Discussion Paper No. 89

Modeling Spatial Structure of Administrative System in Ponneri, India, in the Late Eighteenth Century

Yukio Sadahiro *

June 2001

*Department of Urban Engineering,
University of Tokyo
7-3-1, Bunkyo-ku, Tokyo 113-8656, Japan

July 2, 2001

Modeling the spatial structure of administrative system in Ponneri, India, in the late eighteenth century

Yukio Sadahiro

Department of Urban Engineering, University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

Phone: +81-3-5841-6273
Fax: +81-3-5841-8521
E-mail: sada@okabe.t.u-tokyo.ac.jp

1. Introduction

Until the middle eighteenth century, South India had been under the rule of the Mughal Empire. There were administrative units called *magans* which are almost equivalent to counties of today. In the late eighteenth century, the colonial policy was introduced by the British and the Mughal Empire had gradually lost its power in South India. To govern the area and collect taxes, the British appointed officers called *zamindaris* and sent them to some of the villages. Each zamindari governed twenty-five villages on average, which formed a new administrative system. We call the system the *zamindari system*.

The spatial structure of zamindari system does not completely agree with that of the magan system, as we will see later in Section 5. They are partly similar but different in some places. Why is the zamindari system spatially different from the magan system, though they are both administrative systems of the same area? This paper aims to answer this question by modeling the zamindari system by various factors including the magan system.

2. Data sources

The data sources used in the analysis are the village accounts compiled by Thomas Barnard (Barnard Report: 1760s-70s), the Permanent Settlement Records on Zamindaris, Poligars, and Pagodas in 1801, and the census map in 1971 (for details, see Mizushima, 2000). There were 144 villages recorded in the Barnard Report , whose location was digitized into GIS by ArcInfo ver. 7.2.1 (Figure 1). The location of villages was estimated by comparing the census map in 1971 and the Barnard Report, because village names had often changed during the two hundred years. A huge amount of data are available about villages which include socio-economic data such as population, caste composition, names of landholders, agricultural products, and so forth.

Figure 1. The studied area: Ponneri, India.

3. Review of existing methods

The zamindari system is regarded as a spatial tessellation, that is, a set of non-overlapping space-filling regions (Okabe *et al*., 2000). Many of the variables that seem influential on the introduction and establishment of the zamindari system, which include the magan system, also have tessellation structures. Therefore, to explain the zamindari system by other variables is, in a broader sense, to build a model representing a spatial tessellation by a set of other spatial tessellations. Such a tessellation modeling often occurs in geography, urban analysis, sociology and ecology; school districts can be modeled by administrative units, land uses, regions bounded by transportation facilities and those characterized by socio-economic attributes of residents; electoral districts are explained by administrative units, local communities, census

tracts, and so forth.

Formally, the problem we are facing is described as follows. Suppose a region S which is divided into a set of subregions by a categorical variable, say, school districts or census tracts. The subregions form a spatial tessellation, which we want to explain by other tessellations. We call the categorical variable the *dependent variable*, as we do in regression analysis. Similarly, we call the tessellation given by the dependent variable the *dependent tessellation*, denoted as $Y=\{y_1, y_2, ..., y_n\}$, where y_i is the ith region in S (Figure 2a). To represent the spatial structure of the tessellations, we use a tessellation indicator function

$$1(\mathbf{u}; y_i) = \begin{cases} 1 & \text{if } \mathbf{u} \in y_i \\ 0 & \text{otherwise} \end{cases}. \quad (1)$$

Equation (1) indicates that the point at $\mathbf{u}$ is contained in y_i.

To model the dependent tessellation Y, we have a set of tessellations given by *independent variables*. Let $X_i=\{x_{i1}, x_{i2}, ..., x_{ini}\}$ be the tessellation given by the ith independent variable, say, a set of administrative units (Figure 2b). The tessellation defined by an independent variable is called the *independent tessellation*. The set of independent tessellations is denoted by $X=\{X_1, X_2, ..., X_m\}$. The tessellation indicator function is also defined for independent tessellations.

Figure 2. An example of the dependent tessellation Y and (b) a set of independent tessellations $X=\{X_1, X_2, ..., X_m\}$.

Our objective is to explain the dependent tessellation Y by a set of independent tessellations X. There are at least three existing approaches to this problem. This section briefly overviews the three methods.

3.1 Contingency table

The contingency table is the joint frequency distribution of the two categorical variables given by the tessellations (Llyod, 1999; Powers and Xie, 2000). Let $f(y_i, x_{jk})$ be the joint frequency distribution of the region pair y_i and x_{jk}. It is mathematically given by

$$f(y_i, x_{jk}) = \int_{\mathbf{u}\in S} 1(\mathbf{u}; y_i) 1(\mathbf{u}; x_{jk}) \mathrm{d}\mathbf{u}. \quad (2)$$

Calculating the above equation for all the pairs of regions y_i and x_{jk}, we obtain the contingency table for Y and X_j. Table 1 shows the contingency table of the tessellations Y and X_1 in Figure 2.

Table 1. The contingency table of the tessellations Y and X_1 shown in Figure 2.

We denote $f(y_\cdot, x_{jk})$ and $f(y_i, x_{j\cdot})$ as the sum of $f(y_i, x_{jk})$ for $i \in \{1, 2, ..., m\}$ and that for

$k \in \{1, 2, ..., n_j\}$, respectively. Let $f(y_\cdot, x_{j\cdot})$ be the total frequency, that is,

$$f(y_\cdot, x_{j\cdot}) = \sum_i \sum_k f(t_i, x_{jk}). \tag{3}$$

If the two tessellations are completely independent, the expected frequency of $f(y_i, x_{jk})$ is given by

$$E\left[f(y_i, x_{jk}) \right] = \frac{f(y_\cdot, x_{jk})}{f(y_\cdot, x_{j\cdot})} \frac{f(y_i, x_{j\cdot})}{f(y_\cdot, x_{j\cdot})}. \tag{4}$$

Therefore, the difference between the expected and observed frequencies indicates the similarity between the two tessellations. The similarity is measured by the Pearson χ^2 statistics; it is statistically tested whether the two tessellations are independent.

Calculating the χ^2 statistics for Y and all the $X_i \in X$, we can detect tessellations in X, if exist, that agree well with Y. This enables us to build a model representing Y.

3.2 Multinomial logit model

An alternative to the contingency table is the multinomial logit model which is frequently used in marketing science (Davies and Rogers, 1984; Ghosh and McLafferty, 1987), transportation science (Ben-Akiva and Lerman, 1985), and so forth. The multinomial logit model describes the individual choice behavior by the utility maximization principle and stochastic instability in utility function. The probability that the ith individual chooses the jth alternative is given by

$$P_{ij} = \frac{\exp(V_{ij})}{\sum_k \exp(V_{ik})}, \tag{5}$$

where V_{ij} is the constant component of the utility of the jth alternative for the ith individual. The multinomial logit model can be applied to a wide variety of individual choice behavior.

To apply the multinomial logit model to tessellation modeling, we regard the categories of dependent variable as alternatives, and points distributed over S as individuals who choose a value of the dependent variable. Independent variables compose the constant component of the utility function. The probability that the point at $\mathbf{u}$ is contained in y_i is then given by

$$P_i(\mathbf{u}) = \frac{\exp(V_i(\mathbf{u}))}{\sum_j \exp(V_j(\mathbf{u}))}, \tag{6}$$

where $V_i(\mathbf{u})$ is a function of the independent variables at $\mathbf{u}$. Given the probability distribution and the form of $V_i(\mathbf{u})$, we can estimate $V_i(\mathbf{u})$ and thus build a model representing the dependent tessellation Y. The maximum likelihood method is generally used for model estimation.

Statistical significance of independent variables is tested by the asymptotic t statistic. The goodness-of-fit of the model is measured by ρ, the fraction of an initial log likelihood value explained by the model.

3.3 Decision tree

The decision tree is a tree-like structure consisting of nodes and links, where nodes represent rules for decision making. Traversing the tree from the root to the terminal node, we obtain the final decision. The decision tree is used in data mining, decision analysis, machine learning, and so forth (Berry and Linoff, 1997; Cios *et al*., 1998).

To apply the decision tree to tessellation modeling, we describe the rules assigned to nodes by functions of independent variables, and regard the terminal nodes as the categories of dependent variable. Figure 3 shows an example of the decision tree.

Figure 3. An example of the decision tree.

Given the dependent and independent tessellations, we can estimate the decision tree by computational algorithms such as CHAID (CHi-square Automatic Interaction Detection, Hartigan, 1975), CART (Classification And Regression Tree, Breiman *et al*., 1984), ID3 (Interactive Dichotomizer 3, Quinlan, 1987), C4.5 (Quinlan, 1993), and so forth.

Unfortunately, these three methods lack the concept of space, that is, they do not explicitly consider the spatial structure of variables, so that in their original form they are not appropriate for modeling of spatial tessellations. The contingency table, for instance, does not take into account the 'spatial' distance between categories. Therefore, the tessellations shown in figure 4b are all equivalent in terms of the agreement with the tessellation shown in figure 4a. The multinomial logit model also suffers from the correlation among the probabilistic error terms assumed for individuals, points distributed over a tessellation. There usually exists the correlation among error terms especially when individuals are spatially distributed.

Figure 4. Agreement between tessellations. (a) A tessellation, (b) tessellations that are all equivalent in terms of the agreement with that shown in figure 4a.

One exception that explicitly considers the spatial structure in agreement of categorical variables is Pontius (2000). He proposes several measures of the agreement of two categorical maps, distinguishing locational disagreement of a categorical variable from its quantification error. However, since the method focuses on comparison of maps and its evaluation, it is not directly applicable to modeling of spatial tessellations.

4. Modeling a spatial tessellation by a set of other spatial tessellations

As discussed in the previous section, existing methods cannot be directly applied to our

problem, that is, to explain the zamindari system by other variables including the magan system. We thus adopt a new method for modeling a spatial tessellation by a set of other spatial tessellations originally developed by Sadahiro (2001a).

The method is designed for exploratory spatial analysis rather than confirmatory analysis (Tukey, 1977; Openshaw *et al*., 1987; Openshaw and Openshaw, 1997; Anselin, 1998). Exploratory spatial analysis searches for interesting patterns and plausible hypotheses in spatial phenomena, which helps more sophisticated spatial analysis such as mathematical and statistical modeling. Exploratory spatial analysis is useful at an early stage of spatial analysis, especially when a huge amount of spatial data are available.

In the following we outline Sadahiro's three methods of tessellation modeling: the region-based method, boundary-based method, and hybrid method (for details, see Sadahiro, 2001a). The region-based method focuses on the regions that compose the dependent tessellation, while the boundary-based method emphasizes boundaries dividing the whole region *S*. The hybrid method is a combination of the two methods.

4.1 Region-based method

The region-based method builds a model representing a dependent tessellation by a set of independent tessellations, focusing on the regions that compose the dependent tessellation. Its basic idea is to decompose the region *S* into subregions and to model them separately by small pieces of independent tessellations (Figure 5).

Figure 5. Modeling a dependent tessellation by a combination of regions in independent tessellations.

To model *Y* in this way, we have to find a set of independent tessellations whose combination agrees well with *Y*. The following algorithm detects such a set of independent tessellations *U*, and gives a set of regions *V*, a model representing *Y*.

Algorithm: Region-based method

Input: A dependent tessellation *Y* and a set of independent tessellations $X=\{X_1, X_2, ..., X_m\}$

Output: A set of independent tessellations *U* and a set of regions *V* modeled by *U*.

Step 1: Set *U* and *V* empty.

Step 2: Do 2.1-2.5 while neither *X* nor *Y* is an empty set.

Step 2.1: Evaluate the agreement between *Y* and X_i for all $X_i \in X$.

Step 2.2: Choose the independent tessellation X_i that gives the best agreement.

Step 2.3: Do 2.3.1 and 2.3.2 for all $y_j \in Y$.

2.3.1: Evaluate the fitness of X_i for y_j.

2.3.2: If the fitness is significant, add y_j to V and remove y_j from Y.

Step 2.4: If any $y_j \in Y$ was added to V, add X_i to U.

Step 2.5: Remove X_i from X.

Step 3: Report U and V.

In step 2.1, the region-based method evaluates the agreement between the dependent tessellation Y and the independent tessellation X_i. To this end we use a measure which we call the agreement index defined as follows.

Let $\gamma(d; Y)$ be the spatial autocorrelation function of the tessellation Y:

$$\gamma(d;Y) = \frac{\int_{\mathbf{u}\in S}\int_{\mathbf{v}\in S,|\mathbf{u}-\mathbf{v}|=d} 1(\mathbf{u},\mathbf{v};Y)\mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u}}{\int_{\mathbf{u}\in S}\int_{\mathbf{v}\in S,|\mathbf{u}-\mathbf{v}|=d}\mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u}}, \tag{7}$$

where

$$1(\mathbf{u},\mathbf{v};Y) = \begin{cases} 1 & \text{if } \exists i, 1(\mathbf{u};y_i)1(\mathbf{v};y_i) = 1 \\ 0 & \text{otherwise} \end{cases}. \tag{8}$$

The spatial autocorrelation function is similar to the covariogram used in geostatistics (Isaaks and Srivastava, 1989; Wackernagel, 1995). It shows a large value if there exists a strong spatial autocorrelation in T. Otherwise, it shows a small value. The spatial autocorrelation function is also calculated for the independent tessellations.

The agreement index is then defined by

$$A(X_i;Y) = \int_{\mathbf{u}\in S}\int_{\mathbf{v}\in S} e(\mathbf{u},\mathbf{v};X_i,Y)\mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u}, \tag{9}$$

where

$$e(\mathbf{u},\mathbf{v};X_i,Y) = \begin{cases} 1-\gamma(d;Y) & \text{if } 1(\mathbf{u},\mathbf{v};Y)=1 \text{ and } 1(\mathbf{u},\mathbf{v};X_i)=1 \\ -\gamma(d;Y) & \text{if } 1(\mathbf{u},\mathbf{v};Y)=1 \text{ and } 1(\mathbf{u},\mathbf{v};X_i)=0 \\ \gamma(d;Y)-1 & \text{if } 1(\mathbf{u},\mathbf{v};Y)=0 \text{ and } 1(\mathbf{u},\mathbf{v};X_i)=1 \\ \gamma(d;Y) & \text{if } 1(\mathbf{u},\mathbf{v};Y)=0 \text{ and } 1(\mathbf{u},\mathbf{v};X_i)=0 \end{cases}. \tag{10}$$

In the following we use it in its standardized form

$$\alpha(X_i;Y) = \frac{A(X_i;Y) - \int_{\mathbf{u}\in S}\int_{\mathbf{v}\in S}\min\{-\gamma(|\mathbf{u}-\mathbf{v}|;Y),\gamma(|\mathbf{u}-\mathbf{v}|;Y)-1\}\mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u}}{\int_{\mathbf{u}\in S}\int_{\mathbf{v}\in S}\max\{1-\gamma(|\mathbf{u}-\mathbf{v}|;Y),\gamma(|\mathbf{u}-\mathbf{v}|;Y)\}\mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u} - \int_{\mathbf{u}\in S}\int_{\mathbf{v}\in S}\min\{-\gamma(|\mathbf{u}-\mathbf{v}|;Y),\gamma(|\mathbf{u}-\mathbf{v}|;Y)-1\}\mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u}}, \tag{11}$$

which satisfies $0 \le \alpha(X_i; Y) \le 1$. Since this index explicitly takes into account the spatial structure of the dependent tessellation, it distinguishes the three tessellations shown in figure 5b from that in figure 5a.

Evaluation of the fitness of X_i for y_j (step 2.3.1) is similar to that of the agreement between tessellations. The *fitness index* of X_i with respect to y_j is defined by

$$\beta(X_i;y_j) = \frac{\int_{\mathbf{u}\in y_j}\int_{\mathbf{v}\in S} e(\mathbf{u},\mathbf{v};X_i,Y)\mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u} - \int_{\mathbf{u}\in y_j}\int_{\mathbf{v}\in S}\min\{-\gamma(|\mathbf{u}-\mathbf{v}|;Y),\gamma(|\mathbf{u}-\mathbf{v}|;Y)-1\}\mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u}}{\int_{\mathbf{u}\in y_j}\int_{\mathbf{v}\in S}\max\{1-\gamma(|\mathbf{u}-\mathbf{v}|;Y),\gamma(|\mathbf{u}-\mathbf{v}|;Y)\}\mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u} - \int_{\mathbf{u}\in y_j}\int_{\mathbf{v}\in S}\min\{-\gamma(|\mathbf{u}-\mathbf{v}|;Y),\gamma(|\mathbf{u}-\mathbf{v}|;Y)-1\}\mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u}}. \tag{12}$$

Its significance is determined by a threshold value β_T given by the analyst. The independent tessellation X_i is regarded to fit y_j significantly if

$$\beta\left(X_i; y_j\right) \geq \beta_T . \tag{13}$$

The choice of the threshold β_T depends on the circumstances. If analysis is at an early stage, a small value would be appropriate because it permits a loose fitness of tessellations so that various independent variables can be discussed later. When the focus is on only important variables, the threshold β_T should be large so that independent tessellations closely fit the dependent tessellation.

In addition to categorical variables, numerical variables can be taken into account in the region-based method. We categorize a numerical variable into several classes to obtain a categorical variable, maximizing the agreement between the numerical and dependent variables with respect to their spatial tessellations.

4.2 Boundary-based method

The region-based method focuses on the regions that compose the dependent tessellation, considering them as the essential components of the tessellation. The boundary-based method, on the other hand, emphasizes the boundaries between regions rather than the regions themselves. The basis of the boundary-based method is the viewpoint that the tessellation is generated by dividing a region into subregions, not combining subregions into a larger region.

Suppose an example shown in figure 6. The independent tessellations X_1, X_2, and X_3 are all different from the dependent tessellation Y. However, all the three tessellations partly agree with Y for some of the boundaries, which are shown by the solid lines in the figure. Consequently, the dependent tessellation Y can be modeled by a combination of some of the boundaries in X_1, X_2, and X_3. This is the basic idea of the boundary-based method.

Figure 6. Modeling a dependent tessellation by a combination of boundaries in independent tessellations.

The algorithm of the boundary-based method is substantially the same as that of the region-based method, except it evaluates the agreement between boundaries of regions, not regions themselves. It successively tries the independent tessellations that give the best agreement with the dependent tessellation and finally yields a model representing the dependent tessellation.

Evaluation of agreement, therefore, is performed on a boundary basis. Let $b(y)_{ij}$ be the boundary between the regions y_i and y_j. The set of boundaries that composes the dependent tessellation Y is $B(Y)=\{b(Y)_{ij}, i, j \in N\}$, where $N=\{1, 2, ..., n\}$. We define the boundary indicator function:

$$1(y_i, y_j) = \begin{cases} 1 & \text{if } b(Y)_{ij} \text{ exists} \\ 0 & \text{otherwise} \end{cases}. \quad (14)$$

The boundaries of the independent tessellation X_i are similarly represented. The boundary between the regions x_{ij} and x_{ik} is denoted by $b(X_i)_{jk}$, and the set of boundaries of X_i is $B(X_i)=\{b(X_i)_{jk}, j, k \in N_i\}$, where $N_i=\{1, 2, ..., n_i\}$. The agreement between the independent tessellation $B(X_i)$ and the dependent tessellation $B(Y)$ is measured by

$$A'\left(B(X_i); B(Y)\right) = \sum_{j,k,1(y_j,y_k)=1} \int_{\mathbf{u}\in y_j \cup y_k} \int_{\mathbf{v}\in y_j \cup y_k} e(\mathbf{u}, \mathbf{v}; X_i, Y) \mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u} \quad (15)$$

and its standardized form

$$\alpha'\left(B(X_i); B(Y)\right) = \frac{A'(X_i; Y) - \sum_{j,k,1(y_j,y_k)=1} \int_{\mathbf{u}\in y_j \cup y_k} \int_{\mathbf{v}\in y_j \cup y_k} \min\{-\gamma(|\mathbf{u}-\mathbf{v}|; Y), \gamma(|\mathbf{u}-\mathbf{v}|; Y) - 1\} \mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u}}{\sum_{j,k,1(y_j,y_k)=1} \left[\int_{\mathbf{u}\in y_j \cup y_k} \int_{\mathbf{v}\in y_j \cup y_k} \max\{1-\gamma(|\mathbf{u}-\mathbf{v}|; Y), \gamma(|\mathbf{u}-\mathbf{v}|; Y)\} \mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u} - \int_{\mathbf{u}\in y_j \cup y_k} \int_{\mathbf{v}\in y_j \cup y_k} \min\{-\gamma(|\mathbf{u}-\mathbf{v}|; Y), \gamma(|\mathbf{u}-\mathbf{v}|; Y) - 1\} \mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u} \right]}. \quad (16)$$

The fitness of $B(X_i)$ for $b(Y)_{jk}$ is evaluated by

$$\beta\left(B(X_i); b(Y)_{jk}\right) = \frac{\int_{\mathbf{u}\in y_j \cup y_k} \int_{\mathbf{v}\in y_j \cup y_k} e(\mathbf{u}, \mathbf{v}; X_i, Y) \mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u} - \int_{\mathbf{u}\in y_j \cup y_k} \int_{\mathbf{v}\in y_j \cup y_k} \min\{-\gamma(|\mathbf{u}-\mathbf{v}|; Y), \gamma(|\mathbf{u}-\mathbf{v}|; Y) - 1\} \mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u}}{\int \int \max\{1-\gamma(|\mathbf{u}-\mathbf{v}|; Y), \gamma(|\mathbf{u}-\mathbf{v}|; Y)\} \mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u} - \int \int \min\{-\gamma(|\mathbf{u}-\mathbf{v}|; Y), \gamma(|\mathbf{u}-\mathbf{v}|; Y)}. \quad (17)$$

Its significance is again judged by the threshold value β_T' given by the analyst.

4.3 Hybrid method

The hybrid method is a combination of the region-based and boundary-based methods, which inherits strengths from both of them. It performs the region-based method first and then applies the boundary-based method to a part of the original dependent tessellation not modeled by the region-based method. It is expected that the hybrid method yields better result than the individual methods.

5 Modeling of the zamindari system

Using the methods described in the previous section, this section analyzes the zamindari system using a set of other independent variables available. Before the zamindari system was introduced in the late eighteenth century, the magan system had existed, as mentioned earlier. Figure 7 shows the two administrative systems, the magan and zamindari systems, by Voronoi diagrams. Since the village boundary was not known in map format, it was approximated by the Voronoi diagram in which villages were used as generators.

Figure 7. Two administrative systems in Ponneri in the late eighteenth century. (a) The magan and (b) zamindari systems.

Interestingly, the zamindari system does not completely agree with the magan system. They agree well in the central region of Ponneri, but have different structures in its surroundings. Why is the zamindari system different from the magan system?

To answer this question, we first visually compared the map of zamindari system with maps of other variables that might have affected its introduction and establishment (Aono, 2000; Fuko, 2001). However, this process was quite difficult and inefficient because of a huge amount of attribute data; it took a long time to compare maps visually. This experience led us to develop the exploratory modeling method in Sadahiro (2001a). We should also note that, through the visual analysis, we noticed that the zamindari system might be explained by the spatial combination of the magan system and other factors; a part of the zamindari system not explained by the magan system seemed to be modeled by the tessellations given by other variables. This inspired us to the spatial decomposition approach of the region-based method.

The dependent tessellation is the zamindari system represented by a set of Voronoi regions shown in Figure 7b. From attribute data of villages we chose fifteen variables as independent variables (Table 2). They were also transformed into Voronoi diagrams, the independent tessellations.

Table 2. Independent variables used in analysis.

We first applied the region-based method to model the zamindari system. The threshold value β_T was set to 0.99. As shown in Table 3, the method reported at the first execution of step 2.1 that the magan system is the most influential among all the variables. We thus removed the regions explained by the magan system and investigated the other independent variables in turn. However, since any of the other variables did not show significant fitness for regions left in *Y*, the final result *U* contains only the magan system. Figure 8 shows a set of regions *V* explained by the magan system.

Table 3. Modeling the zamindari system by the region-based method.

Figure 8. Zamindari regions modeled by the region-based method.

The result is quite reasonable because it is unlikely that the zamindari system completely ignored the existing magan system. However, the area of zamindari regions explained by the magan system accounts for only 22.43 %, which is not satisfactory. We then applied the boundary-based method to the same data, with the threshold value β_T' 0.99. Unfortunately, it did not improve the result significantly, so we applied the hybrid method and obtained the result shown in Table 4 and Figure 9.

Table 4. Modeling the zamindari system by the hybrid method.

Figure 9. Zamindari boundaries modeled by the hybrid method.

Combination of the two methods yielded better result than that given by the individual methods. Among all the boundaries 76.77 % are explained by four independent variables: the magan system, dominant caste, poligar system, and population. Poligars were the military who were assigned the role to keep safe and order, so the poligar system was in a sense another administrative system in those days. Consequently, it is understandable that the poligar system affected the spatial structure of the zamindari system. Analysis also detected the ratio of dominant caste as an influential factor. There were a number of villages in which a certain caste accounted for a large proportion of residents, say, pariah (untouchable), vellalar (farmer), and idaiyar (cowkeeper). Such villages were usually characterized by their dominant castes, and thus it is possible that the existence of dominant caste affected the zamindari system.

From the above result we can at least say that the magan system is the most influential among various factors on the spatial structure of zamindari system. This naturally leads to a question of the formation of magan system: how was it formed? To answer this, we next analyze the magan system using other variables except the zamindari system. Among the three methods we chose the hybrid method because it yielded the best result in the previous analysis. The threshold values β_T and β_T' were set to 0.99. The result is shown in Table 5 and Figure 10.

Table 5. Modeling the magan system by the hybrid method.

Figure 10. Magan boundaries modeled by the hybrid method.

The hybrid method explains 84.26 % of the boundaries of magan system by four independent variables: the poligar system, irrigated farmland, dominant caste, and population. The poligar system (Figure 11) is the most influential among the variables, which suggests that the magan system has its root in the poligar system. This seems to reflect a chain of administrative systems: the poligar system, the magan system, and finally the zamindari system. Administrative systems are always partly inherited by newer systems.

Figure 11. The poligar system.

7 Conclusion

In this paper we have built a model representing the zamindari system in Ponneri, India, in the late eighteenth century. As reviewed in Section 3, there are at least three existing methods

available for modelling spatial tessellations: the contingency table, multinomial logit model, and decision tree. These methods, however, lack the concept of space, that is, they do not explicitly consider the spatial structure of variables, so that in their original form they are not appropriate for modeling of spatial tessellations. We thus used the exploratory modeling methods proposed by Sadahiro (2001a). Among the three methods the hybrid method gave the best result in terms of model fitting; 76.77 % of the zamindari boundary was explained by four independent variables: the magan system, dominant caste, poligar system, and population. We then applied the hybrid method to the magan system to explain it by other factors expect the zamindari system. The result showed that the magan system is mainly described by the poligar system. These results reflect a chain of administrative systems: the poligar system, the magan system, and finally the zamindari system.

We finally discuss some limitations of our study for future research. First, spatial tessellations are not only formed by other spatial tessellations but also affected by other spatial objects such as points, lines, scalar and vector fields. For instance, when school districts are determined, the location of transportation facilities is always taken into consideration. The zamindari and magan systems might have been affected by other types of spatial objects, say, road networks, water system, terrain elevation, and so forth. A more flexible model that treats a wide variety of spatial objects should be developed and applied. Second, benchmark values for the thresholds β_T and β_T' used in fitness evaluation should be discussed further. In our study we found that values around 0.99 work successfully in terms of the balance between the rigidity and looseness of model fitting. However, this value may not be appropriate for other applications. Though the thresholds can be determined arbitrarily, it is convenient if some benchmark values are presented through empirical studies. Third, we analyzed the zamindari and magan systems, administrative systems in the late eighteenth century. We found that the zamindari system is closely related to the magan system, and it seems that the magan system has its root in the poligar system. Administrative systems are always partly inherited by newer systems. Consequently, it is quite important to investigate them in the spatio-temporal domain, that is, spatio-temporal analysis of administrative systems. Unfortunately, there are few methods applicable to analysis of tessellation change (Langran, 1992; Sadahiro, 2001b, 2001c). The methodology of spatio-temporal analysis of spatial tessellations should be developed in future. In addition, the studied area and period of this paper should be extended to make the results of analysis more general and persuasive.

Acknowledgment

The author is grateful to A. Okabe and T. Mizushima for fruitful discussion. He also thanks A. Masuyama, E. Shimizu and T. Sato for their valuable comments. This research was partly supported by the Ministry of Education, Culture, Sports, Science and Technology,

Grant-in-Aid for Creative Basic Research, 09NP1301, 1997-2001.

References

Anselin, L. (1998): "Exploratory spatial data analysis in a geocomputational environment", in *Geocomputation: A Primer*, Longley, P. A., S. M. Brooks, R. McDonnell, and B. Macmillan (eds), Chichester: John Wiley, pp 77-94.

Aono, S. (2000): *Spatial Structure of Caste System in the Late Eighteenth Century in Ponneri, India*. Unpublished graduation thesis, Department of Urban Engineering, University of Tokyo, Tokyo.

Ben-Akiva, M. and S. Lerman (1985): *Discrete Choice Analysis: Theory and Application to Travel Demand*. Massachusetts: MIT Press.

Berry, M. J. A., G. S. Linoff (1997): *Data Mining Techniques: For marketing, Sales, and Customer Support*. New York: John Wiley.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1984): *Classification and Regression Tree*s. New York: Chapman & Hall.

Cios, K., W. Pedrycz, and R. Swiniarski (1998): *Data Mining Methods for Knowledge Discovery*. Boston: Kluwer.

Davies, R. L. and D. S. Rogers (1984): *Store Location and Store Assessment Research*. New York: John Wiley.

Fuko, S. (2001): *Establishment of Administrative System in the Late Eighteenth Century in Ponneri, India*. Unpublished graduation thesis, Department of Urban Engineering, University of Tokyo, Tokyo.

Ghosh, A. and S. McLafferty (1987): *Location Strategies for Retail and Service Firms*. Lexington, Massachusetts: D. C. Heath and Co.

Hartigan, J. A. (1975): *Clustering Algorithms*. New York: John Wiley.

Isaak, E. H. and R. M. Srivastava (1989): *Applied Geostatistics*. New York: Oxford University Press.

Langran, G. (1992): *Time in Geographic Information Systems*. London: Taylor & Francis.

Lloyd, C. J. (1999): *Statistical Analysis of Categorical Data*. New York: John Wiley.

Mizushima, T, 2000 *Mirasi System as Social Grammar - State, Local Society, and Raiyat in the 18th-19th South India* - a report available on the website: http://www.l.u-tokyo.ac.jp/~zushima9/Archive/1-28.doc.

Openshaw, S., M. Charlton, C. Wymer, and A. Craft (1987): "A Mark 1 Geographical Analysis Machine for the Automated Analysis of Point Data Sets," *International Journal of Geographical Information Systems*, **1**, 335-358.

Openshaw, S. and C. Openshaw C. (1997): *Artificial Intelligence in Geography*. Chichester: John Wiley.

Pontius, R. G. Jr. (2000): "Quantification Error Versus Location Error in Comparison of Categorical Maps," *Photogrammetric Engineering and Remote Sensing*, **66**, 1011-1016.

Powers, D. A. and Y. Xie (2000): *Statistical Methods for Categorical Data Analysis*. San Diego: Academic Press.

Okabe, A., B. Boots, K. Sugihara, and S.-N. Chiu (2000): *Spatial Tessellations: Concepts and Applications of Voronoi Diagram*. Chichester: John Wiley.

Quinlan, J. R. (1987): "Simplifying Decision Trees," *International Journal of Man-Machine Studies*, **27**, 221-334.

Quinlan, J. R. (1993): *C4.5 Programs for Machine Learning*. San Mateo: Morgan Kaufmann.

Sadahiro , Y. (2001a): "Exploratory Modeling of a Spatial Tessellation by a Set of Other Spatial Tessellations," *Discussion Paper Series*, **88**, Department of Urban Engineering, University of Tokyo.

Sadahiro, Y. (2001b): "Analysis of Surface Changes Using Primitive Events," *International Journal of Geographical Information Science*, **15**, to appear.

Sadahiro, Y. (2001c): "Spatio-temporal Analysis of Surface Changes by Tessellations," *Geographical Analysis*, **33**, to appear.

Tukey, J. W. (1977): *Exploratory Data Analysis*. New York: Addison-Wesley.

Wackernagel, H. (1995): *Multivariate Geostatistics*. Berlin: Springer.

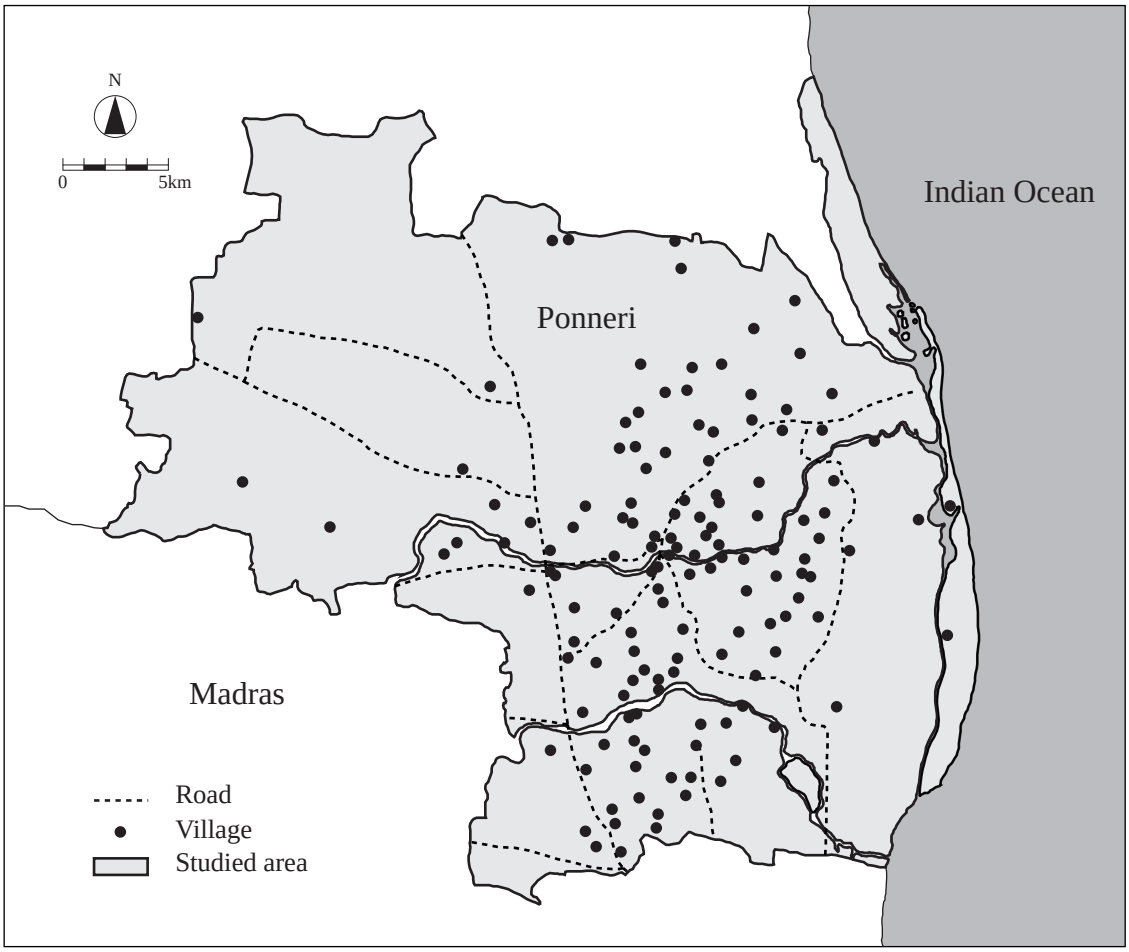
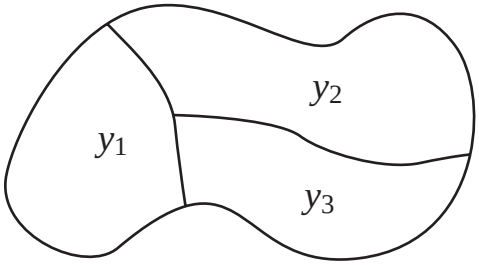

N
0
5km
Indian Ocean
Ponneri
Madras
Road
Village
Studied area

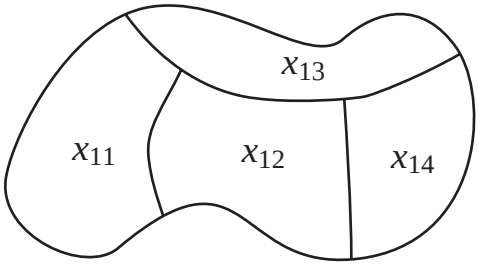
Figure 1

Dependent tessellation Y

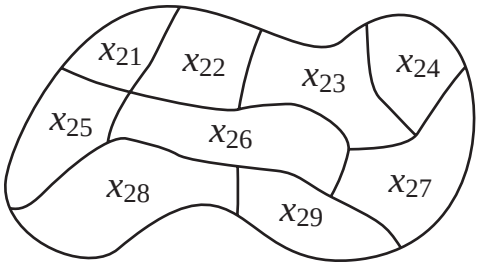


(a)

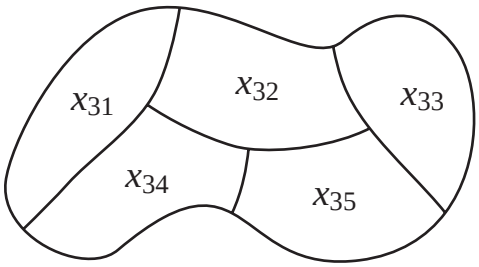
Independent tessellations X_1



Independent tessellations X_2



Independent tessellations X_3



(b)

Figure 2

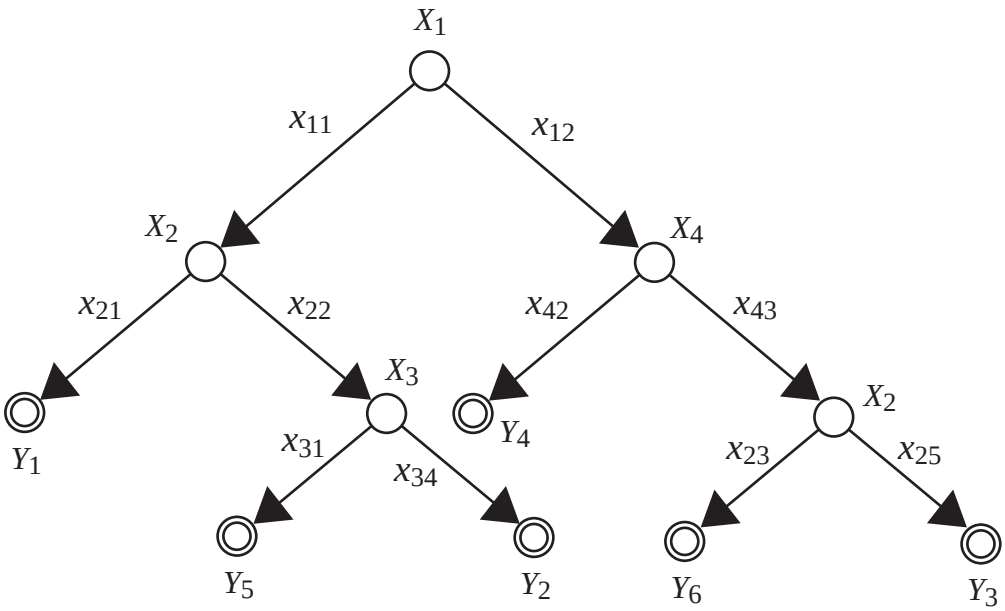
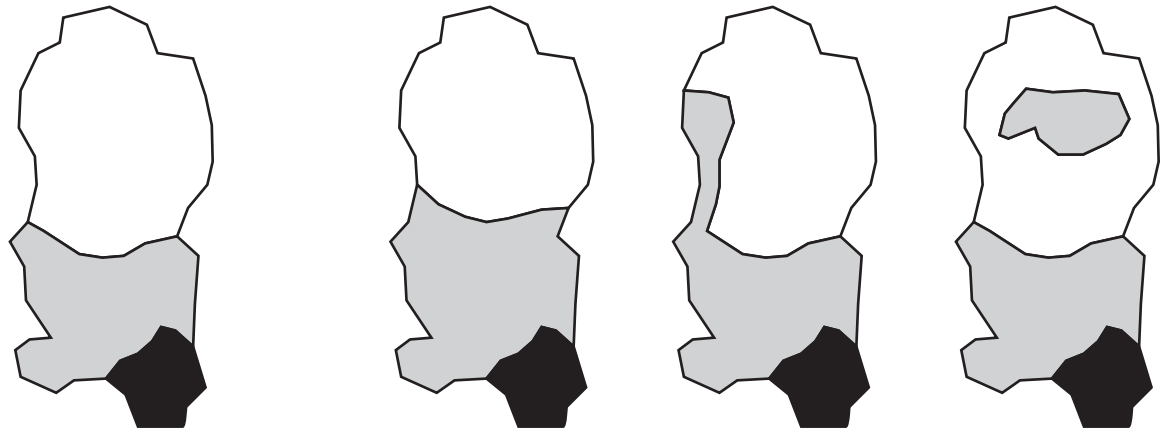

X_1
x_{11}
x_{12}
X_2
X_4
x_{21}
x_{22}
x_{42}
x_{43}
X_3
X_2
Y_1
Y_4
x_{31}
x_{34}
x_{23}
x_{25}
Y_5
Y_2
Y_6
Y_3

Figure 3



(a) (b)

Figure 4

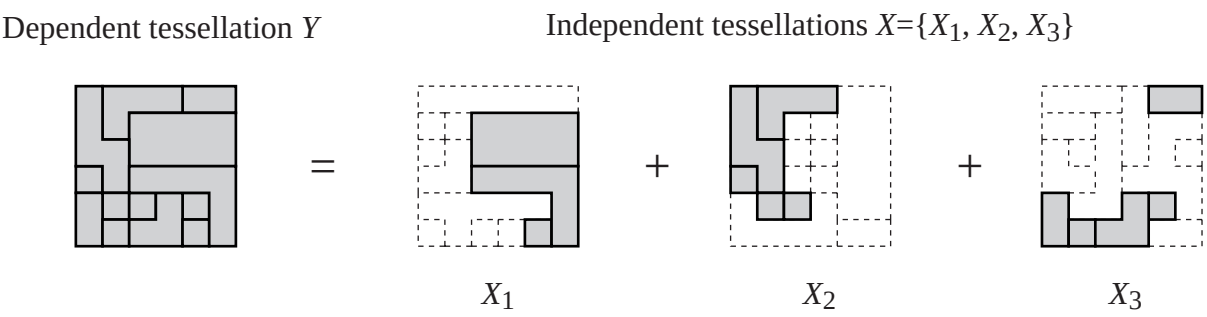

Dependent tessellation Y
Independent tessellations $X=\{X_1, X_2, X_3\}$
=
+
+
X_1
X_2
X_3

Figure 5

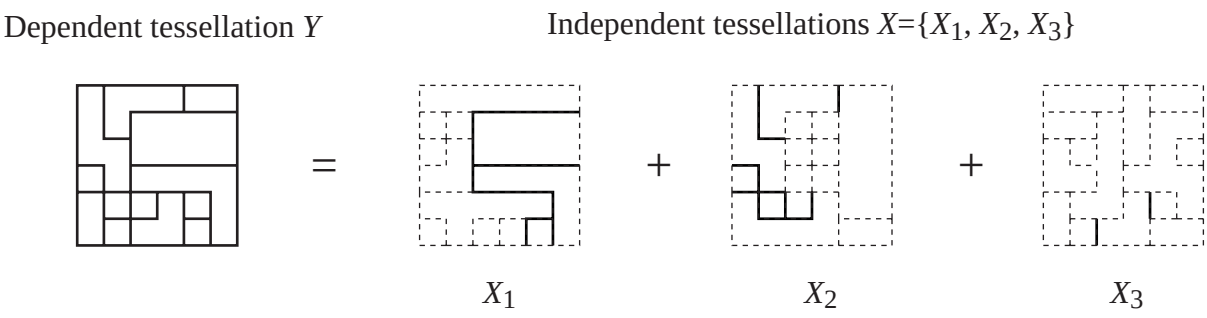
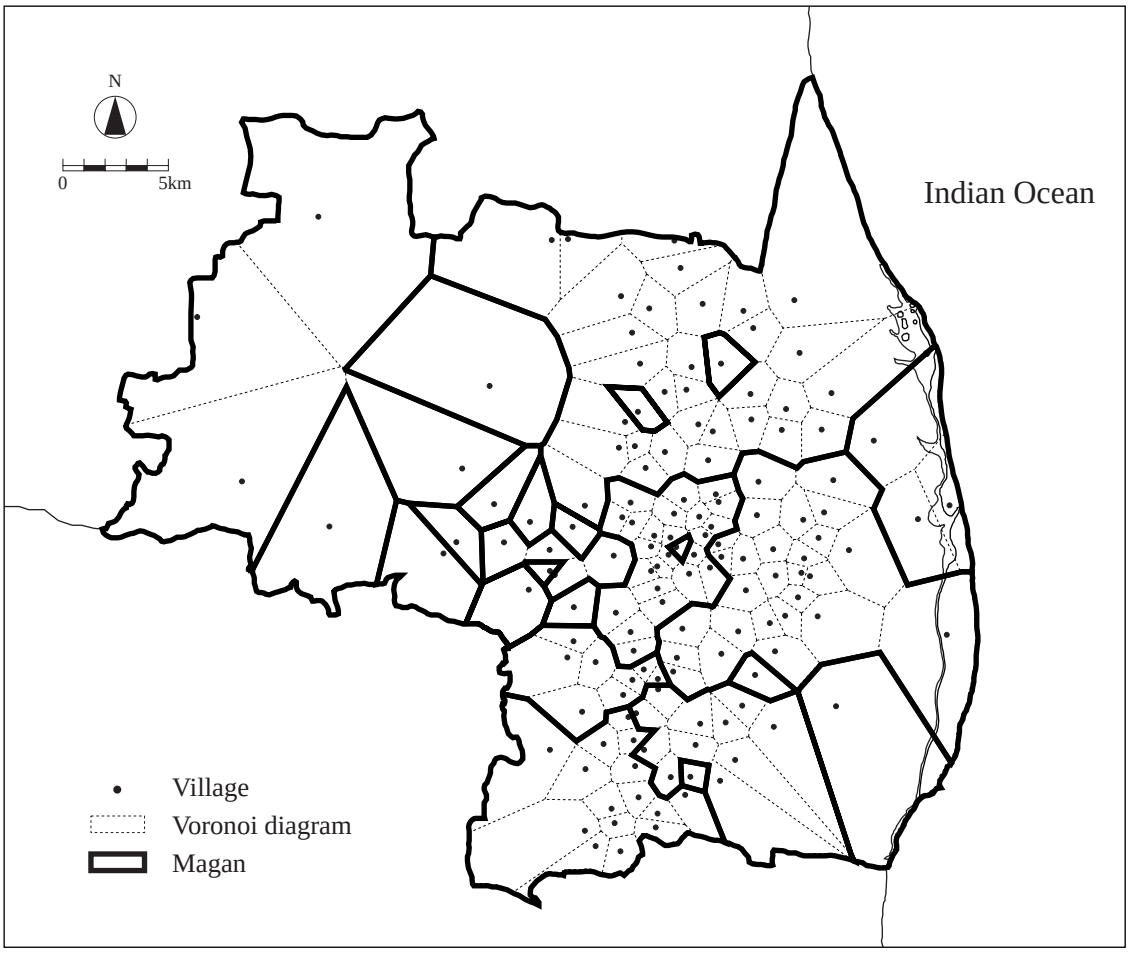
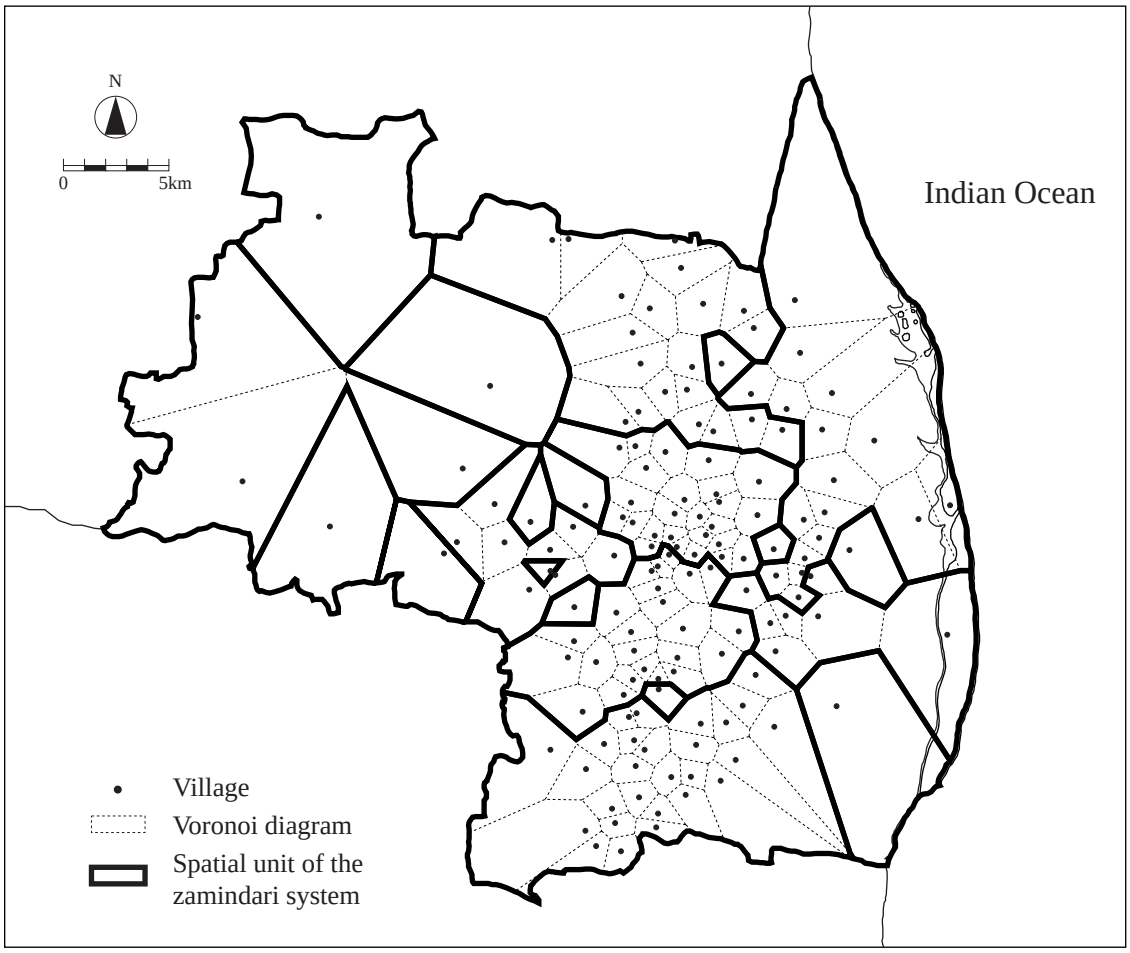

Dependent tessellation Y
Independent tessellations $X=\{X_1, X_2, X_3\}$
=
+
+
X_1
X_2
X_3

Figure 6



(a)

Figure 7a


N
0
5km
Indian Ocean
Village
Voronoi diagram
Spatial unit of the zamindari system

(b)

Figure 7b

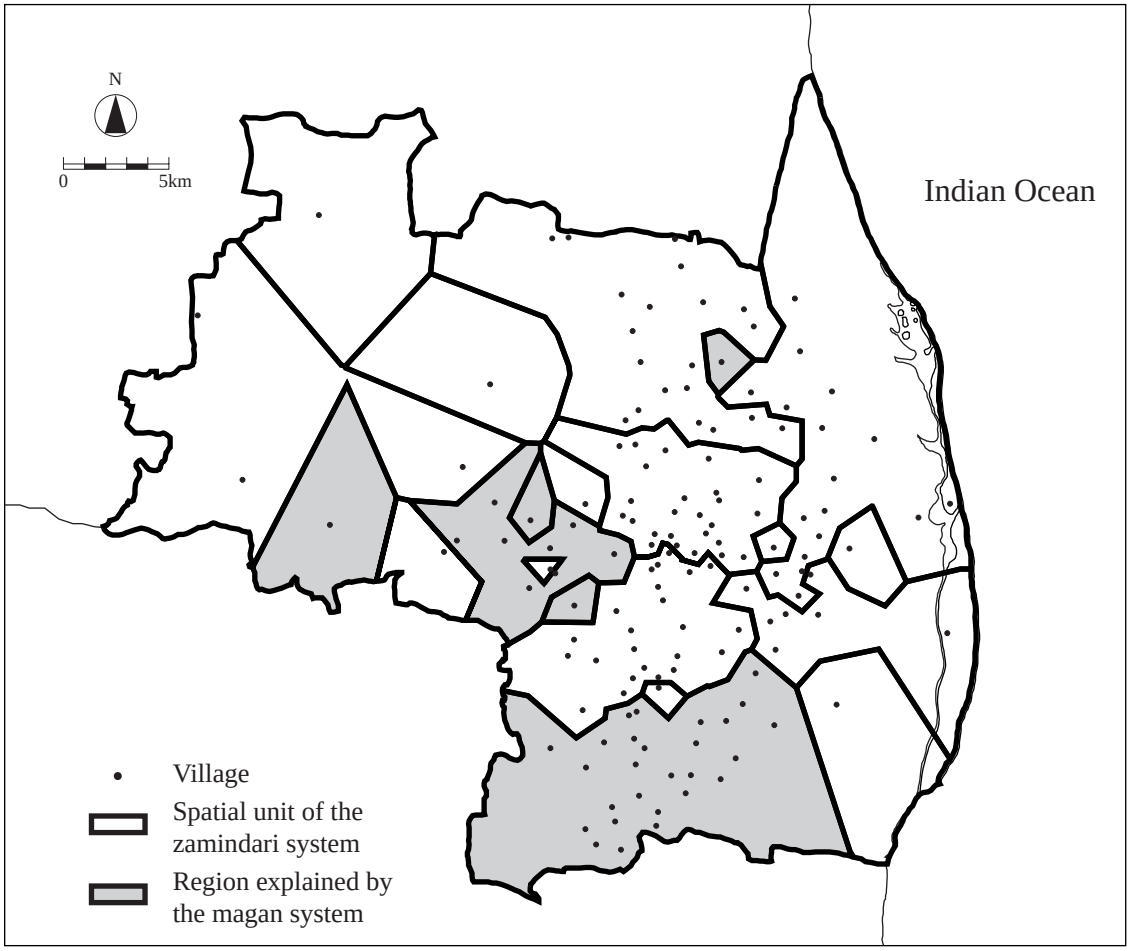

N
0
5km
Indian Ocean
Village
Spatial unit of the zamindari system
Region explained by the magan system

Figure 8

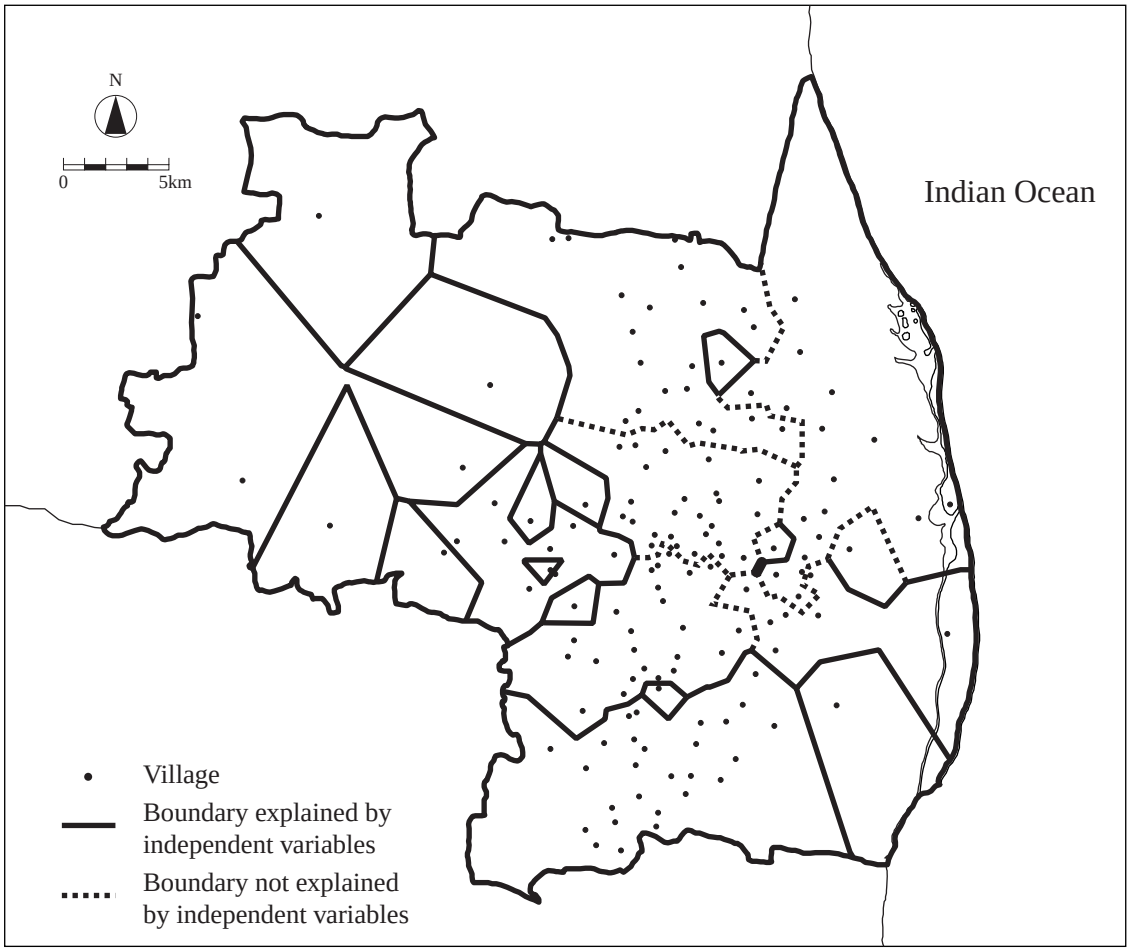

N
0
5km
Indian Ocean
Village
Boundary explained by independent variables
Boundary not explained by independent variables

Figure 9

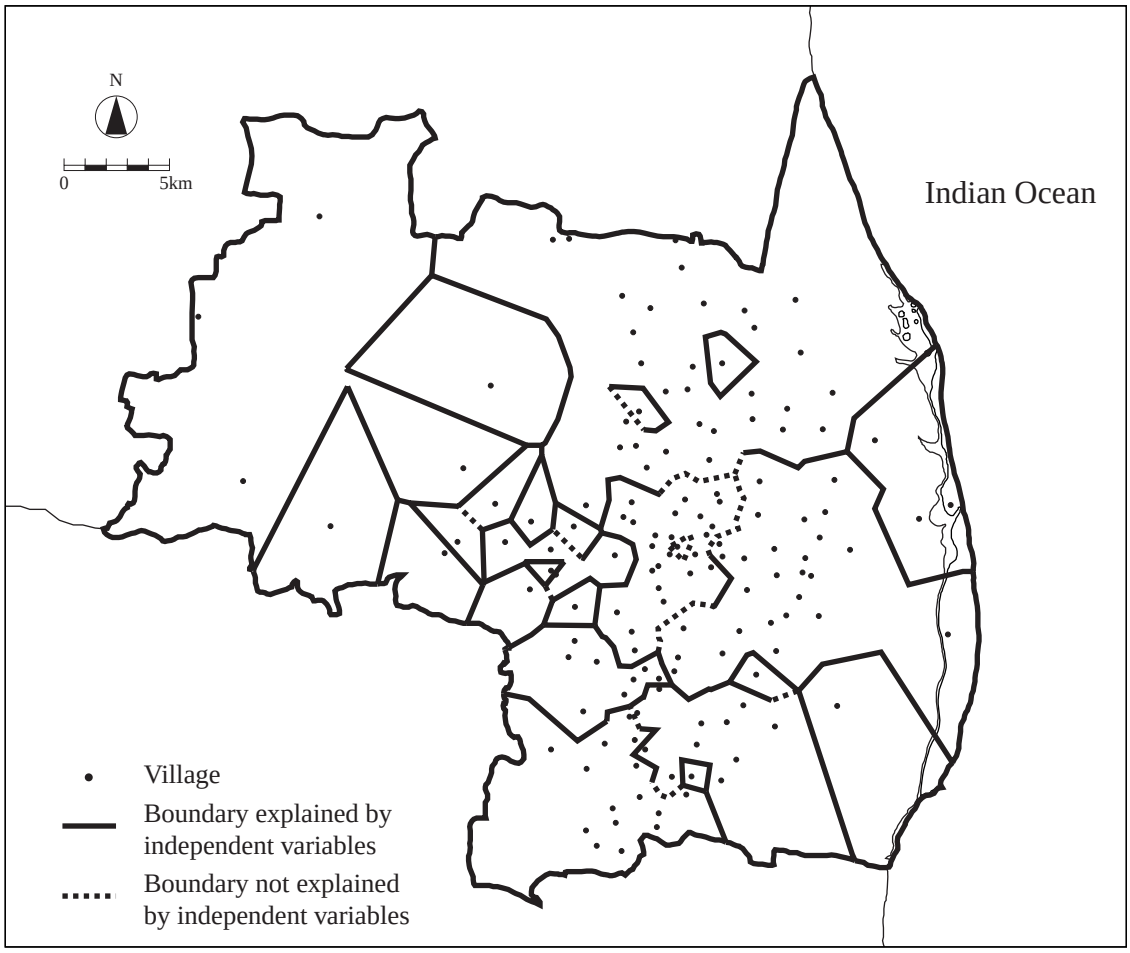

N
0
5km
Indian Ocean
Village
Boundary explained by independent variables
Boundary not explained by independent variables

Figure 10

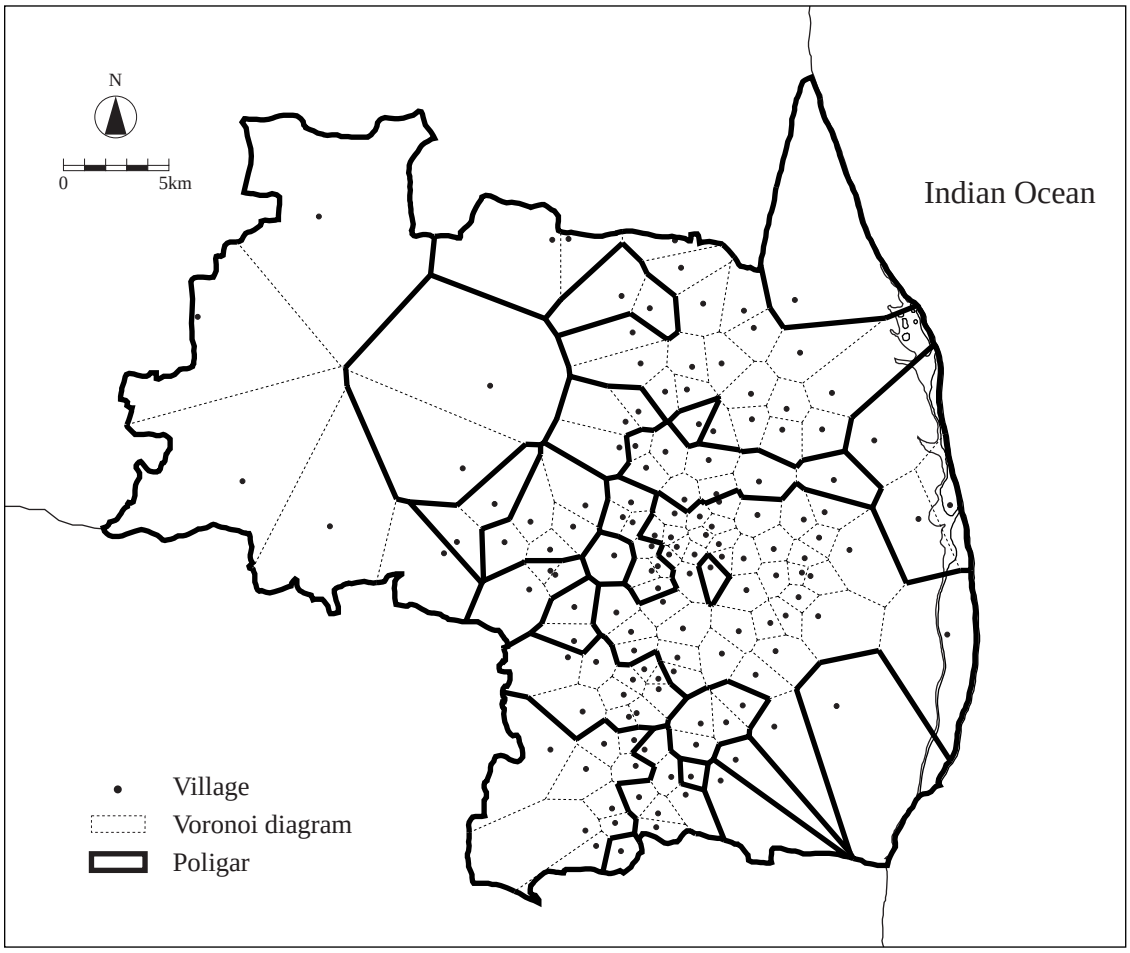


Figure 11

		Tessellation Y		
		y_1	y_2	y_3
Tessellation X_1	x_{11}	674	38	0
	x_{12}	78	164	437
	x_{13}	0	514	0
	x_{14}	0	282	201

Table 1

Categorical variables

Magan	Administrative unit in the middle eighteenth century
Poligar	Military assigned the role to keep safe and order
Caste composition	Village category based on caste composition
Dominant caste	Existence of a dominant caste (binary variable)
Brahman	The highest priest of Hindu
Crop type	Village category based on crops cultivated

Numerical variables

Population	Number of residents
Area	Area of a village
Caste homogeneity	Homogeneity in caste composition measured by the extropy index
Irrigated farmland	Ratio of the area of irrigated farmland to that of the total farmland
Wasteland	Ratio of the area of wasteland to that of the whole village
Forest	Ratio of the area of forest to that of the whole village
State-owned land	Ratio of the area of land owned by state to that of the whole village
Hoe	Number of hoes
Tree	Number of trees

Table 2

Agreement index $\alpha(X_i; Y)$ at the first execution of Step 2.1

Magan	0.8249
Poligar	0.7682
Caste composition	0.4648
Dominant caste	0.6076
Brahman	0.3497
Crop type	0.4105
Population	0.7512
Area	0.7548
Caste homogeneity	0.7401
Irrigated farmland	0.7740
Wasteland	0.7669
Forest	0.7623
State-owned land	0.7495
Hoe	0.4681
Tree	0.4676

The best agreement index $\alpha(X_i; Y)$ after the first execution of Step 2.1

Irrigated farmland	0.7591
Wasteland	0.7496
Forest	0.7245
State-owned land	0.7231
Dominant caste	0.7189
Population	0.7158
.	.
.	.
.	.

Table 3

The best agreement index $\alpha(X_i; Y)$ reported in the region-based method

Magan	0.8249

The best agreement index $\alpha(X_i; Y)$ reported in the boundary-based method

Dominant caste	0.8974
Poligar	0.8952
Population	0.8911
Wasteland	0.8902
Area	0.8900
State-owned land	0.8892
.	.
.	.
.	.

Table 4

The best agreement index $\alpha(X_i; Y)$ reported in the region-based method

Poligar	0.7759

The best agreement index $\alpha(X_i; Y)$ reported in the boundary-based method

Irrigated farmland	0.8523
Poligar	0.8952
Population	0.8911
Wasteland	0.8902
Area	0.8900
State-owned land	0.8892
.	.
.	.
.	.

Table 5