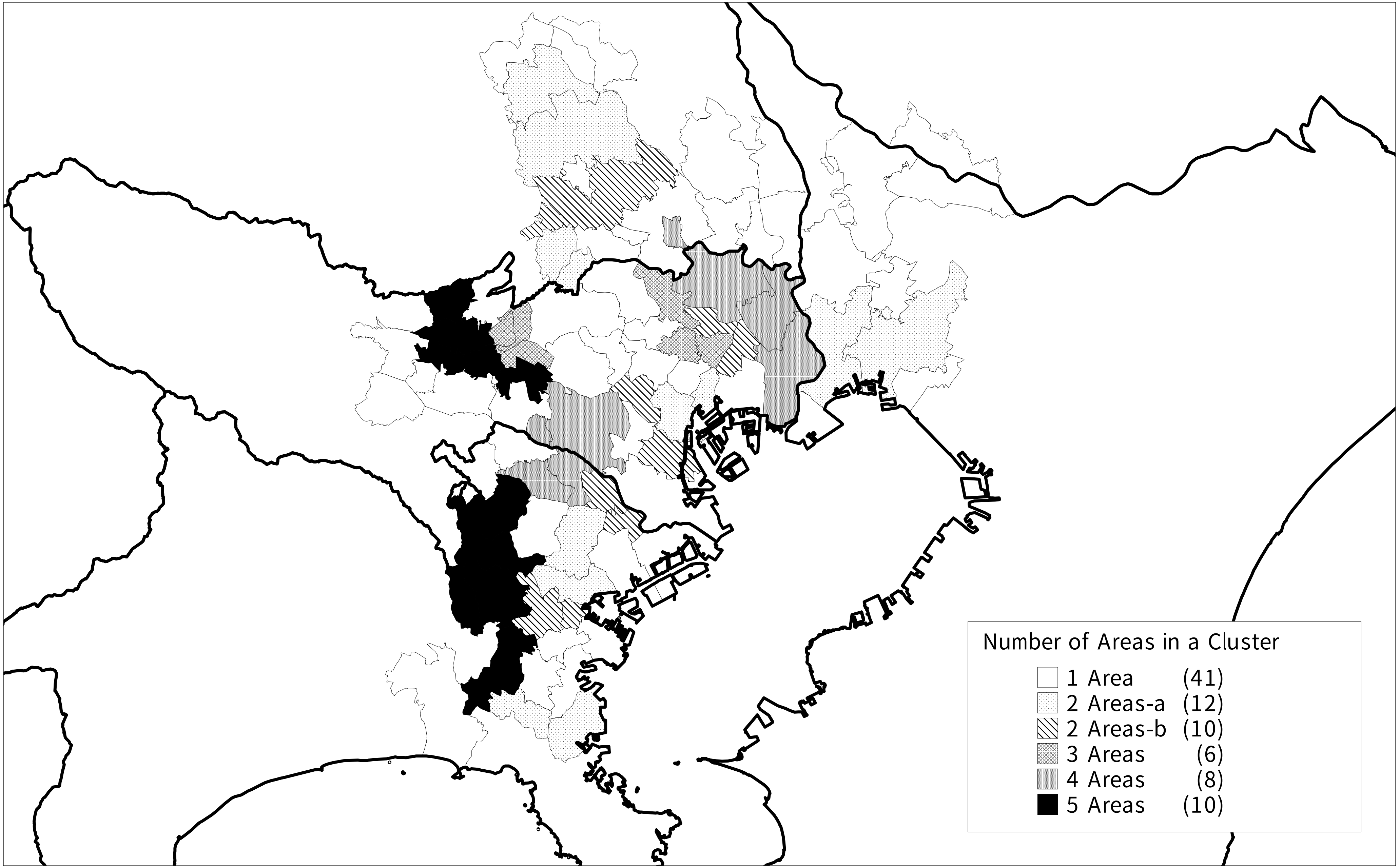

Fig.2 Clusters in the Number of Housing Start

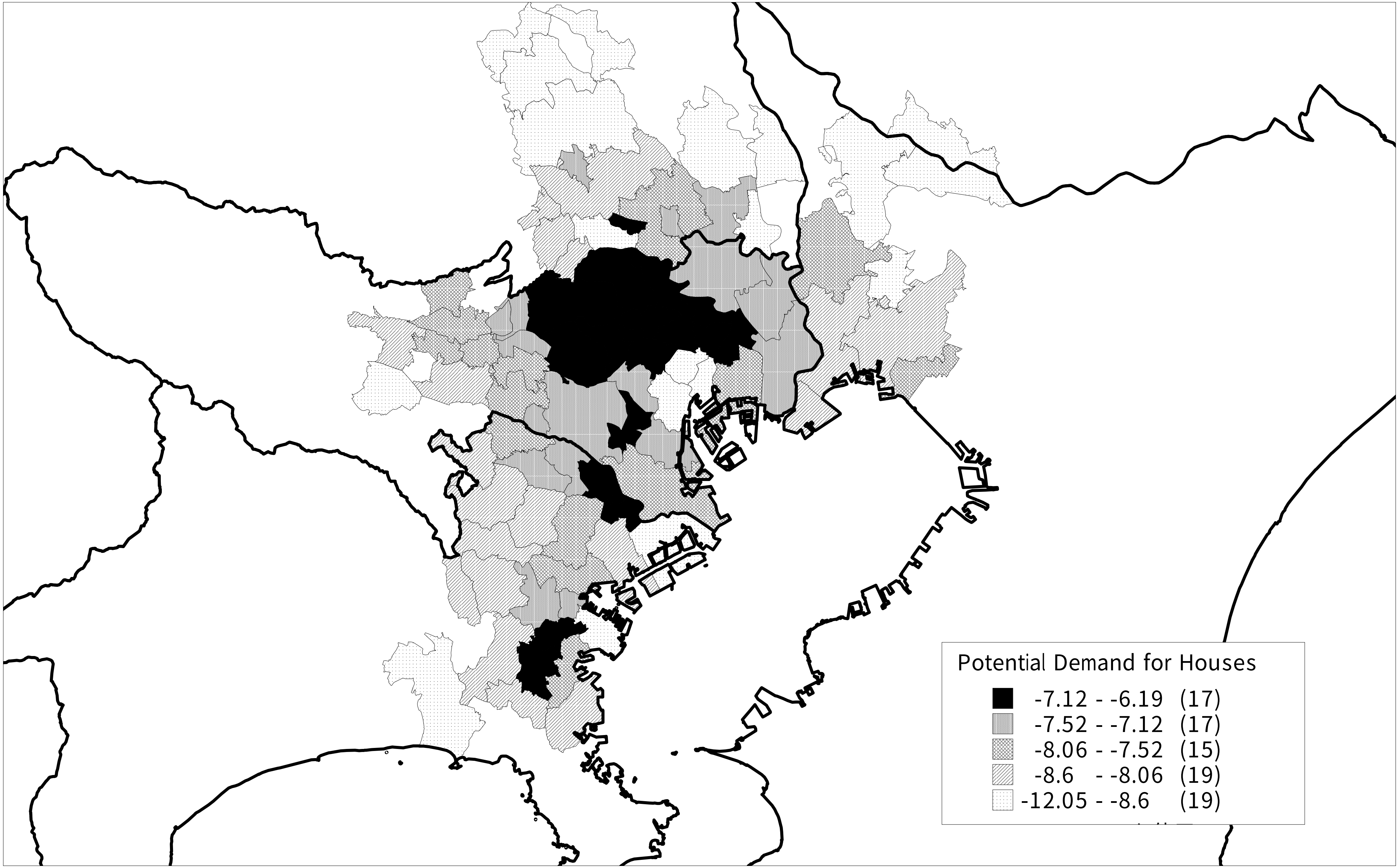

Fig.3 Spatial Distribution of Unobservable Potential Demand for Houses

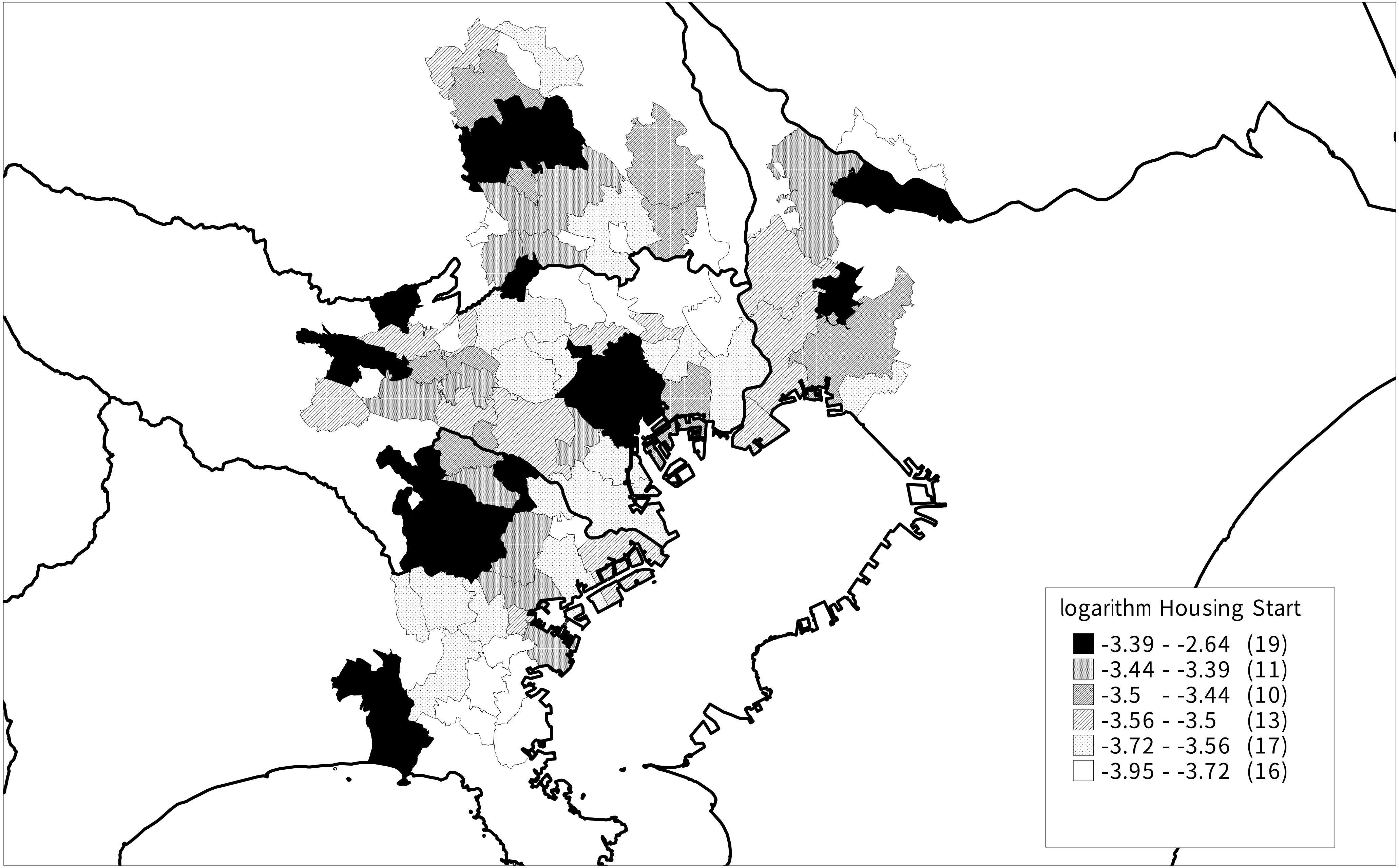

Fig.4 Crude Map of Logarithm of Housing Start per Household