Discussion Paper No. 88

Exploratory Modeling of a Spatial Tessellation by a Set of Other Spatial Tessellations

Yukio Sadahiro *

June 2001

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

June 29, 2001

Exploratory modeling of a spatial tessellation by a set of other spatial tessellations

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

Exploratory modeling of a spatial tessellation by a set of other spatial tessellations

Abstract

Spatial tessellation is one of the most important spatial structure in geography. There are various types of spatial tessellations such as administrative units, school districts, census tracts, and so forth. Spatial tessellations are often closely related to each other; school districts are determined by, say, administrative units and land uses; electoral districts are based on administrative units, local communities, census tracts, and so forth. Such relationships among spatial tessellations have drawn attention of geographers; why and how are they formed? To answer this question, this paper proposes three methods for modeling a spatial tessellation by a set of other tessellations: the region-based method, boundary-based method, and hybrid method. They are all designed for exploratory spatial analysis rather than confirmatory analysis. The methods are evaluated through an empirical study, analysis of the administrative system in Ponneri, India, in the late eighteenth century.

1 Introduction

Spatial tessellation is one of the most important spatial structure in geography. There are various types of spatial tessellations such as administrative units, school districts, electoral districts, census tracts, vegetation pattern, land uses and land covers. Artificial tessellations are also used in geography such as Voronoi diagrams, Delaunay tessellations, lattices and grids, in order to approximate catchment areas of urban facilities or to aggregate a set of spatial data on a common zonal system (Cressie, 1993; Sadahiro, 1999; Okabe *et al*., 2000).

Spatial tessellations are often closely related to each other. Electoral districts are based on administrative units, local communities, census tracts, and so forth. Administrative units are related to land uses, local communities, school districts, and a tessellation given by transportation facilities such as roads and railways. School districts are determined by administrative units, land uses, regions bounded by transportation facilities and those characterized by socio-economic attributes of residents; it even happens that the root of today's school districts go back to past administrative units. Such relationship among spatial tessellations has drawn attention of geographers; why and how are they formed?

To model a spatial tessellation by other tessellations, we usually begin with visual comparison of tessellation maps in order to find hypotheses for further sophisticated analysis. Visual analysis is an indispensable step of exploratory spatial analysis, especially in exploring plausible hypotheses (Openshaw *et al*., 1987; Openshaw and Openshaw, 1997). Comparing two maps, we may find similarity between the tessellations which suggests an influence of one tessellation on the other.

Visual analysis, though it is quite important, is not efficient when numerous maps have to be compared. To treat a great number of maps is not an unrealistic supposition, because today we have numerous spatial databases and not a few data are aggregated across spatial units to form spatial tessellations. In addition, visual analysis heavily depends on human perception of images so that it tends to be subjective and ambiguous, and consequently the result is often unreliable. A more efficient and objective method that helps us find interesting and reasonable hypotheses is necessary.

The use of quantitative methods and GIS makes exploratory analysis more efficient and persuasive. Unfortunately, there exists only a few quantitative methods available for modeling a spatial tessellation by a set of other spatial tessellations. If tessellations are based on categorical variables, one option is to use the contingency table (Llyod, 1999; Powers and Xie, 2000) and calculate the Kappa index (Cohen, 1960; Landis and Koch, 1977; Cartersen, 1987). Considering two spatial tessellations as a representation of two categorical variables, we can measure the similarity between them and test its statistical significance. An alternative is to apply the multinomial logit model where one tessellation represents the dependent variable while others are independent variables (Ben-Akiva and Lerman, 1985). The multinomial logit model

has wider applicability than the contingency table because it can incorporate not only categorical variables but also numerical variables as independent variables. 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. The contingency table, for instance, does not take into account the 'spatial' distance between categories. Therefore, the tessellations shown in figure 1b are all equivalent in terms of the agreement with the tessellation shown in figure 1a. 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 1. Agreement between tessellations. (a) A tessellation, (b) tessellations that are all equivalent in terms of the agreement with that shown in figure 1a.

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.

To fill the gap of the research, this paper proposes a method for modeling a spatial tessellation by a set of other spatial tessellations. One of the motivations of this study was provided by an experience of joint study with researchers in Islamic area studies. As described in section 6, we had to compare numerous maps of spatial tessellations visually in order to propose hypotheses for modeling a spatial tessellation, an administrative system in the late eighteenth century in a county in India. It was a time-consuming task which promoted us to develop an exploratory method for modeling a spatial tessellation. Consequently, the method is designed for exploratory rather than confirmatory spatial analysis (Tukey, 1977; Openshaw and Openshaw, 1997; Anselin, 1998) which assumes a huge amount of spatial data.

In section 2 we formulate the problem to solve in this paper, and discuss the methodology of tessellation modeling. From sections 3 to 5 we propose three methods for modeling a spatial tessellation by a set of other spatial tessellations successively. To test the validity of the methods we perform an empirical study in section 6. Section 7 summarizes the conclusions with discussion.

2 Methodology

Suppose a region S which is divided into a set of subregions, say, school districts or census tracts, by a categorical variable. 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. The regions by definition satisfy

$$\bigcap_i y_i = \varnothing \tag{1}$$

and

$$\bigcup_i y_i = S. \tag{2}$$

To represent the spatial structure of the dependent tessellation, we use a tessellation indicator function

$$i(\mathbf{u}; Y) = y_i. \tag{3}$$

Equation (3) indicates that the point at **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. The tessellation defined by an independent variable is called the *independent tessellation*. Independent tessellations are also represented by tessellation indicator functions:

$$i(\mathbf{u}; X_i) = x_{ij}. \tag{4}$$

Our objective is to build a model representing the dependent tessellation Y by the set of independent tessellations $X=\{X_1, X_2, ..., X_m\}$. There are at least two possible approaches to the modeling, whose choice depends on the circumstances. Suppose a dependent tessellation Y shown in figure 2a. If we have a set of independent tessellations $X=\{X_1, X_2, X_3\}$ in figure 2b, we can obtain Y by overlaying X_1, X_2, and X_3. We call this the *overlay approach*. We next consider another set of independent tessellations $X'=\{X_1', X_2', X_3'\}$ shown in figure 2c. We can also obtain Y by combining the gray-shaded regions in X' as shown in figure 2d. We call this the *combination approach*. The overlay approach yields a tessellation by overlaying a set of tessellations while the combination approach generates a tessellation by combining small pieces of tessellations.

Figure 2. An example of the dependent and independent tessellations. (a) The dependent tessellation *Y*, (b) a set of independent tessellations *X*, (c) another set of independent tessellations *X'*, (d) regions composing *Y*.

As seen in figure 2, the choice of the method depends on the dependent and independent tessellations; the approach that gives a better explanation of the dependent tessellation should be chosen. In this paper, we follow the combination approach and leave the overlay approach for future research, because the combination approach is appropriate for our empirical study

described in section 6. In the following we propose three methods of tessellation modeling: the region-based method, boundary-based method, and hybrid method. 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.

3 Region-based method

3.1 Outline

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. Consider, for instance, an example shown in figure 3. The independent tessellations X_1, X_2, and X_3 are different from the dependent tessellation *Y*. However, for some regions, they completely agree with *Y*; X_1, X_2, and X_3 agree with *Y* with respect to the gray-shaded regions. The dependent tessellation *Y*, therefore, can be modeled by a spatial combination of X_1, X_2, and X_3. This is the basic idea of the region-based method; it decomposes the region *S* into subregions and models them separately by independent tessellations.

Figure 3. 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*.

Figure 4 illustrates the above algorithm. As shown in figure 4, the region-based method successively tries the independent tessellations that give the best agreement with the dependent tessellation and finally yields a model representing the dependent tessellation.

Figure 4. The region-based method.

3.2 Evaluation of the agreement between dependent and independent tessellations

The region-based method evaluates the agreement between the dependent tessellation *Y* and the independent tessellation X_i (step 2.1). To this end we propose a measure which we call the agreement index.

Suppose a pairwise indicator function defined by

$$1(\mathbf{u},\mathbf{v};T)=\begin{cases}1 & \text{if } i(\mathbf{u};T)=i(\mathbf{v};T)\\ 0 & \text{otherwise}\end{cases}, \tag{5}$$

where *T* is a spatial tessellation of the region *S*. If two locations belong to the same region in *T*, the function is equal to one. Otherwise, it becomes zero.

Using the pairwise indicator function, we describe the spatial autocorrelation of the tessellation *T* by a function of distance *d* between two locations:

$$\gamma(d;T)=\frac{\int_{\mathbf{u}\in S}\int_{\mathbf{v}\in S,|\mathbf{u}-\mathbf{v}|=d}1(\mathbf{u},\mathbf{v};T)\mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u}}{\int_{\mathbf{u}\in S}\int_{\mathbf{v}\in S,|\mathbf{u}-\mathbf{v}|=d}\mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u}}. \tag{6}$$

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. Figure 5 shows a typical example of the spatial autocorrelation function.

Figure 5. An example of the spatial autocorrelation function.

We then consider a hypothetical trial of guessing the value of the pairwise indicator function for a given pair of locations. Given two locations **u** and **v** in *S* and the spatial autocorrelation function $\gamma(d; Y)$, we guess the value of the pairwise indicator function $1(\mathbf{u}, \mathbf{v}; Y)$. A simple strategy is to give one or zero following their occurrence probabilities represented by $\gamma(|\mathbf{u}\text{-}\mathbf{v}|; Y)$. We choose one with a probability of $\gamma(|\mathbf{u}\text{-}\mathbf{v}|; Y)$, and zero with a probability of $1\text{-}\gamma(|\mathbf{u}\text{-}\mathbf{v}|; Y)$. If $1(\mathbf{u}, \mathbf{v}; Y)=1$, we give the right answer with a probability of $\gamma(|\mathbf{u}\text{-}\mathbf{v}|; Y)$. Otherwise, we are correct with a probability of $1\text{-}\gamma(|\mathbf{u}\text{-}\mathbf{v}|; Y)$. The probability of giving the right answer reflects the degree of difficulty of the trial.

We apply the above stochastic trial to the evaluation of agreement between Y and X_i. We regard the process of fitting X_i to Y as a set of the above trials, that is, X_i as a set of answers for trials, and evaluate their agreement by the total score, taking into account the degree of difficulty of individual trials. If $1(\mathbf{u}, \mathbf{v}; Y)=1$ and $\gamma(|\mathbf{u}\text{-}\mathbf{v}|; Y)$ is small, it is highly evaluated to give the right answer. If the answer is wrong, a light penalty is imposed. On the other hand, if $1(\mathbf{u}, \mathbf{v}; Y)=1$ and $\gamma(|\mathbf{u}\text{-}\mathbf{v}|; Y)$ is large, it is easy to give the right answer and thus its evaluation is low. To formalize this evaluation system, we consider a pairwise evaluation function

$$e(\mathbf{u},\mathbf{v};X_i,Y)=\begin{cases}1-\gamma(d;Y) & \text{if } 1(\mathbf{u},\mathbf{v};Y)=1 \text{ and } 1(\mathbf{u},\mathbf{v};X_i)=1\\ -\gamma(d;Y) & \text{if } 1(\mathbf{u},\mathbf{v};Y)=1 \text{ and } 1(\mathbf{u},\mathbf{v};X_i)=0\\ \gamma(d;Y)-1 & \text{if } 1(\mathbf{u},\mathbf{v};Y)=0 \text{ and } 1(\mathbf{u},\mathbf{v};X_i)=1\\ \gamma(d;Y) & \text{if } 1(\mathbf{u},\mathbf{v};Y)=0 \text{ and } 1(\mathbf{u},\mathbf{v};X_i)=0\end{cases}, \tag{7}$$

and integrate it for all $\mathbf{u}$ and $\mathbf{v}$ in S to evaluate the agreement between X_i and Y:

$$A(X_i;Y)=\int_{\mathbf{u}\in S}\int_{\mathbf{v}\in S} e(\mathbf{u},\mathbf{v};X_i,Y)\mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u}. \tag{8}$$

This function shows a large value if the independent tessellation X_i agrees well with the dependent tessellation Y. In the following we use it in its standardized form

$$\alpha(X_i;Y)=\frac{A(X_i;Y)-\int_{\mathbf{u}\in S}\int_{\mathbf{v}\in S}\min\{-\gamma(|\mathbf{u}-\mathbf{v}|;Y),\gamma(|\mathbf{u}-\mathbf{v}|;Y)-1\}\mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u}}{\int_{\mathbf{u}\in S}\int_{\mathbf{v}\in S}\max\{1-\gamma(|\mathbf{u}-\mathbf{v}|;Y),\gamma(|\mathbf{u}-\mathbf{v}|;Y)\}\mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u}-\int_{\mathbf{u}\in S}\int_{\mathbf{v}\in S}\min\{-\gamma(|\mathbf{u}-\mathbf{v}|;Y),\gamma(|\mathbf{u}-\mathbf{v}|;Y)-1\}\mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u}}, \tag{9}$$

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

3.3 Evaluation of the fitness of a tessellation for a region in a different tessellation

Another evaluation step in the region-based method is to measure the fitness of X_i for y_j (step 2.3.1). This evaluation is similar to that of the agreement between tessellations. It is also based on the pairwise evaluation function defined by equation (7); the difference lies in the domain of integration. The *fitness index* of X_i with respect to y_j is defined by

$$\beta(X_i;y_j)=\frac{\int_{\mathbf{u}\in y_j}\int_{\mathbf{v}\in S} e(\mathbf{u},\mathbf{v};X_i,Y)\mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u}-\int_{\mathbf{u}\in y_j}\int_{\mathbf{v}\in S}\min\{-\gamma(|\mathbf{u}-\mathbf{v}|;Y),\gamma(|\mathbf{u}-\mathbf{v}|;Y)-1\}\mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u}}{\int_{\mathbf{u}\in y_j}\int_{\mathbf{v}\in S}\max\{1-\gamma(|\mathbf{u}-\mathbf{v}|;Y),\gamma(|\mathbf{u}-\mathbf{v}|;Y)\}\mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u}-\int_{\mathbf{u}\in y_j}\int_{\mathbf{v}\in S}\min\{-\gamma(|\mathbf{u}-\mathbf{v}|;Y),\gamma(|\mathbf{u}-\mathbf{v}|;Y)-1\}\mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u}}. \tag{10}$$

Significance of the fitness, which appears in step 2.3.2, 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

$$\beta(X_i;y_j)\ge\beta_T. \tag{11}$$

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.

3.4 Numerical variables

It often happens that, in addition to categorical variables forming tessellations, numerical variables are available over the same region and seem influential on the dependent tessellation. It is desirable to take numerical variables into account in tessellation modeling.

Suppose a scalar function $f(\mathbf{x})$ representing a numerical variable defined over the region S. Let $C=\{c_1, c_2, ..., c_k\}$ be a set of values satisfying $\min\{f(\mathbf{x})\}\le c_1\le c_2\le ...c_k\le\max\{f(\mathbf{x})\}$. This is the set of boundary values used for categorizing the numerical variable into k+1 classes. They yield the tessellation $F=\{f_1, f_2, ..., f_{k+1}\}$ where

$$f_i = \{\mathbf{x}, c_{i-1} \le f(\mathbf{x}) < c_i\}. \tag{12}$$

The pairwise indicator function is given by

$$1(\mathbf{u},\mathbf{v}; f) = \begin{cases} 1 & \text{if } \exists i, \mathbf{u},\mathbf{v} \in f_i \\ 0 & \text{otherwise} \end{cases}. \tag{13}$$

The pairwise evaluation function and the agreement and fitness indices are defined in a way similar to those of categorical variables.

The boundary set C is determined so that it gives the best agreement between the numerical and dependent variables in terms of their spatial tessellations. Mathematically, we solve

$$\max_B \alpha(F;Y). \tag{14}$$

The number of boundaries k is optimized if we use a large initial value for k; redundancy in the set C emerges as a repetition of the same value, say, $c_1=c_2=c_3$.

4 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.

The scale of analysis is also somewhat different between the two methods. The region-based method has a global point of view on tessellations. Its agreement index evaluates all the point pairs in S in that whether or not they are equivalent with respect to the pairwise indicator function, even if they are located at a great distance. The boundary-based method, on the other hand, by its nature, looks at tessellations locally because the division process is usually a local phenomenon. As seen later, it considers only pairs of adjacent regions in Y in agreement evaluation.

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 third row of 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.

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:

$$1\left(y_i, y_j\right)=\begin{cases}1 & \text{if } b(Y)_{ij} \text{ exists}\\ 0 & \text{otherwise}\end{cases}. \tag{15}$$

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 boundary indicator function of X_i is defined by

$$1\left(x_{ij}, x_{ik}\right)=\begin{cases}1 & \text{if } b\left(X_i\right)_{jk} \text{ exists}\\ 0 & \text{otherwise}\end{cases}. \tag{16}$$

Each tessellation is represented by either its composing regions or boundaries hereafter; the dependent tessellation, for instance, is represented by either Y or $B(Y)$. The algorithm of the boundary-based method is as follows:

Algorithm: Boundary-based method

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

Output: A set of independent tessellations $B(U)$ and a set of boundaries $B(V)$ modeled by $B(U)$.

Step 1: Set $B(U)$ and $B(V)$ empty.

Step 2: Do 2.1-2.5 while neither $B(X)$ nor $B(Y)$ is an empty set.

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

Step 2.2: Choose the independent tessellation $B(X_i)$ that gives the best agreement.

Step 2.3: Do 2.3.1 and 2.3.2 for all $b(Y)_{jk}\in B(Y)$.

2.3.1: Evaluate the fitness of $B(X_i)$ for $b(Y)_{jk}$.

2.3.2: If the fitness is significant, add $b(Y)_{jk}$ to $B(V)$ and remove $b(Y)_{jk}$ from $B(Y)$.

Step 2.4: If any $b(Y)_{jk}\in B(Y)$ was added to $B(V)$, add $B(X_i)$ to $B(U)$.

Step 2.5: Remove $B(X_i)$ from $B(X)$.

Step 3: Report $B(U)$ and $B(V)$.

The algorithm is substantially the same as that of the region-based method, except it evaluates the agreement between boundaries. Figure 7 illustrates the above algorithm.

Figure 7. The boundary-based method

The agreement between the independent tessellation $B(X_i)$ and the dependent tessellation $B(Y)$ is measured by

$$A'\left(B(X_i);B(Y)\right)=\sum_{j,k,1(y_j,y_k)=1}\int_{\mathbf{u}\in y_j\cup y_k}\int_{\mathbf{v}\in y_j\cup y_k}e(\mathbf{u},\mathbf{v};X_i,Y)\mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u} \tag{17}$$

and its standardized form

$$\alpha'\left(B(X_i);B(Y)\right)=\frac{A'(X_i;Y)-\sum_{j,k,1(y_j,y_k)=1}\int_{\mathbf{u}\in y_j\cup y_k}\int_{\mathbf{v}\in y_j\cup y_k}\min\{-\gamma(|\mathbf{u}-\mathbf{v}|;Y),\gamma(|\mathbf{u}-\mathbf{v}|;Y)-1\}\mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u}}{\sum_{j,k,1(y_j,y_k)=1}\left[\int_{\mathbf{u}\in y_j\cup y_k}\int_{\mathbf{v}\in y_j\cup y_k}\max\{1-\gamma(|\mathbf{u}-\mathbf{v}|;Y),\gamma(|\mathbf{u}-\mathbf{v}|;Y)\}\mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u}-\int_{\mathbf{u}\in y_j\cup y_k}\int_{\mathbf{v}\in y_j\cup y_k}\min\{-\gamma(|\mathbf{u}-\mathbf{v}|;Y),\gamma(|\mathbf{u}-\mathbf{v}|;Y)-1\}\mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u}\right]}. \tag{18}$$

The difference between equations (8) and (17) lies in the domain of integration; the former integrates the pairwise evaluation function for all **u** and **v** in S while the latter considers only the point pairs in regions adjacent in Y. This reflects the different views of the methods mentioned earlier.

The fitness of $B(X_i)$ for $b(Y)_{jk}$ (step 2.3.1) is evaluated by

$$\beta\left(B(X_i);b(Y)_{jk}\right)=\frac{\int_{\mathbf{u}\in y_j\cup y_k}\int_{\mathbf{v}\in y_j\cup y_k}e(\mathbf{u},\mathbf{v};X_i,Y)\mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u}-\int_{\mathbf{u}\in y_j\cup y_k}\int_{\mathbf{v}\in y_j\cup y_k}\min\{-\gamma(|\mathbf{u}-\mathbf{v}|;Y),\gamma(|\mathbf{u}-\mathbf{v}|;Y)-1\}\mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u}}{\int\int\max\{1-\gamma(|\mathbf{u}-\mathbf{v}|;Y),\gamma(|\mathbf{u}-\mathbf{v}|;Y)\}\mathrm{d}\mathbf{v}\mathrm{d}\mathbf{u}-\int\int\min\{-\gamma(|\mathbf{u}-\mathbf{v}|;Y),\gamma(|\mathbf{u}-\mathbf{v}|;Y)\}}. \tag{19}$$

Its significance is again judged by the threshold value β_T' given by the analyst. Numerical variables can also be included in analysis as well as in the region-based method.

5 Hybrid method

The region-based and boundary-based methods are dual in the sense that regions are defined by boundaries while boundaries are given by regions. However, they are different in scale of analysis as mentioned earlier; the region-based method treats tessellations globally while the boundary-based method focuses on local heterogeneity in variables.

The difference in scale naturally leads to the idea that they can complement with each other. Since the region-based method is suitable for global analysis, it should be performed first and then followed by the boundary-based method. The boundary-based method is applied to a part of the original dependent tessellation that is not modeled by the region-based method. We call this the *hybrid method*.

The input of the hybrid method is a dependent tessellation and a set of independent tessellations represented by both regions and boundaries. Its output is the set of regions and boundaries modeled by the two methods, and the set of independent tessellations used for

modeling.

6 Empirical study

This section describes an empirical study of modeling a tessellation by a set of other tessellations, in order to test the validity of the method proposed in the previous sections. The studied area is Ponneri, located to the north of Madras, India, in the late eighteenth century (figure 8).

Figure 8. Ponneri, India.

The data sources 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 (figure 8), whose location was digitized into GIS by ArcInfo ver. 7.2.1. 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.

The objective of analysis is to explain the spatial structure of administrative system called the *zamindari* system in the late eighteenth century. 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.

Figure 9 shows the two administrative systems, the magan and zamindari systems, by Voronoi diagrams. Since the map of village boundary was not available, it was approximated by the Voronoi diagram in which villages were used as generators.

Figure 9. 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. Their spatial structures are partly similar but different in some places. 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 proposed in the previous sections. We have developed it to extract possible influential factors among numerous variables, to evaluate them in terms of the agreement with the zamindari system, and finally to build a model representing the zamindari system, in an effective and objective way. 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. That is why we chose the combination approach discussed in section 2.

The dependent tessellation is, therefore, the zamindari system represented by a set of Voronoi regions as shown in figure 9b. From attribute data of villages we chose fifteen variables as independent variables (table 1). They were also transformed into Voronoi diagrams, the independent tessellations.

Table 1. 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 2, 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 10 shows a set of regions *V* explained by the magan system.

Table 2. Modeling the zamindari system by the region-based method.

Figure 10. 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 thus applied the boundary-based method to the same data, with the threshold value β_T' 0.99. The result is shown in table 3.

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

As shown in table 3, the boundary-based method also detected the magan system as the most influential variable. We then removed the regions explained by the magan system and investigated the other independent variables in turn. The next variable we obtained was the forest ratio. After removing the regions explained by the forest ratio, we could not obtain further influential variables. Figure 11 shows the boundaries explained by the magan system and the forest ratio.

Figure 11. Zamindari boundaries modeled by the boundary-based method.

Figure 11 is more satisfactory than figure 10, because 69.89 % of the boundaries are explained by the two variables. However, there still remains 30.11 % boundaries not explained by the independent variables. We thus finally applied the hybrid method and obtained the result shown in table 4 and figure 12.

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

Figure 12. Zamindari boundaries modeled by the hybrid method.

Combination of the two methods yielded better result than that given by 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.

7 Conclusion

In this paper we have developed a method for modeling a spatial tessellation by a set of other tessellations, motivated by an experience of joint study with researchers in Islamic Area Studies. We proposed three methods, that is, the region-based method, boundary-based method, and hybrid method, all designed for exploratory spatial analysis rather than confirmatory analysis. The region-based method focuses on the regions that compose the dependent tessellation while the boundary-based method emphasizes the boundaries between

regions rather than the regions themselves. The hybrid method, a combination of the two methods, inherits strengths from both of them. In these methods the agreement between tessellations is measured by the agreement indices that suppose a hypothetical stochastic process of fitting an independent tessellation to the dependent tessellation. Since the indices explicitly take into account the spatial structure of the dependent tessellation, they can distinguish the three tessellations shown in figure 1b from that in figure 1a. To test the validity of the method, we analyzed the administrative system called the zamindari system in Ponneri, India, in the late eighteenth century. The empirical study yielded some interesting findings that help us understand the spatial structure of zamindari system.

We finally discuss some limitations of our method for further research. First, as discussed in section 2, the method proposed treats only one aspect of tessellation formation - it implicitly assumes that the dependent tessellation is spatially decomposable, that is, the tessellation is obtained by combining small pieces of tessellations. The combination approach, however, does not always work successfully because there are cases where its underlying assumption does not hold; tessellations generated by overlay of multiple tessellations. To deal with such cases, it is important to pursue the overlay approach, to develop a method for modeling a tessellation by overlay of independent tessellations. In addition, it is also desirable to combine the two approaches to extend their applicability in exploratory spatial analysis. Second, 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. A more flexible model that treats a wide variety of spatial objects is necessary. Third, benchmark values for the thresholds β_T and β_T' used in fitness evaluation should be discussed further. Though the thresholds can be determined arbitrarily, it is convenient if some benchmark values are presented. To this end, more empirical studies have to be performed.

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* Eds P A Longley, S M Brooks, R McDonnell, B Macmillan (Wiley, Chichester) 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, Lerman S, 1985 *Discrete Choice Analysis: Theory and Application to Travel Demand* (MIT Press, Massachusetts)

Cartersen L W Jr., 1987 "A measure of similarity for cellular maps" *American Geographer* **14** 345-358

Cohen J, 1960 "A coefficient of agreement for nominal scales" *Educational and Psychological Measurement* **20** 37-46

Cressie N, 1993 *Statistics for Spatial Data*. (Wiley, New York)

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

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

Landis J R, Koch G C 1977 "The measurement of observer agreement for categorical data" *Biometrics* **33** 159-174

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

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.

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

Openshaw S, Charlton M, Wymer C, Craft A, 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, Openshaw C, 1997 *Artificial Intelligence in Geography*. (Wiley, Chichester)

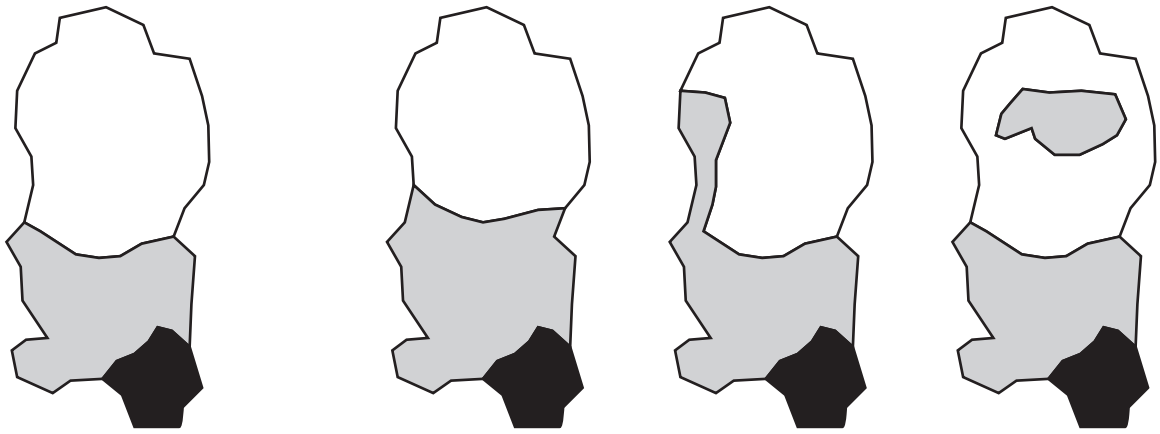
Pontius R G Jr., 2000 "Quantification error versus location error in comparison of categorical maps" *Photogrammetric Engineering and Remote Sensing* **66** 1011-1016

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

Sadahiro Y, 1999 "Accuracy of areal interpolation: a comparison of alternative methods" *Journal of Geographical Systems* **1** 323-346

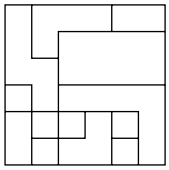
Tukey J W, 1977 *Exploratory Data Analysis* (Addison-Wesley, New York)

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



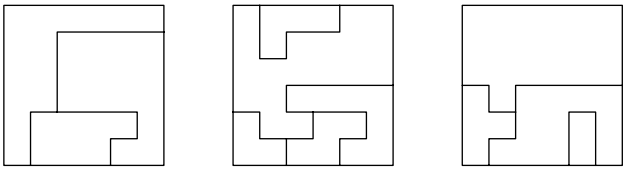
(a) (b)

Figure 1



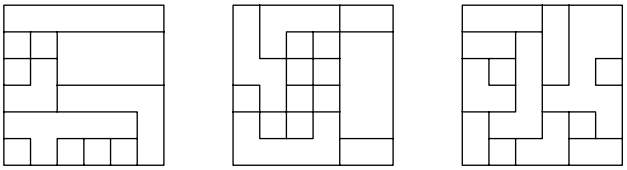
Y

(a)



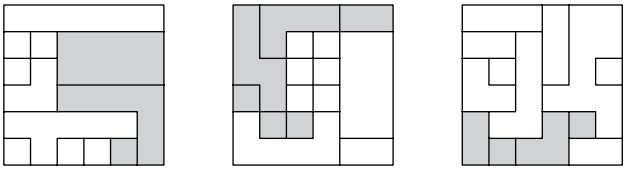
X_1 X_2 X_3

(b)



X_1' X_2' X_3'

(c)



X_1' X_2' X_3'

(d)

Figure 2

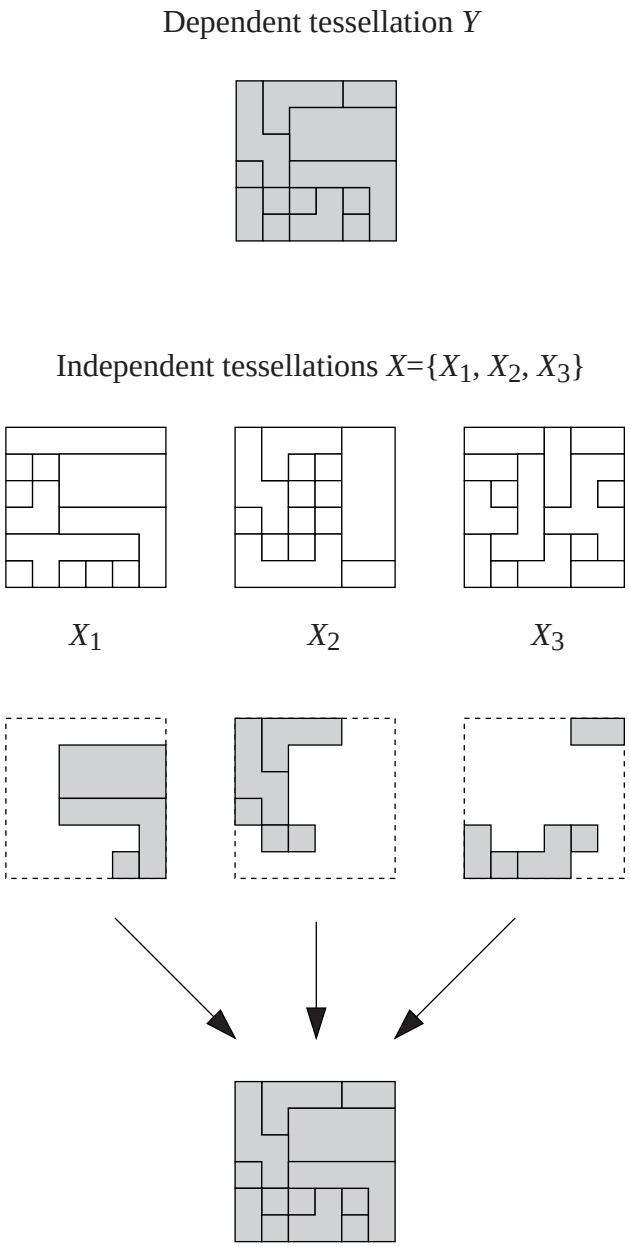

Dependent tessellation Y
Independent tessellations $X=\{X_1, X_2, X_3\}$
X_1
X_2
X_3

Figure 3

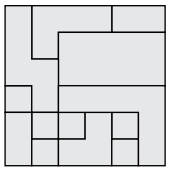
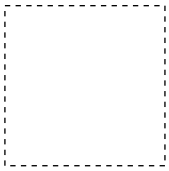
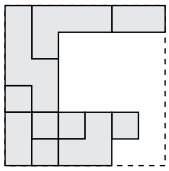
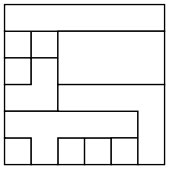
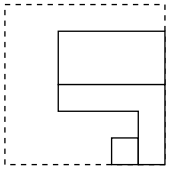
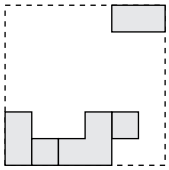
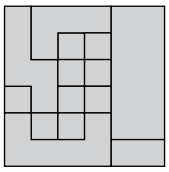
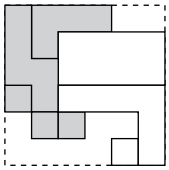
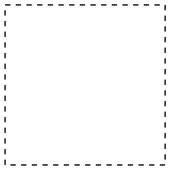
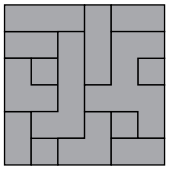
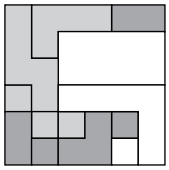
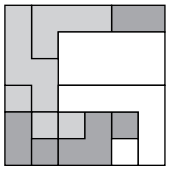
	Y	X	X_i that gives the best agreement with Y	V	U
Step 1		$\{X_1, X_2, X_3, X_4\}$			$\varnothing$
Step 2		$\{X_2, X_3, X_4\}$	 X_1		$\{X_1\}$
		$\{X_3, X_4\}$	 X_2		$\{X_1, X_2\}$
		$\{X_4\}$	 X_3		$\{X_1, X_2, X_3\}$
Step 3					$\{X_1, X_2, X_3\}$

Figure 4

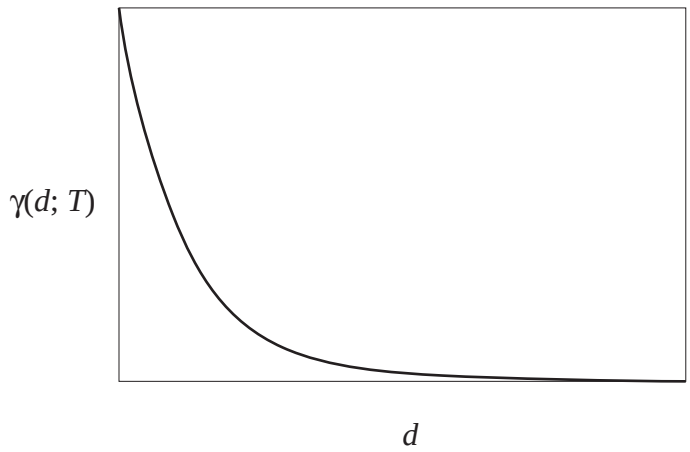
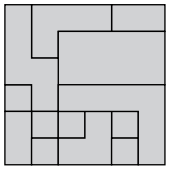


Figure 5

Dependent tessellation Y



Independent tessellations $X=\{X_1, X_2, X_3\}$

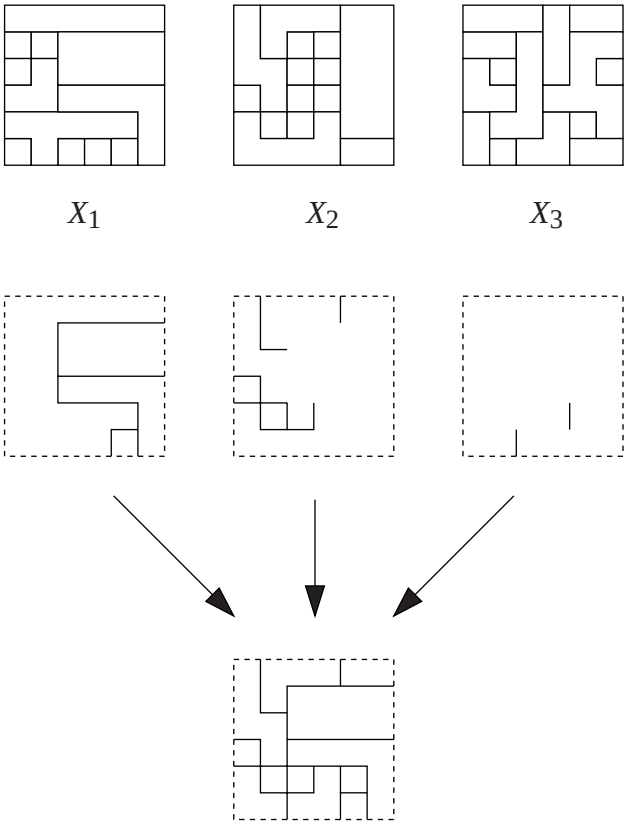


Figure 6

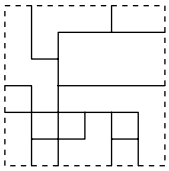
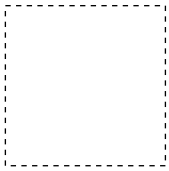
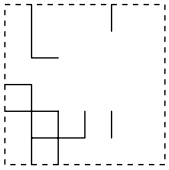
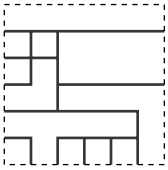
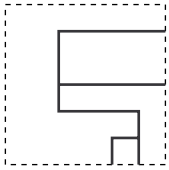
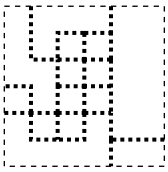
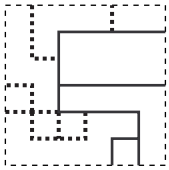
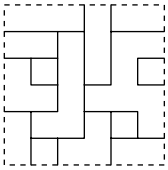
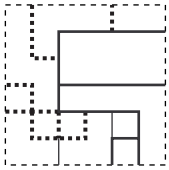
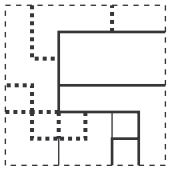
	$B(Y)$	$B(X)$	$B(X_i)$ that gives the best agreement with $B(Y)$	$B(V)$	$B(U)$
Step 1		$\{B(X_1), B(X_2), B(X_3), B(X_4)\}$			$\varnothing$
Step 2		$\{B(X_2), B(X_3), B(X_4)\}$	 $B(X_1)$		$\{B(X_1)\}$
		$\{B(X_3), B(X_4)\}$	 $B(X_2)$		$\{B(X_1), B(X_2)\}$
		$\{B(X_4)\}$	 $B(X_3)$		$\{B(X_1), B(X_2), B(X_3)\}$
Step 3					$\{B(X_1), B(X_2), B(X_3)\}$

Figure 7

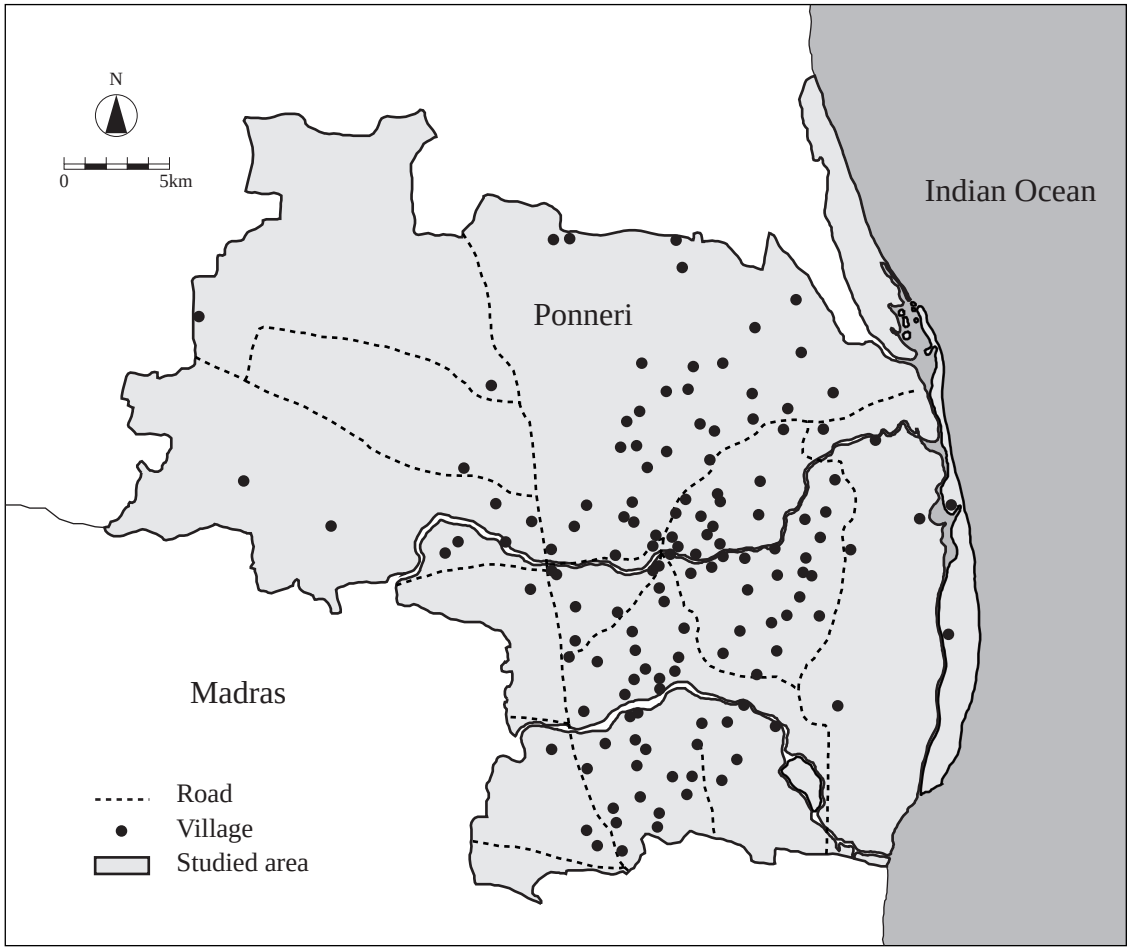
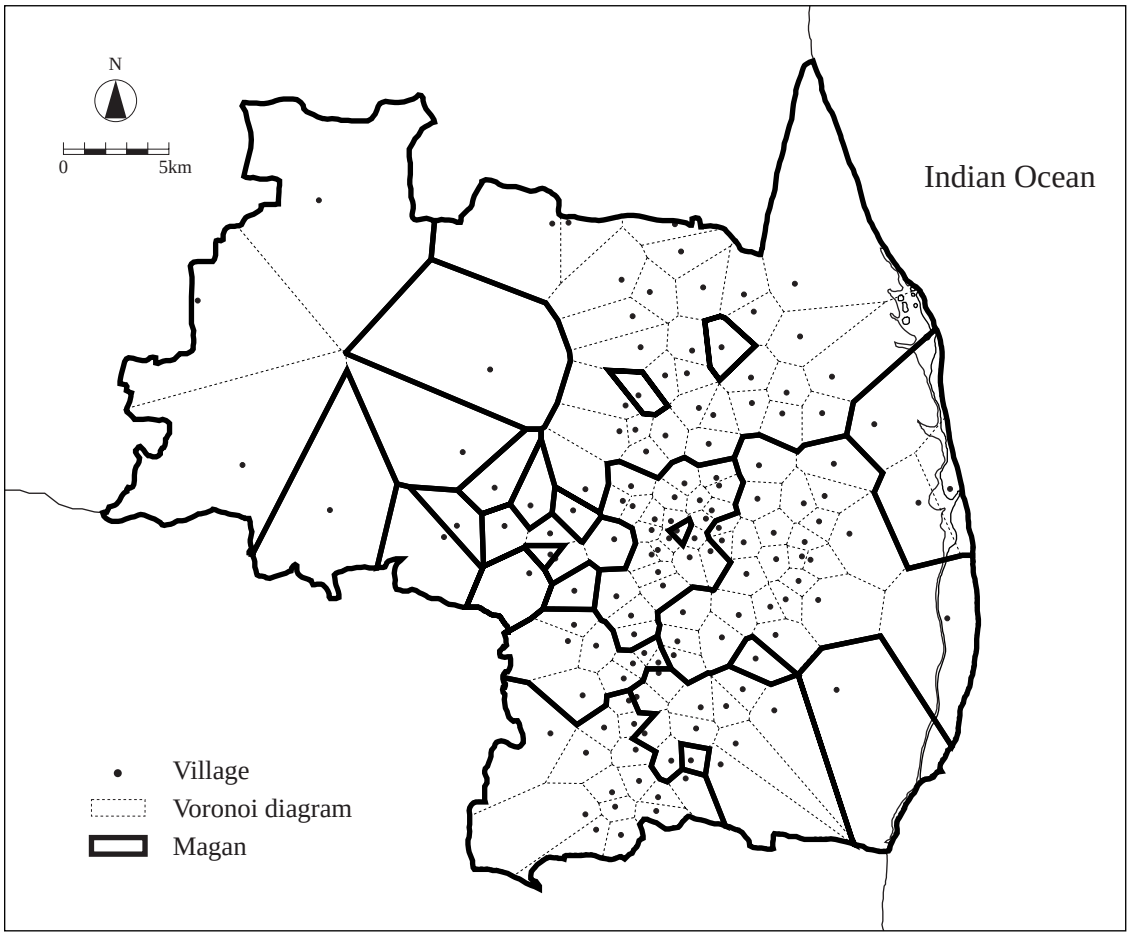
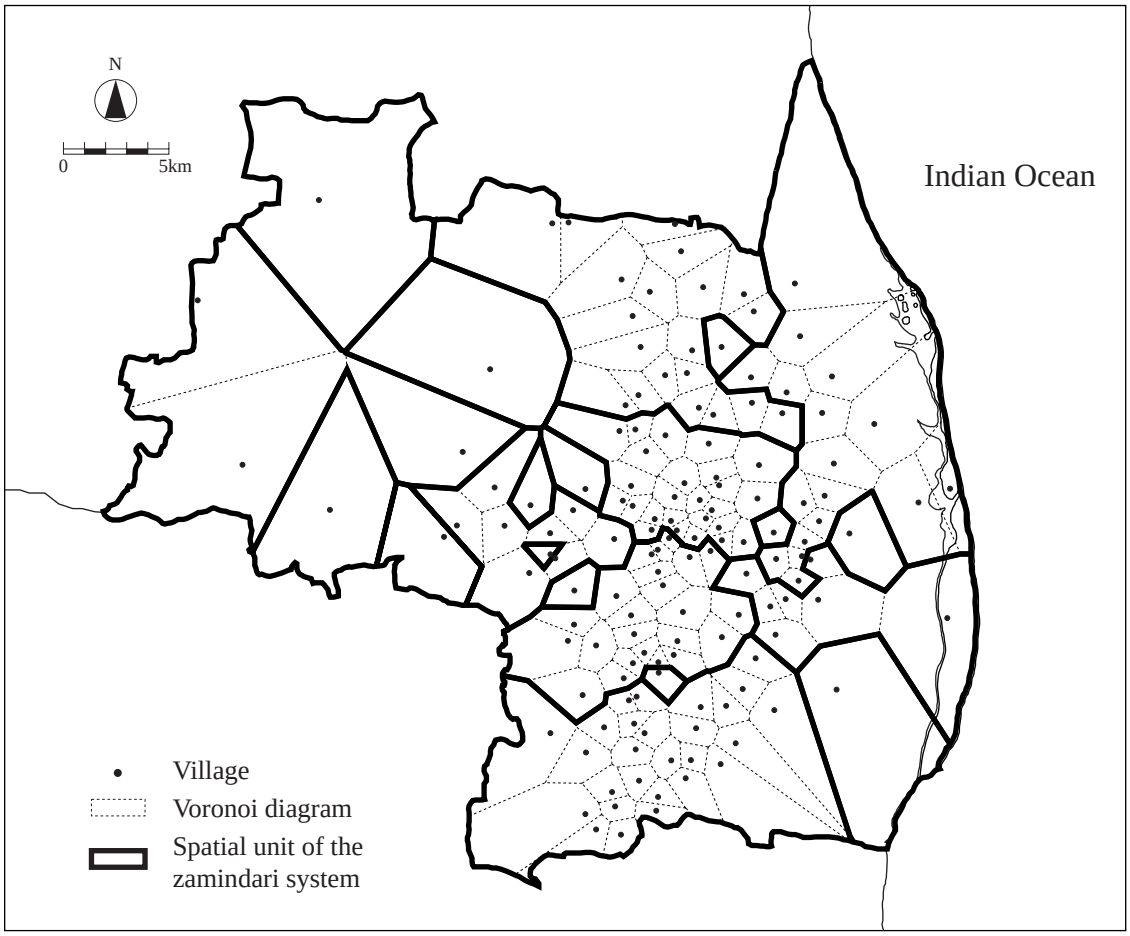

N
0
5km
Indian Ocean
Ponneri
Madras
Road
Village
Studied area

Figure 8


N
0
5km
Indian Ocean
Village
Voronoi diagram
Magan

(a)


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

(b)

Figure 9

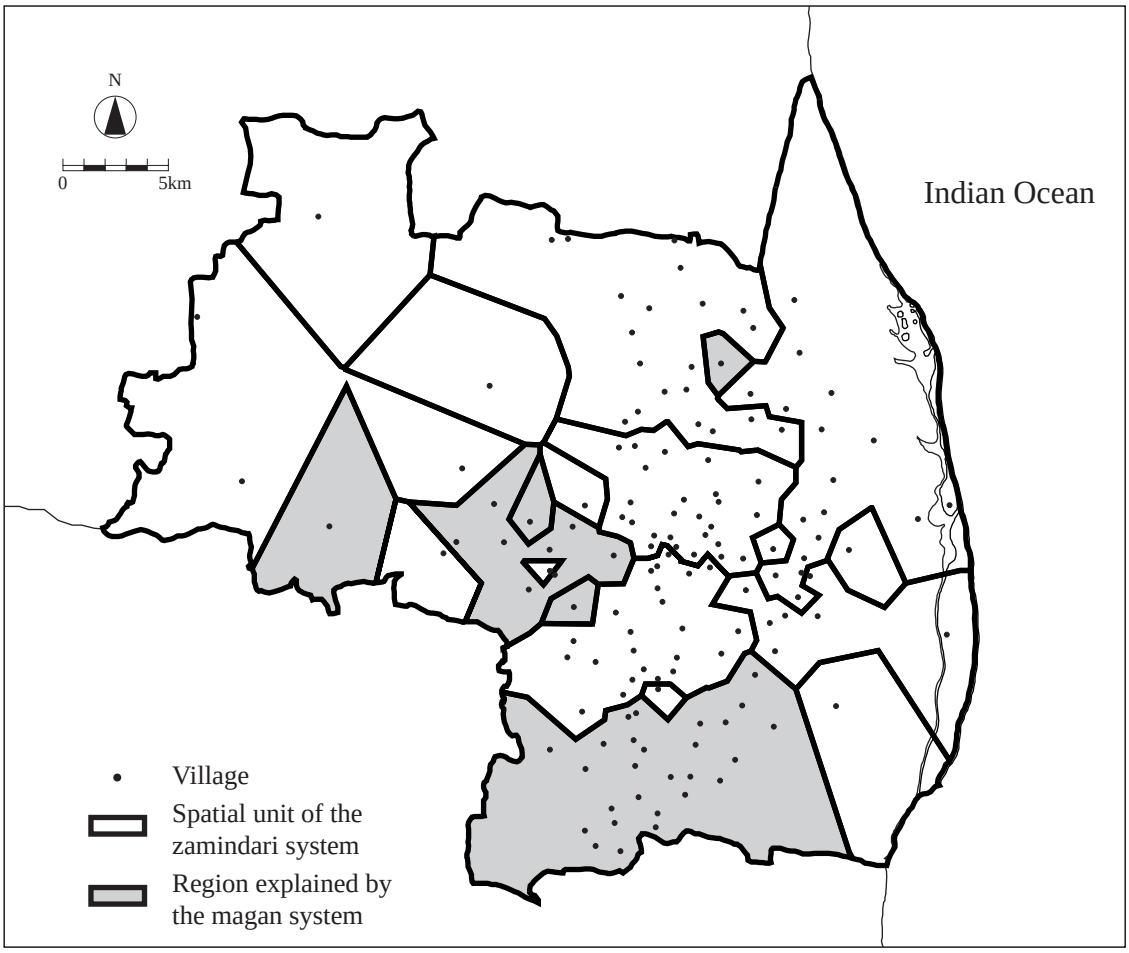

N
0
5km
Indian Ocean
Village
Spatial unit of the zamindari system
Region explained by the magan system

Figure 10

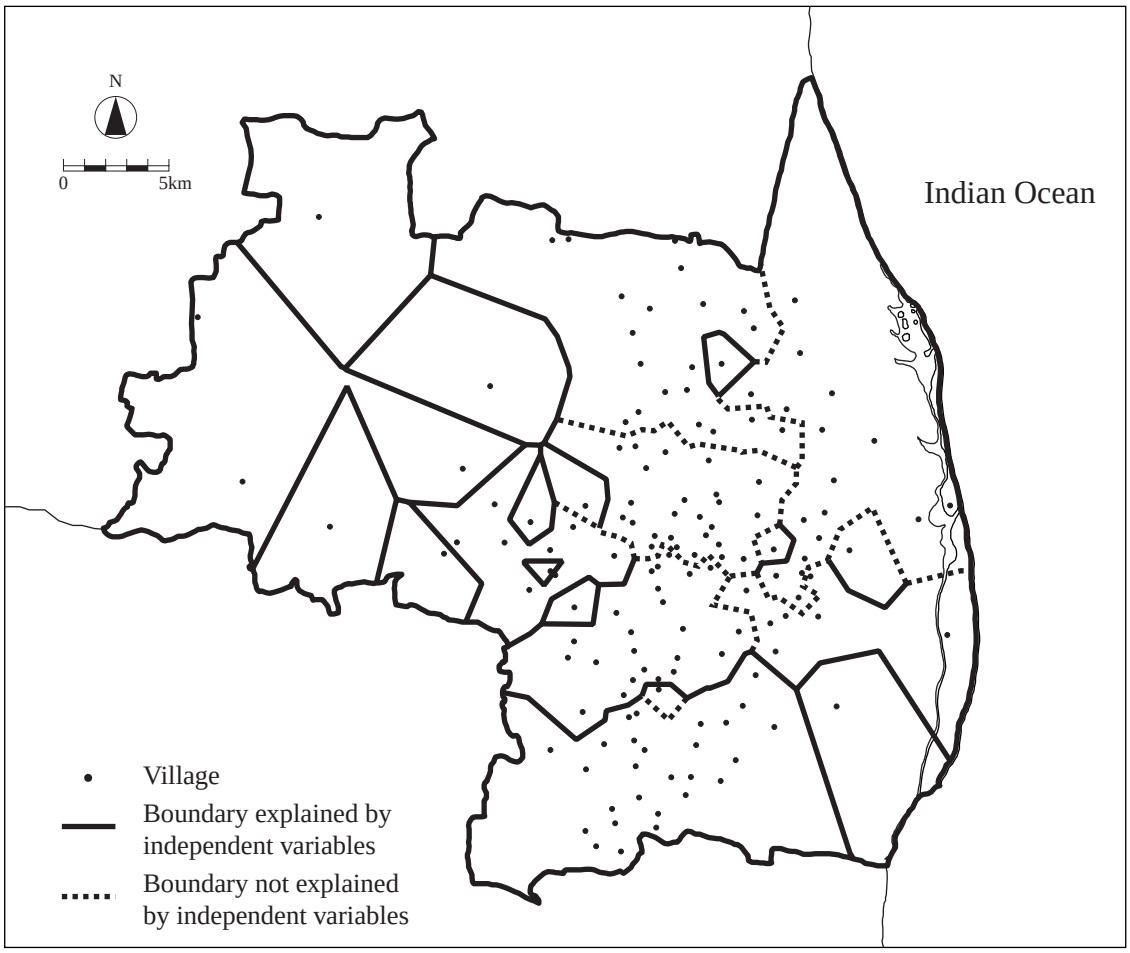

N
0
5km
Indian Ocean
Village
Boundary explained by independent variables
Boundary not explained by independent variables

Figure 11

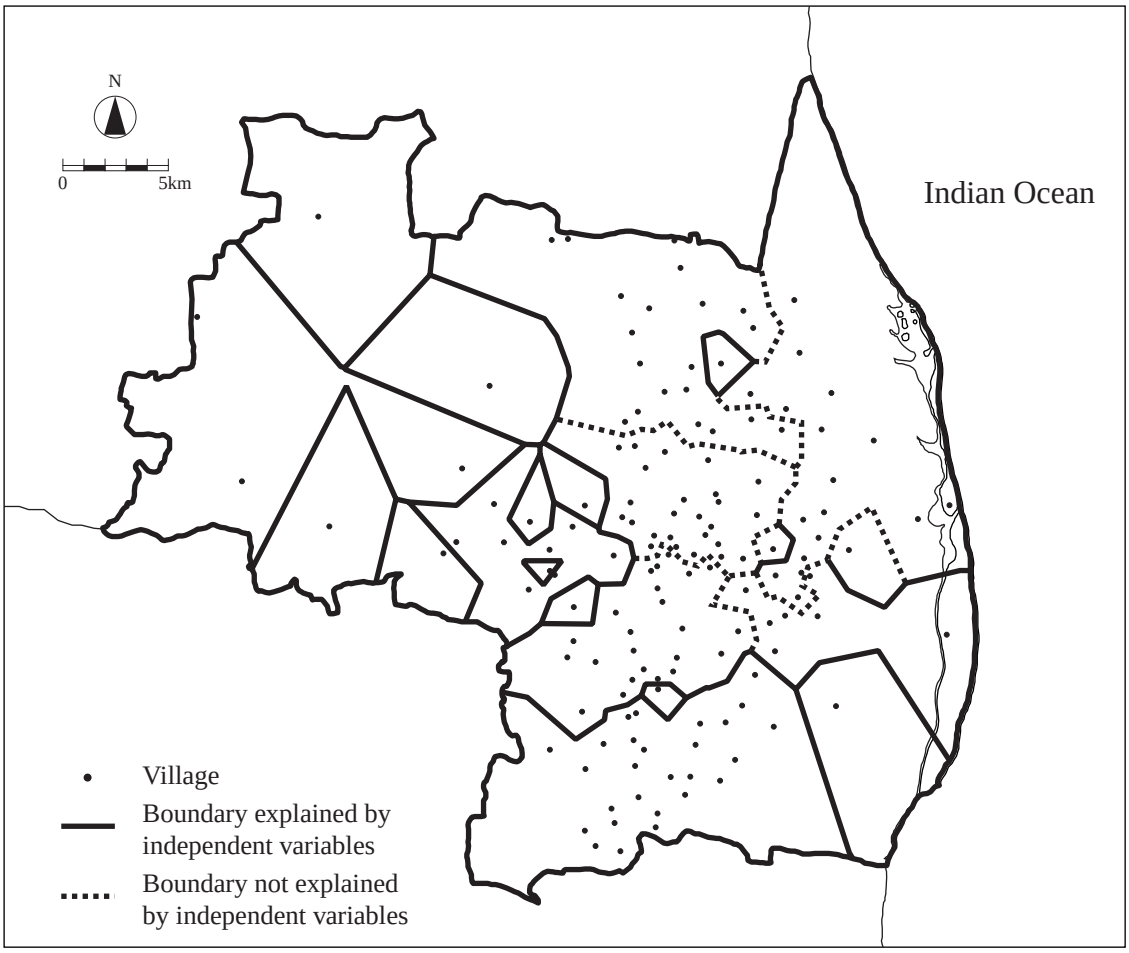

N
0
5km
Indian Ocean
Village
Boundary explained by independent variables
Boundary not explained by independent variables

Figure 12

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 1

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 2

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

Magan	0.9131
Poligar	0.8707
Caste composition	0.8442
Dominant caste	0.8570
Brahman	0.7608
Crop type	0.7125
Population	0.9010
Area	0.8984
Caste homogeneity	0.8777
Irrigated farmland	0.8969
Wasteland	0.8945
Forest	0.9006
State-owned land	0.8941
Hoe	0.8037
Tree	0.8034

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

Forest	0.9087
Population	0.8938
Dominant caste	0.8913
Tree	0.8909
Caste composition	0.8908
Hoe	0.8867
State-owned land	0.8859
.	.
.	.
.	.

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