CSIS Discussion Paper No. **136R**

Configuration of sample points for the reduction of multicollinearity in regression models with distance variables

Yukio Sadahiro* and Yan Wang**

June 2015

*Center for Spatial Information Science, The University of Tokyo
**Department of Urban Engineering, The University of Tokyo

Center for Spatial Information Science, The University of Tokyo
5-1-5, Kashiwanoha, Kashiwa-shi, Chiba 277-8568, Japan
E-mail: sada@csis.u-tokyo.ac.jp

Abstract

Keywords: sample points, multicollinearity, distance variables, regression models

Regression models often suffer from multicollinearity that greatly reduces the reliability of estimated coefficients and hinders an appropriate understanding of the role of independent variables. It occurs in regional science especially when independent variables include the distances from urban facilities. This paper proposes a new method for deriving the configuration of sample points that reduces multicollinearity in regression models with distance variables. Multicollinearity is evaluated by the maximum absolute correlation coefficient between distance variables. A spatial optimization technique is utilized to calculate the optimal configuration of sample points. The method permits us not only to locate sample points appropriately but also to evaluate the location of facilities from which the distance is measured in terms of the correlation between distance variables in a systematic way. Numerical experiments and empirical applications are performed to test the validity of the method. The results support the technical soundness of the proposed method, and provided some useful implications for the design of sample location.

1. Introduction

Regression models often suffer from multicollinearity, a state of high correlation between two or more independent variables. Multicollinearity greatly reduces the reliability of estimated coefficients, and consequently, hinders an appropriate understanding of the role of independent variables. Spatial analysis often utilizes the distances from urban facilities as independent variables that can be highly correlated with each other. Hedonic housing price models consider the distance from housing units to CBD, schools, urban parks, and hazardous waste sites to evaluate the spatial environment around each house (Berry (1976); Li and Brown (1980); Bender and Hwang (1985); Harrison Jr and Rubinfeld (1978); Ihlanfeldt and Taylor (2004); Noonan, Krupka, and Baden (2007)). A close relationship between obesity and neighborhood shopping environment is represented as a regression model that utilizes the accessibility to supermarkets, farmers' markets and local food stores as independent variables (Morland, Diez Roux, and Wing (2006); Rundle et al. (2009); Jilcott Pitts et al. (2013)). Spatial distribution of air quality is often represented as a regression model that includes the distance from chemical plants, power plants, smelters, and airports (Gordon and Gorham (1963); Riga-Karandinos and Karandinos (1998); Yu et al. (2004)). This paper calls these facilities *landmarks* hereafter. Since distance variables are measured on the same two-dimensional space, they are inevitably correlated at least to some extent with each other.

A wide variety of methods have been developed to handle with multicollinearity in regression analysis (Farrar and Glauber (1967); Asteriou and Hall (2007); Chatterjee and Hadi (2013)). A simple method is to remove independent variables that are highly correlated with others. Numerical measures are available for this purpose including correlation coefficient, determinant of correlation matrix, VIF, and more sophisticated indices (Weissfeld and Sereika (1991); Kovàcs, Petres, and Tóth (2005); Curto and Pinto (2007); Dormann et al. (2013)). Ridge regression, principal component regression, and their extensions are also useful for mitigate the multicollinearity among distance variables (Mansfield, Webster, and Gunst (1977); Vigneau et al. (1997); Kashid and Kulkarni (2002); Ni (2011); Chen (2012); Miller (2012)).

The above methods implicitly assume that the observation data used in regression models have already been prepared. On the other hand, if data are not yet collected, or a plenty of data are available from which we can select a subset of observations, we can reduce the multicollinearity among distance variables by carefully choosing the location of sample points. The former case often happens in physical geography and environmental science, where analysts collect their own data on the quality of air and water, soil and vegetation, and the distribution of animal species by a field survey. The latter includes spatial data of real estate such as multiple listing system in the US, land value data in Germany, and land and property survey data in Japan.

To derive an appropriate location of sample points, Heikkila (1988) evaluates the multicollinearity among distance variables using numerical experiments under hypothetical circumstances. The paper assumes a unit interval and a unit square on which either two or three landmarks

are located. Given a population distribution from which sample data are drawn, the paper evaluates the location of landmarks in terms of the correlation between distance variables. The results provide useful implications for a desirable location of sample domain in relation to the location of landmarks. Focusing on the case of two landmarks on a one-dimensional unit interval, Dewhurst (1993) also discusses the choice of sample domain. Instead of distance correlation, the paper utilizes the variances of estimated parameters to evaluate the degree of multicollinearity. Numerical simulations yield the location of sample domain that minimizes the variances and hence maximizes the stability of model estimation.

Following the line of Heikkila (1988) and Dewhurst (1993), this paper discusses more extensively the configuration of sample points that reduces the multicollinearity among distance variables. Our focus is on the case of two-dimensional space where two or more landmarks are located. We propose a mathematical method for deriving the optimal location of sample points that minimizes the correlation between distance variables. The method permits us not only to locate sample points appropriately but also to evaluate the location of landmarks in terms of the correlation between distance variables in a systematic way.

Section 2 proposes a method for locating sample points that reduces the multicollinearity among distance variables as small as possible. Description of the method is accompanied with numerical experiments and empirical applications that test the validity of the method and provide findings useful for the design of sample location. Section 3 summarizes the conclusions with discussion.

2. Method and applications

Suppose a spatial phenomenon represented as a continuous function $y(\mathbf{x})$ defined over a two-dimensional space, such as land price, ground temperature, and elevation. The function is measured at N sample points. Let P_i, $\mathbf{p}_i$, and y_i be ith sample point, its location, and the function value observed at P_i, respectively, where $i \in \mathbf{N} = \{1, 2, ..., N\}$. Function $y(\mathbf{x})$ is determined by both spatial and aspatial factors, the former of which include the distance to landmarks. Let F_j and $\mathbf{z}_j$ be jth landmark and its location, respectively ($j \in \mathbf{L} = \{1, 2, ..., L\}$). The distance between location $\mathbf{p}$ and $\mathbf{z}_j$ is denoted by $d(\mathbf{p}, \mathbf{z}_j)$. Aspatial factors are represented by functions $f_1(\mathbf{x})$, ..., $f_M(\mathbf{x})$.

We build a regression model that explains $y(\mathbf{x})$ by its determinants based on the data observed at sample points:

$$y_i = \alpha_0 + \alpha_1 f_1(\mathbf{p}_i) + ... + \alpha_M f_M(\mathbf{p}_i) + \beta_1 d(\mathbf{p}_i, \mathbf{z}_1) + ... + \beta_L d(\mathbf{p}_i, \mathbf{z}_L) + \varepsilon(\mathbf{p}_i) \tag{1}$$

where α_k's and β_k's are the parameters to be estimated ($k \in \mathbf{L}$) and $\varepsilon(\mathbf{p}_i)$ is an error term. We omit aspatial factors in the model for the present to focus on the multicollinearity among distance variables:

$$y_i = \beta_0 + \beta_1 d(\mathbf{p}_i, \mathbf{z}_1) + \beta_2 d(\mathbf{p}_i, \mathbf{z}_2) + ... + \beta_N d(\mathbf{p}_i, \mathbf{z}_L) + \varepsilon(\mathbf{p}_i) \tag{2}$$

where β_k's are the parameters to be estimated ($k \in \mathbf{L}$). The multicollinearity among distance variables is unavoidable since the landmarks are distributed on the same two-dimensional space.

Our objective is to locate sample points with keeping the multicollinearity among distance variables as small as possible. This requires us to evaluate the multicollinearity in an objective way, for the purpose of which various methods are available at present. A simple measure is the absolute value of correlation coefficient between two variables. A rule of thumb tells us that the correlation coefficient smaller than 0.5 is acceptable. The variance inflation factor (VIF) is another popular index that assesses multicollinearity in terms of the stability of model estimation. Further sophisticated approaches have also been developed in the literature that evaluate and mitigate the multicollinearity (Belsley, Kuh, and Welsch (1980); Spanos and McGuirk (2002); Kovàcs, Petres, and Tóth (2005); Curto and Pinto (2007); Dormann et al. (2013)).

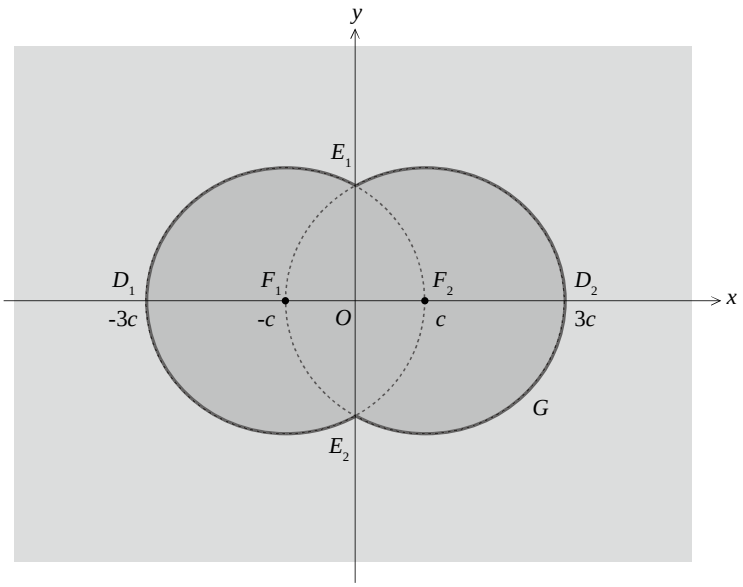
Among those methods, we initially choose the maximum absolute correlation coefficient between distance variables as the measure of multicollinearity because of its simplicity and tractability. We derive the optimal configuration of sample points that minimizes the maximum absolute correlation coefficient denoted by $\max\{|r_{jk}|\}$ ($j, k \in \mathbf{L}, j \neq k$), where r_{jk} is the correlation coefficient between $d(\mathbf{p}, \mathbf{z}_j)$ and $d(\mathbf{p}, \mathbf{z}_k)$ observed in sample points. We should note, however, that other measures of multicollinearity are also applicable to the method proposed in the following with a slight modification as discussed later.

2.1 Two landmarks

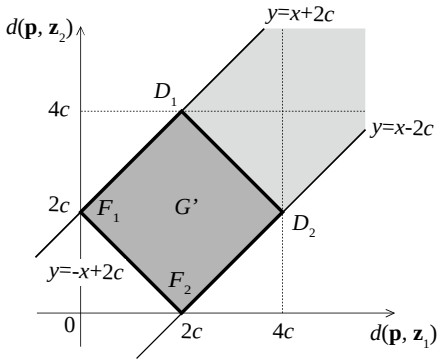
We start with the optimal configuration of sample points when $y(\mathbf{x})$ depends only on the distance to two landmarks, i.e., L=2. This case is rather trivial since we can easily keep r_{12} at zero as follows. We locate the first sample point P_1 arbitrarily and draw the circle of radius $d(\mathbf{p}_1, \mathbf{z}_1)$ centered at F_1. Placing other sample points on the circle, we can change $d(\mathbf{p}, \mathbf{z}_2)$ with keeping $d(\mathbf{p}, \mathbf{z}_1)$. The correlation coefficient r_{12} becomes zero so that we can estimate β_2 without being disturbed by the correlation between $d(\mathbf{p}, \mathbf{z}_1)$ and $d(\mathbf{p}, \mathbf{z}_2)$. We can similarly estimate β_1 by using a circle of radius $d(\mathbf{p}_1, \mathbf{z}_2)$ centered at F_2.

This method, unfortunately, limits the location of sample points only on two circles. Such a tight restriction is practically undesirable because the clustering of sample points often increases the correlation between spatial and aspatial variables in regression models, which also greatly reduces the stability and reliability of model estimation. We thus propose an alternative location of sample points that spreads over a wider area.

Suppose XY coordinate system as shown in Figure 1a, where the location of two landmarks F_1 and F_2 are denoted as $\mathbf{z}_1$=(-c, 0) and $\mathbf{z}_2$=(c, 0), respectively. We also consider another space whose XY axes indicate $d(\mathbf{p}, \mathbf{z}_1)$ and $d(\mathbf{p}, \mathbf{z}_2)$ as shown in Figure 1b. We call the latter *distance space* to distinguish from *real space* shown in Figure 1a. Any location in the real space is projected in light and dark gray regions in Figure 1b, which is bounded by three lines $y=x+2c$, $y=x-2c$, and $y=-x+2c$.



(a)



(b)

Figure 1 The relationship between the real space and the distance space defined by $d(\mathbf{p}, \mathbf{z}_1)$ and $d(\mathbf{p}, \mathbf{z}_2)$. (a) The real space. Landmarks F_1 and F_2 and circles of radius $2c$ centered at F_1 and F_2 are drawn. (b) The distance space. Light and dark gray regions indicate the possible relationship between $d(\mathbf{p}, \mathbf{z}_1)$ and $d(\mathbf{p}, \mathbf{z}_2)$ in the distance space. Dark gray square corresponds to the dark gray region in Figure 1a. No correlation appears between the two distances if sample points are distributed symmetrically in both vertical and horizontal directions.

The correlation coefficient r_{12} becomes zero if sample points are distributed symmetrically in both vertical and horizontal directions in the distance space. Clearly, numerous distributions satisfy this condition. Among such distributions, we take a uniform distribution in the dark gray square denoted by G' in Figure 1b that corresponds to the dark gray region G in Figure 1a. We choose this distribution since sample points are spread relatively widely around the landmarks. Suppose a uniform function in G' whose

integral in G is equal to one. The function corresponds to the continuous function defined by

$$f_{12}(\mathbf{p}; F_1, F_2) = \frac{L(\mathbf{p}; F_1, F_2)}{2\{d(\mathbf{z}_1, \mathbf{z}_2)\}^2 d(\mathbf{p}, \mathbf{z}_1) d(\mathbf{p}, \mathbf{z}_2)}, \tag{3}$$

where $L(\mathbf{p}; F_1, F_2)$ is the length of perpendicular line from location $\mathbf{p}$ to F_1F_2 (see Sadahiro and Wang (2015) for the derivation of the equation). We call this function the *density function* of sample points, which is illustrated in Figure 2. The density is high around the circle whose diameter is F_1F_2. We denote the circle as C_{12} indicated by the white broken line in the figure. Two distances $d(\mathbf{p}, \mathbf{z}_1)$ and $d(\mathbf{p}, \mathbf{z}_2)$ are not highly correlated around the circle. The density function $f_{12}(\mathbf{p}; F_1, F_2)$ decreases as $\mathbf{p}$ approaches the X-axis since the correlation between $d(\mathbf{p}, \mathbf{z}_1)$ and $d(\mathbf{p}, \mathbf{z}_2)$ is highly positive around D_1F_1 and F_2D_2 while highly negative around F_1F_2.

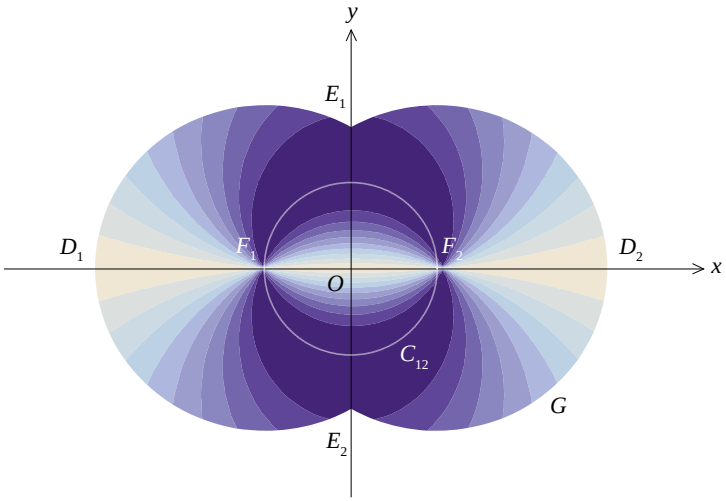


Figure 2 Density function $f_{12}(\mathbf{p}; F_1, F_2)$ defined by Equation (3). Darker shades indicate larger values.

We can reduce r_{12} by locating sample points according to $f_{12}(\mathbf{p}; F_1, F_2)$. We approximate a point distribution that follows a probability distribution on a two-dimensional as follows. We first divide the region in which we locate sample points by a square lattice of a high resolution, say, 10000 or more cells, and calculate the probability of a sample point located in each cell by using numerical integration (Davis and Rabinowitz (2007)). Let λ_i be the probability of a sample point located in ith cell. Cumulative distribution function of ith cell, which is denoted by Λ_i, is given by

$$\Lambda_i = \sum_{j=1}^{i-1} \lambda_j \,. \tag{4}$$

We then generate a random number x between zero and one, and locate a sample point in ith cell that satisfies

$$\Lambda_i \le x < \Lambda_{i+1}. \tag{5}$$

Repeating this process N times, we can approximate a point distribution that follows $f_{12}(\mathbf{p}; F_1, F_2)$. The below is a computational algorithm of this procedure.

1. Prepare a square lattice of K cells.
2. Set i=1.
3. Calculate λ_i, the probability of a sample point located in ith cell.
4. i=i+1.
5. Repeat the steps 3-4 until i=K.
6. Set i=1.
7. Calculate the cumulative distribution function of ith cell defined by Equation (4).
8. i=i+1.
9. Repeat the steps 7-8 until i=K.
10. Set i=1.
11. Generate a random number x between zero and one.
12. Locate a sample point in ith cell that satisfies Equation (5).
13. i=i+1.
14. Repeat the steps 11-13 until i=N.

To test the effectiveness of the density function derived above, we perform a numerical experiment. We distribute 1,000 points according to $f_{12}(\mathbf{p}; F_1, F_2)$ in G (Dark gray region in Figure 1) and calculate the correlation coefficient r_{12}. Repeating this process for 10,000 times, we obtain the probability distribution of r_{12}. We also calculate the probability distribution of r_{12} when sample points are randomly distributed in G for comparison purposes.

Figure 3 shows the result of this numerical experiment. The figure clearly indicates that $f_{12}(\mathbf{p}; F_1, F_2)$ greatly reduces r_{12} compared with the random distribution. The probability of r_{12} falling between -0.1 and 0.1 is only four percent in the random distribution while it is 95 percent when sample points follow $f_{12}(\mathbf{p}; F_1, F_2)$. Density function $f_{12}(\mathbf{p}; F_1, F_2)$ sounds safe enough for locating sample points in the sense that no serious correlation occurs between distance variables.

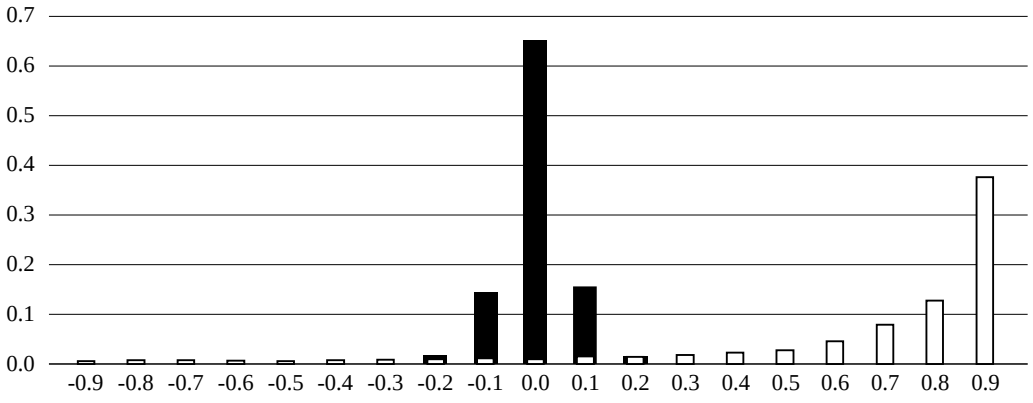


Figure 3 Probability distributions of correlation coefficient r_{12}. Black bars indicate the probability distribution when sample points are distributed according to $f_{12}(\mathbf{p}; F_1, F_2)$, while white bars indicate the probability distribution when sample points are randomly distributed.

The above procedure of locating sample points is effective when sample points can be located only in a specific region, which is often called a *sample domain*. The density function of sample points in sample domain H can be calculated as follows. We first project H from the real space onto the distance space by calculating the distance from each vertex of H to two landmarks and connecting the vertices in the distance space to obtain polygon H'. We then choose a subdomain H_S' inside H' in which sample points can be distributed symmetrically in both $d(\mathbf{p}, \mathbf{z}_1)$ and $d(\mathbf{p}, \mathbf{z}_2)$ directions. We finally determine the configuration of sample points in H_S' and project it back onto the real space.

We can choose subdomain H_S' in various ways. A simple method is to draw a small circle or an axis-parallel square inside H_S'. To locate sample points more widely, we can employ computational algorithms that calculate the largest axis-parallel rectangle inside a polygon (Daniels, Milenkovic, and Roth (1997); Boland and Urrutia (2001)). The below is a computational algorithm of the above procedure.

1. Approximate H by a polygon of vertices V_1, ..., V_K.
2. Set i=1.
3. Calculate the distance from V_i to landmarks F_1 and F_2.
4. Locate the point corresponding to V_i in the distance space.
5. i=i+1.
6. Repeat the steps 3-5 until i=K.
7. Connect the obtained points in the distance space to generate polygon H'.
8. Choose a rectangular subdomain H_S' inside H'.
9. Locate N sample points symmetrically in H'. Let z_1 and z_2 be the coordinates of ith point.
10. Set i=1.

11. Draw two circles of radius z_1 and z_2 centered at F_1 and F_2, respectively, in the real space.
12. Locate a sample point at an intersection of the circles.
13. i=i+1.
14. Repeat the steps 11-14 until i=N.

If sample points are already prepared, we extract a subset of points according to $f_{12}(\mathbf{p}; F_1, F_2)$. We first calculate $f_{12}(\mathbf{p}; F_1, F_2)$ based on a square lattice as mentioned earlier. We then choose sample points in each cell in proportion to its assigned probability. If a cell does not contain enough number of points, we leave the cell and examine the next cell. We finally choose sample points from the cells that still have enough points in proportion to $f_{12}(\mathbf{p}; F_1, F_2)$.

2.2 More than two landmarks

We then consider regression models whose independent variables contain the distances from more than two landmarks. Unfortunately, it is almost impossible to avoid the correlation between more than two distance variables simultaneously. We thus aim to reduce the correlation as small as possible, i.e., to minimize the maximum absolute correlation coefficient between distance variables, by carefully choosing sample location. We formulate the derivation of the optimal configuration of sample points as a spatial optimization problem:

Problem CM (Correlation Minimization):

$$\underset{\mathbf{p}_i, i\in\mathcal{N}}{\text{minimize}} \max\left\{\left|r_{jk}\right|\right\} \quad (j,k\in\mathcal{L}, k\neq j)$$

To solve Problem CM, we initially distribute sample points according to the following density function:

$$g(\mathbf{p}) = \frac{2}{L(L-1)}\sum_{j}\sum_{k\neq j} f_{jk}(\mathbf{p}; F_j, F_k). \tag{6}$$

This function is the average of density function $f_{jk}(\mathbf{p}; F_j, F_k)$ for all the pair of landmarks. We choose this function in expectation that $g(\mathbf{p})$ reduces r_{jk} of every pair of j and k equally and thus yields a small $\max\{|r_{jk}|\}$. We determine the initial location of sample points in the way similar to that discussed in the previous subsection. We then move every sample point from P_1 to P_N in turn by the steepest descent method (Hamacher and Drezner (2002); Avriel (2003); Snyman (2005); Fletcher (2013)). The method moves each point in the direction that decreases the objective function most rapidly. We repeat this process until the movement of sample points converges within a predetermined threshold. The obtained $\max\{|r_{jk}|\}$ is denoted by $r_{\max}$.

Using this method, we perform a numerical experiment. We locate L landmarks randomly in a circle of radius 1.0 and derive an optimal configuration of sample points. We repeat this process 1000 times to obtain the probability distribution of r_{max}. Table 1 describes the summary statistics of the distribution. The mean of r_{max} increases monotonically with L, which implies that correlation between distance variables becomes more unavoidable. If we accept r_{max} smaller than 0.5, we can always optimize the configuration of sample points when $L \leq 3$ since Prob($r_{max}<0.5$) is 100 percent. The probability Prob($r_{max}<0.5$) decreases to 62.3 percent when $L=4$, and then 8.7 percent when $L=5$. A serious gap between $L=4$ and $L=5$ cases suggests that it is safe to consider less than five landmarks simultaneously in terms of multicollinearity in regression analysis.

Table 1 Summary statistics of r_{max} under a random configuration. Prob($r_{max}<0.5$) indicates the probability of r_{max} being smaller than 0.5.

L	2	3	4	5	6	7
Mean	0.0000	0.2738	0.5263	0.7182	0.8155	0.8890
Standard deviation	0.0000	0.2041	0.1284	0.1507	0.1070	0.0722
Prob($r_{max}<0.5$)	100.0	100.0	62.3	8.7	0.0	0.0

In practice, we often have to limit the number of distance variables equal or less than three due to the high correlation between the variables. Considering this, we can regard the proposed method effective enough since it always assures $r_{max}<0.5$ when $L=3$, and permits four landmarks with a probability of 62%. We even recommend the method when $5 \leq L$ since we can reduce r_{max} at an acceptable level in some situations.

The above procedure is also applicable when sample points are already prepared. We determine the initial set of sample points according to $g(\mathbf{p})$. We then optimize the combination of sample points to minimize the objective function of Problem CM. Since it is a combinatorial optimization problem, we use a heuristic algorithm such as simulated annealing, tabu search, and genetic algorithms to reach the final result (Nemhauser and Wolsey (1988); Wolsey (1998); Karlof (2005)).

2.3 Extensions

This subsection briefly discusses several extensions of the proposed method. Due to space limitations, we will examine them more thoroughly elsewhere in the near future.

Problem CM utilizes $\max\{|r_{jk}|\}$ as its objective function. The problem, however, can utilize other measures such as VIF, tolerance value, and eigenvalue instead of $\max\{|r_{jk}|\}$ as a measure of multicollinearity. Replacing the objective function by a different measure, we can optimize the configuration of sample points from a different perspective. We can even incorporate principal component regression by replacing r_{max} with that obtained after principal component analysis. The optimal location

of sample points under the use of principal component regression is highly probable to reduce $r_{\max}$ to 0.5 even when $5 \leq L$.

We can also incorporate aspatial variables into our framework as follows. Aspatial variables are often correlated with distance variables under the existence of spatial autocorrelation, which typically emerges in the clusters of sample points. A practical solution is thus to add the minimum distance between sample points as a constraint in Problem CM. We may use the nearest neighbor distance between points for a certain significance level that are randomly distributed in sample region (Pielou (1977); Ripley (2005)). If we can build a mathematical model that represents the spatial distribution of aspatial variables, we can directly incorporate the model into the optimization procedure. The model gives the correlation coefficients between spatial and aspatial variables and those between aspatial variables. Replacing $\{|r_{jk}|\}$ with the set of all the absolute correlation coefficients, we solve Problem CM to derive the optimal location of sample points that minimize the correlation between all the variables.

2.4 Three landmarks

The numerical experiment in the previous subsection showed that we can keep the correlation between distance variables at a reasonable level when $L \leq 4$. This subsection discusses further in detail the relationship between the location of landmarks and the maximum correlation $r_{\max}$ with a focus on the case of $L=3$.

Clusters of sample points are undesirable as mentioned earlier. Since careful consideration is indispensable on the spatial pattern of sample points, we introduce two measures to evaluate the degree of spatial clustering. One is *standard distance* defined by

$$S_D = \sqrt{\frac{\sum_i \left\{ d\left(\mathbf{p}_i, \mathbf{p}_C\right) \right\}^2}{M}} \quad , \tag{7}$$

where $\mathbf{p}_C$ is the mean center of sample points (Bachi (1962); Rogerson and Yamada (2008); Burt, Barber, and Rigby (2009)). Standard distance S_D evaluates the overall degree of spatial dispersion of sample points. It becomes large when sample points are widely dispersed. Another measure is *relative nearest neighbor distance* defined by

$$R_D = \frac{\dfrac{\sum_i \min_l \left\{ d\left(\mathbf{p}_i, \mathbf{p}_l\right) \right\}}{N}}{S_D} \quad . \tag{8}$$

It is the ratio of the average nearest neighbor distance of sample points to S_D. Relative nearest neighbor distance becomes small when points form local clusters, while it is large if points are widely scattered.

The two measures complement with each other by evaluating the degree of spatial clustering from different perspectives.

Using three measures r_{max}, S_D, and R_D, we discuss the relationship between the location of landmarks and the maximum correlation through numerical experiments. Three landmarks are arranged in two ways. They form an isosceles triangle in Case 1, while a sheared triangle in Case 2. Landmarks F_1 and F_2 are located at (-1, 0) and (1, 0) in both cases, respectively. The third landmark F_3 is located on the y-axis and horizontal line $y = \sqrt{3}$ as shown in Figure 4, respectively. The location of F_3 is denoted as $\left(0, \frac{\sqrt{3}}{15}m\right)$ and $\left(\frac{\sqrt{3}}{15}m, \sqrt{3}\right)$ in the two cases, respectively. Parameter m changes from 0 to 75 in both cases. In Case 1, triangle $\Delta F_1F_2F_3$ becomes an equilateral triangle when m=15. In Case 2, $\Delta F_1F_2F_3$ is an equilateral triangle when m=0, and gradually becomes a sheared triangle with an increase in m.

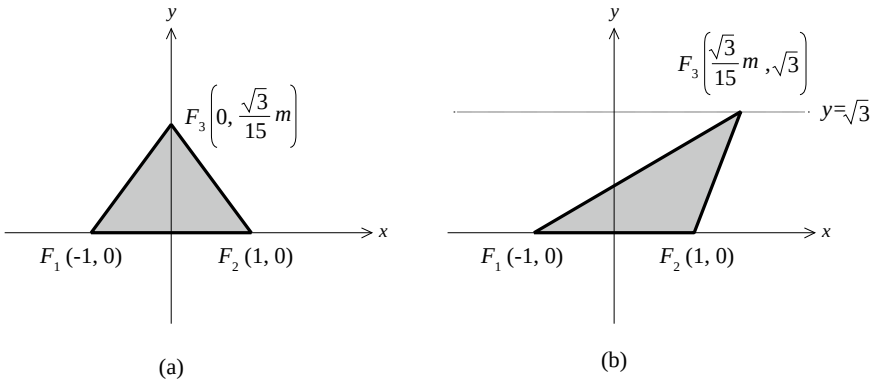
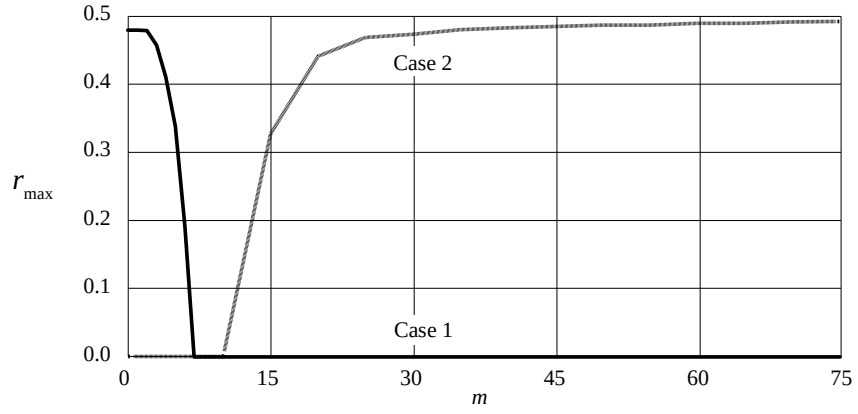


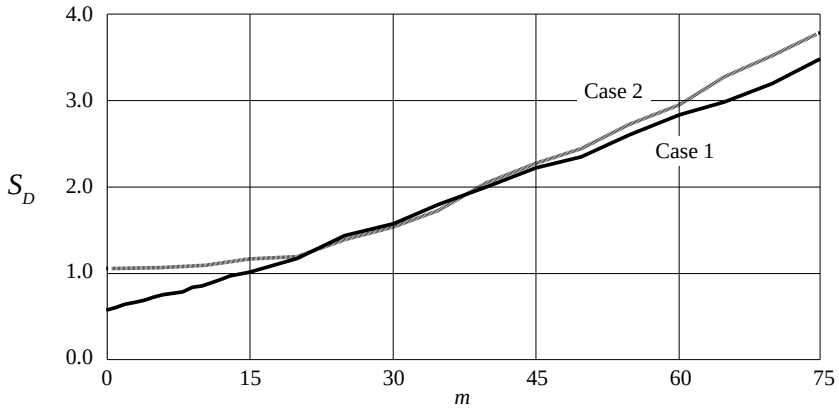
Figure 4 The location of three landmarks F_1, F_2, and F_3. (a) Case 1: The landmarks form an isosceles triangle, (b) Case 2: The landmarks form a sheared triangle.

We first discuss Case 1, where $\Delta F_1F_2F_3$ forms an isosceles triangle. Figure 5 shows the relationship between m and numerical measures. As seen in the figure, r_{max} initially decreases greatly with an increase of m, and stays zero when $7 \leq m$. The correlation between distance variables completely vanishes when vertex angle $\angle F_1F_3F_2$ is smaller than 0.567π=102.1°. Figure 6 shows the final configuration of sample points when m=5, 10, 15, 30, 45, and 75, where C_{ij} is the circle of diameter F_iF_j centered at the midpoint of F_i and F_j. While Figure 6a and Figure 6b contain clusters, points are distributed rather uniformly in Figure 6c, where $\Delta F_1F_2F_3$ forms an equilateral triangle. Points become gathered along a circle centered at the midpoint of F_1 and F_2 with an increase in m. This is clearly reflected in Figure 5c where R_D first increases and then decreases monotonically. Figure 5b shows that standard distance S_D captures the global pattern of sample points successfully, where S_D increases monotonically with a spread

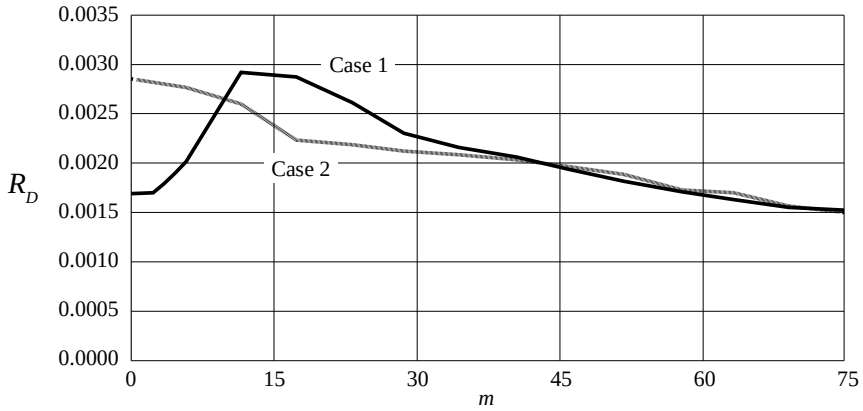
of sample points over a wider area.



(a)



(b)



(c)

Figure 5 Measures of the configuration of sample points. (a) The maximum correlation r_{max}, (b) standard distance S_D, (c) relative nearest neighbor distance R_D.

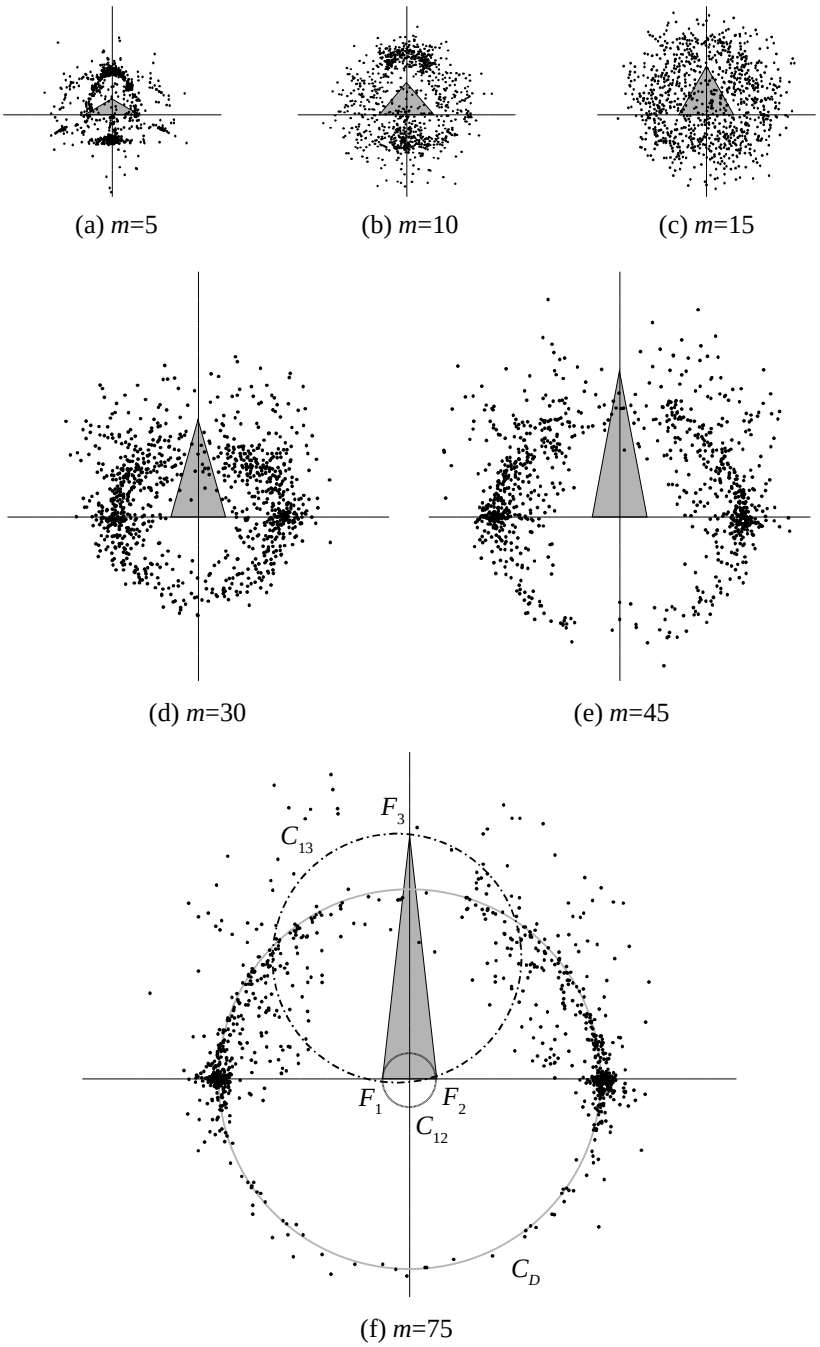


Figure 6 The final configuration of sample points in Case 1. Gray circle denoted as C_D in Figure 6f indicates the clustering of sample points. Broken lines indicate circles C_{12} and C_{13}.

To examine the clustering of sample points around a circle in detail, we discuss the relationship between sample location and distance variables in the case of two sample points. Suppose Figure 7a in which the first sample point P_1 is located. Broken and solid lines indicate arcs centered at F_1 and F_2 and their tangents at P_1, respectively. Let us consider the location of second point P_2 around the neighborhood

of P_1.If we place P_2 between the tangents, the area indicated by gray-shades, correlation coefficient r_{12} becomes positive. Figure 7b depicts the angle of P_2 with respect to F_1 that yields a positive r_{12}. Black area in small circles indicates the range of the angle giving positive r_{12}. As seen in the figure, r_{12} tends to be positive outside circle C_{12} while it becomes negative inside C_{12}. If we locate P_2 randomly around F_1, the probability of positive r_{12} is given by

$$\mathrm{Prob}\left(r_{12}>0\right)=\begin{cases}\dfrac{1}{\pi}\left(\arctan\dfrac{|y_1|}{x_1-c}-\arctan\dfrac{|y_1|}{x_1+c}\right) & \left(x_1<-c, c<x\right)\\ 1-\dfrac{1}{\pi}\left(\arctan\dfrac{|y_1|}{x_1-c}+\arctan\dfrac{|y_1|}{x_1+c}\right) & \left(-c\le x_1\le c\right)\end{cases}. \tag{9}$$

Figure 7c show the distribution of Prob(r_{12}>0). Probability Prob(r_{12}>0) is 1/2 on C_{12}, which implies that points clustered around C_{12} do not exhibit strong correlation between distance variables. We can confirm in this figure that the correlation r_{12} is negative inside C_{12}, while it becomes highly positive as sample points move away from C_{12} toward its outside.

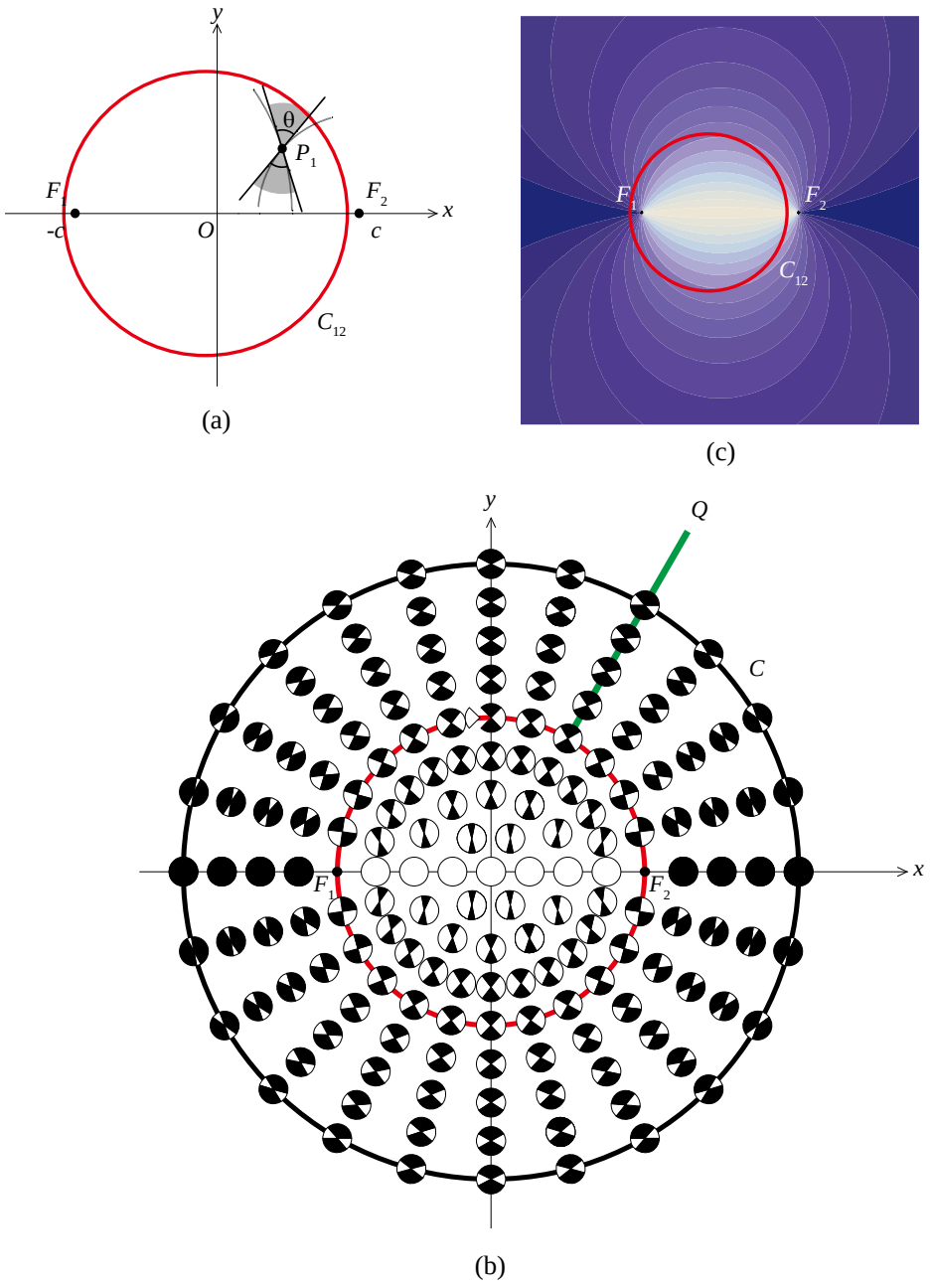


Figure 7 Calculation of the probability distribution of positive r_{12}. Red circles indicate circle C_{12}. (a) The location of the first sample point P_1. (b) The angle of P_2 with respect to P_1 that yields a positive r_{12}. Black area in small circles indicate the range of the angle giving positive r_{12}. (c) The distribution of probability Prob(r_{12}>0). Darker shades indicate that distance variables tend to be highly positively correlated.

Figure 7b also suggests that a specific form of sample clusters causes a strong correlation, either positive or negative. Suppose sample points aligned on an extension of a diameter of C_{12} such as green

line Q in Figure 7b. The line is running through the black area of small circles at every location in the figure. This indicates that $d(\mathbf{p}, \mathbf{z}_1)$ and $d(\mathbf{p}, \mathbf{z}_2)$ are positively correlated at all these locations, and consequently, r_{12} becomes 1.0. In contrast, sample points located on F_1F_2 yields r_{12}=-1.0. Negative correlation also occurs when sample points are aligned on a circle centered at the midpoint of F_1 and F_2 such as circle C in Figure 7b. The circle is passing through the white area of small circles at every location, which yields a highly negative correlation.

Keeping the above patterns in mind, let us return to Figure 6f. We examine correlation coefficients r_{12} and r_{13} in detail to reveal how the sample points accomplish r_{max}=0(=r_{12}=r_{13}). Many points are concentrated along the circle denoted by C_D, whose center is located at the midpoint of F_1 and F_2. We expect the points to cause a highly negative correlation, which is supported by Figure 8a showing the relationship between $d(\mathbf{p}, \mathbf{z}_1)$ and $d(\mathbf{p}, \mathbf{z}_2)$ on C_D. Other points are all dispersed outside C_{12}, which leads to a positive correlation as confirmed in Figure 7c. The negative and positive correlations cancel out with each other, and consequently, r_{12} becomes zero as a whole in Figure 6f. We then turn to the case of r_{13}. Figure 8b shows the relationship between $d(\mathbf{p}, \mathbf{z}_3)$ and the other distance variables on C_D. The figure exhibits no significant correlation. Other points are distributed equally inside and outside of C_{13} in Figure 6f so that negative and positive correlations similarly emerge and cancel out with each other. Correlation r_{13} also becomes zero, so does r_{max}.

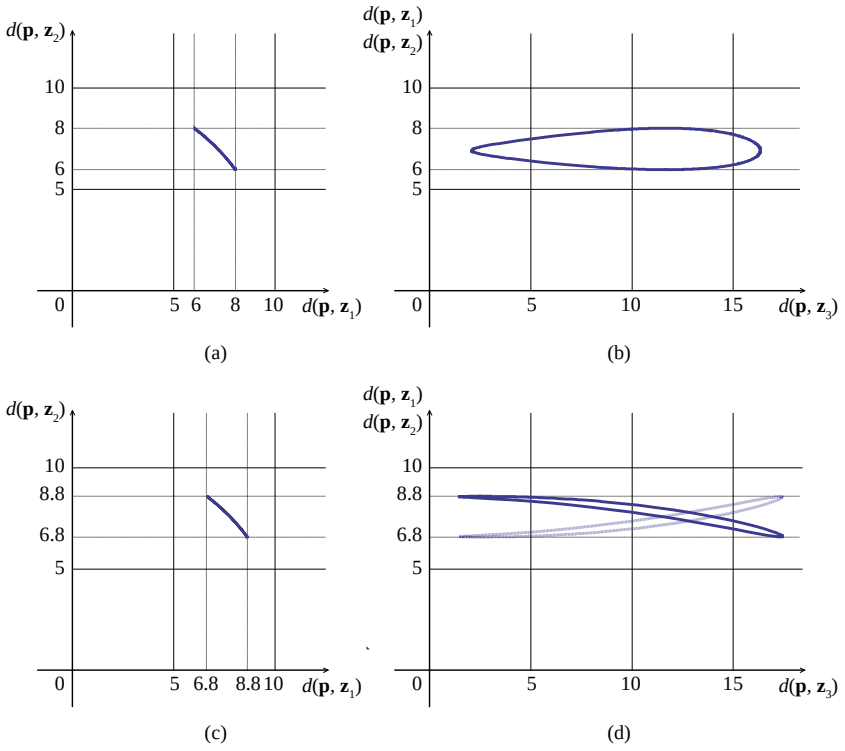


Figure 8 The relationship between distance variables of sample points located on circles. Points are located on (a)(b) circle C_D in Figure 6f, and (c) (d) C_D' in Figure 9e. Bold and broken lines in (b) (d)

indicate $d(\mathbf{p}, \mathbf{z}_1)$ and $d(\mathbf{p}, \mathbf{z}_2)$, respectively.

We then discuss Case 2, where $\Delta F_1F_2F_3$ forms a sheared triangle. Figure 5a shows that $r_{\max}$ stays at zero until m reaches at 11, where $\angle F_1F_3F_2 \approx 0.287\pi = 51.6°$, and then monotonically increases as $\Delta F_1F_2F_3$ becomes further sheared. The final configuration of sample points is shown in Figure 9. Points are distributed uniformly around $\Delta F_1F_2F_3$ when m=0, and then gradually concentrated along a circle centered at the midpoint between F_1 and F_2 with an increase of m. This change is reflected in S_D and R_D shown in Figure 5. Interestingly enough, a clear concentration along two circles appears in Figure 9b as indicated by gray lines, where $\Delta F_1F_2F_3$ forms an isosceles triangle. Similar but less clear circles can also be observed in Figure 6a. Figure 9 suggests that sample points are initially distributed uniformly, then gathered along two circles, and finally clustered along a single circle.

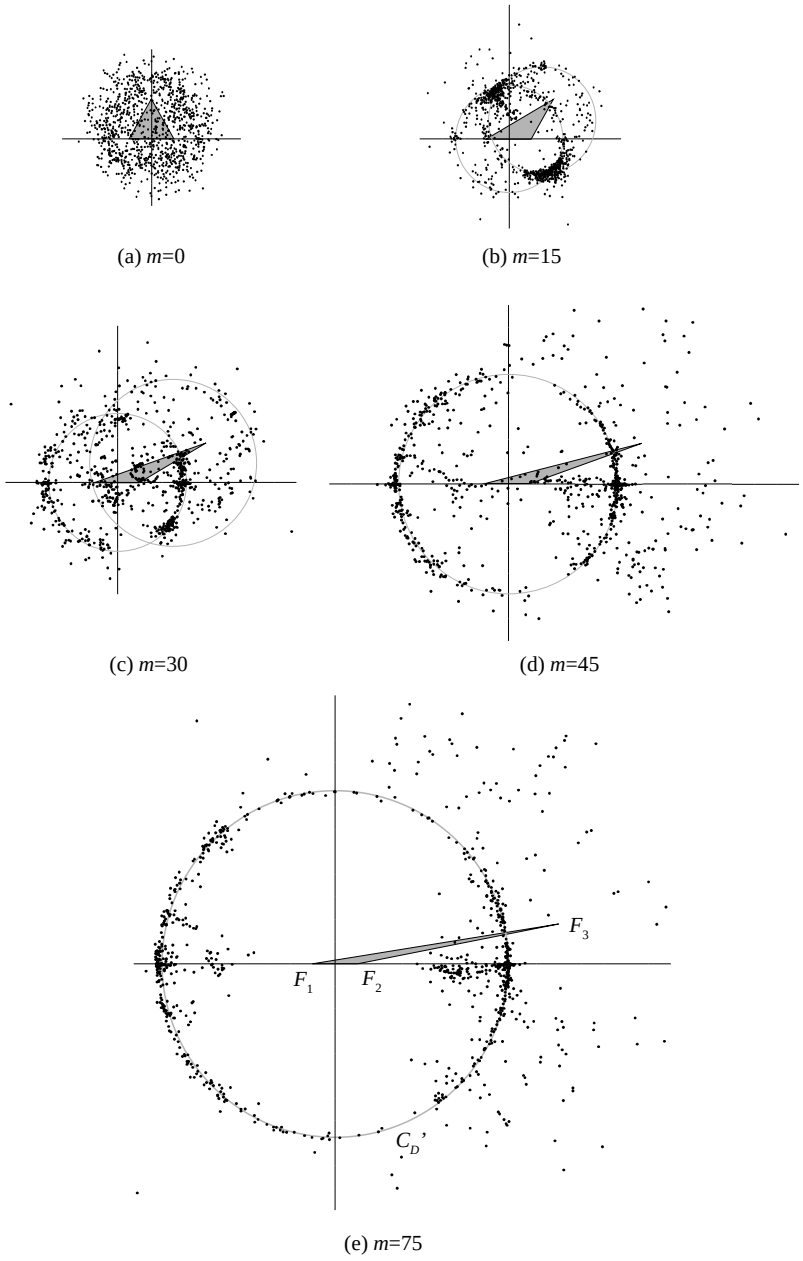


Figure 9 The final configuration of sample points in Case 2. Gray circles indicate the clustering of sample points.

The above applications suggests sufficient conditions for r_{ij}=0. Correlation r_{ij} becomes positive if sample points are either 1) dispersed outside C_{ij}, or 2) aligned on an extension of a diameter of C_{ij}. It becomes negative when sample points are either 3) dispersed inside C_{ij}, or 4) aligned on a diameter of C_{ij}, or 5) aligned on a circle that is larger than C_{ij} and centered at the midpoints of F_iF_j. The combination of one positive and one negative conditions yields r_{ij}=0. Figure 6c and Figure 9a are the cases of conditions 1) and 3), where sample points are dispersed equally inside and outside C_{ij}. Figure 6f, Figure 6e, Figure

9d, and Figure 9e are the cases of 1) and 5). Correlation r_{max} becomes zero only if all the correlations are equal to zero by satisfying some of the above conditions.

What is interesting here is that condition 5) is found only for the shortest edge of $\Delta F_1F_2F_3$ in the above examples. This is because of the following reason. Suppose that F_2F_3 is the shortest among the three edges. Correlation r_{12} is zero if sample points are concentrated along circle C centered at the midpoints of F_1F_2 while other points are distributed outside C_{12} as seen in Figure 10. A necessary condition for $r_{23}=r_{31}=0$ in this case is that the sample points are distributed equally both inside and outside C_{23} and C_{31}. However, since C_{23} is smaller than C, many points are inevitably located outside C_{23}. When F_2F_3 is the shortest edge, therefore, a necessary condition for $r_{max}=0$ is that sample points are concentrated along circle C centered at the midpoints of F_2F_3 while other points are distributed outside C_{23}.

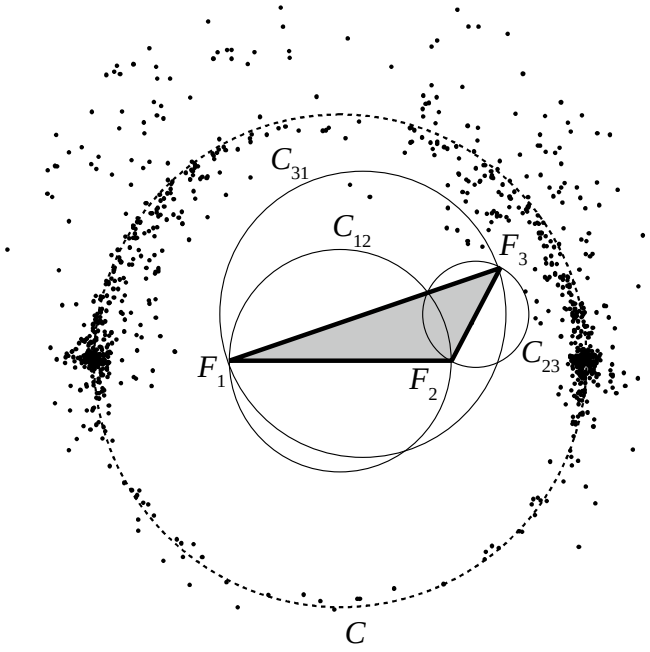


Figure 10 The location of sample points, $\Delta F_1F_2F_3$ and its related circles.

We finally discuss a desirable configuration of three landmarks in terms of the correlation between distance variables. If the concentration of sample points does not cause a strong correlation between spatial and aspatial variables, we can adopt any configuration discussed in this subsection on the assumption that $r_{max}<0.5$ is acceptable. If point clusters are undesirable, it is safe to choose landmarks that form an almost equilateral triangle such as those in Figure 6c and Figure 9a where sample points exhibit no significant clusters. If sample points need to be dispersed more widely, we may choose the configuration of landmarks such as those in Figure 6e and Figure 9d, though we have to allow local clusters of sample points to a certain degree.

2.5 Applications to real data

This subsection applies the proposed method to the analysis of real data. We chose two regions, each of which is located in Tochigi and Fukuoka Prefectures, Japan. Source data are land price data in 2014 collected at 23380 locations by the Ministry of Land, Infrastructure, Transport and Tourism. We consider a regression model that describes the land price at each location by the distance to landmarks.

In the first region in Tochigi Prefecture we took three cities as landmarks: Utsunomiya, Tochigi, and Kanuma (Figure 11). Sample domain H is a circle of radius 25 kilometers centered at the centroid of the three cities. The domain contains 277 sample points, whose $r_{\max}$ is 0.7425 (Figure 11a). To reduce $r_{\max}$ smaller than 0.5, we located sample points in two different ways: one is to newly locate 277 sample points in H without any restriction, and the other is to extract 50 percent of the original points, i.e., 138 from 277 points.

Figure 11b is the distribution of sample points located without any restriction. Sample points are uniformly distributed in H, which is almost the same as the initial location of sample points given by Equation (3). If we further continue optimization, we could reduce $r_{\max}$ to 0.00. Figure 11c and Figure 11d indicate the points randomly chosen from the original data and those obtained after optimization, respectively, where $r_{\max}$ reduced from 0.7310 to 0.50. Figure 11d looks reasonable since it contains no extreme cluster of points. If we further continue optimization, we could decrease $r_{\max}$ to 0.2988. Otherwise, we may increase sample points to reduce the loss of information with keeping $r_{\max}$=0.5. This option is more effective in this case since we could increase sample points up to 80 percent of the original ones, i.e., we could attain $r_{\max}$=0.5 with the loss of only 20 percent of the original information.

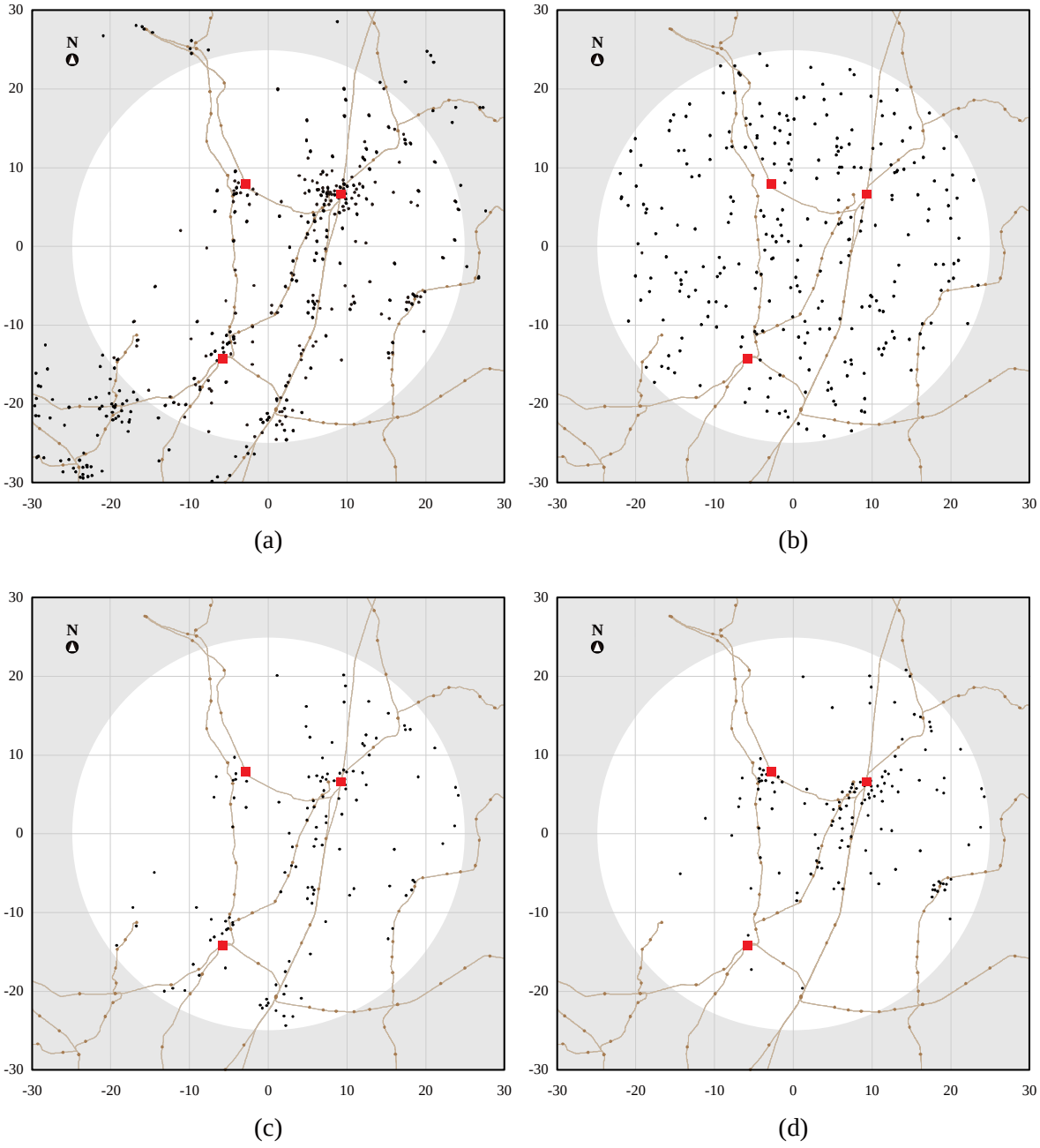


Figure 11 The location of sample points and landmarks in Tochigi Prefecture, indicated by black and red points, respectively. Brown lines are railway lines. (a) The actual location in 2014, (b) sample points located without any restriction, (c) sample points extracted randomly from their original location, (d) sample points obtained after optimization. The maximum correlation r_{max} is 0.5 in Figure 11b and Figure 11d.

Sample domain *H* in the second example is located in Fukuoka Prefecture. (Figure 12). It is a circle of radius 20 kilometers facing Hakata Bay centered at the centroid of three landmarks, i.e.,

downtown area of Fukuoka City, Higashi-Hirao Park that contains football and athletic stadiums, and a big shopping mall called Torius. Domain H contains 381 sample points, where r_{max} is 0.8445 (Figure 12a). We optimize the location of sample points in the same way as that in Tochigi Prefecture.

Figure 12b shows the distribution of sample points located without any restriction for r_{max}=0.5. We can see a rather obscure cluster of points around the three landmarks. It would be a circular cluster such as the one observed in Figure 6, a part of which is substituted by the cluster on the peninsula in Hakata Bay. This distribution seems practically infeasible since the peninsula is not densely inhabited. Figure 12c and Figure 12d are the points randomly chosen from the original ones and those obtained after optimization, respectively, where r_{max} reduced from 0.6827 to 0.5. We could either further reduce r_{max} to 0.2768, or increase sample points up to 95 percent of the original ones with keeping r_{max}=0.5. The latter is very effective in a practical sense as we can decrease r_{max} from 0.8445 to 0.5 by discarding only five percent of the original information. We can use original data more effectively than in the case of Tochigi Prefecture because Fukuoka Prefecture has more sample points that assure the flexibility of sample location.

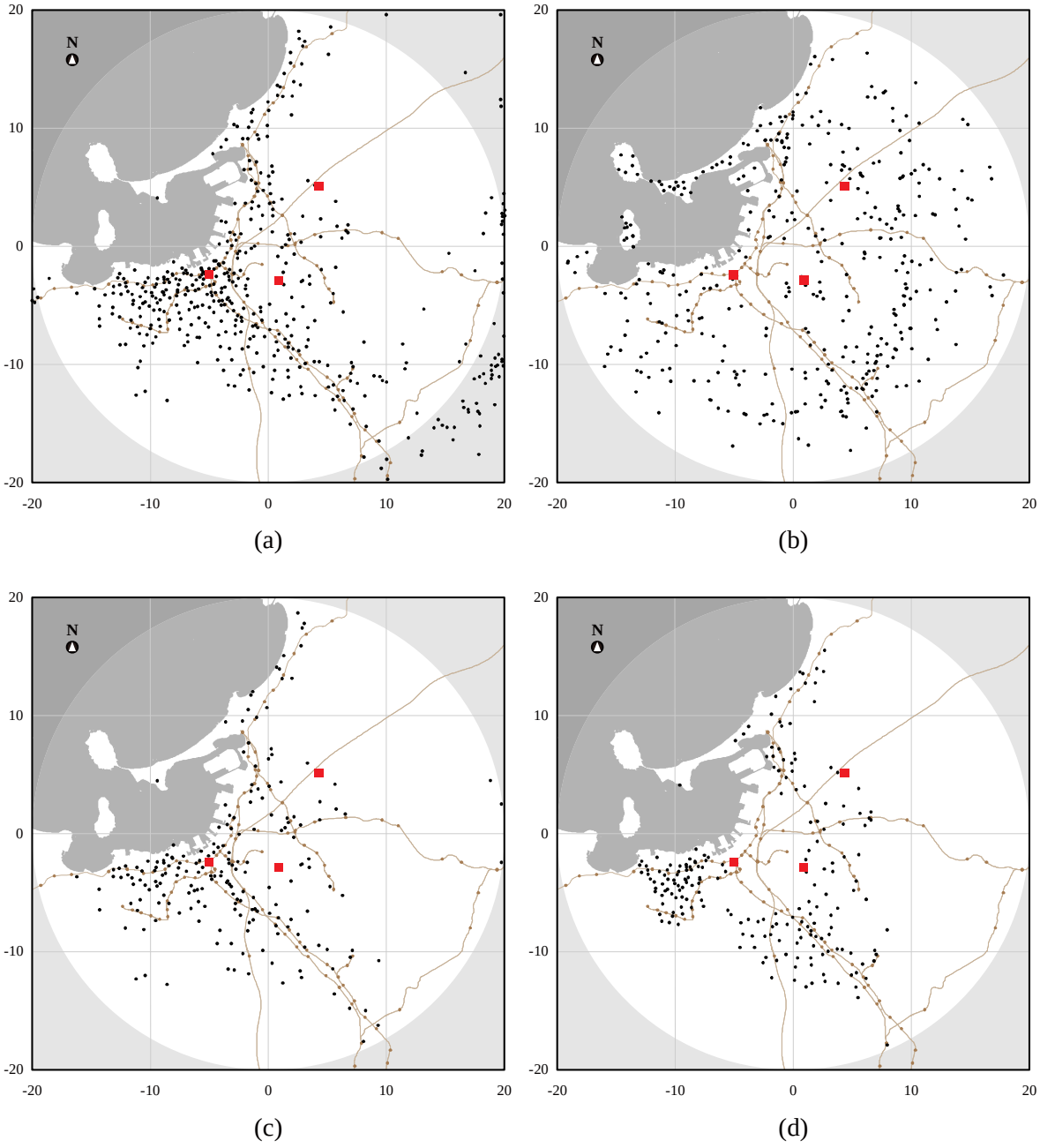


Figure 12 The location of sample points and landmarks in Fukuoka Prefecture, indicated by black and red points, respectively. Brown lines are railway lines. The darker gray region indicates Hakata bay. (a) The actual location in 2014, (b) sample points located without any restriction, (c) sample points extracted randomly from their original location, (d) sample points obtained after optimization. The maximum correlation r_{max} is 0.5 in Figure 12b and Figure 12d.

We then evaluate the location of sample points in a more realistic context. We built a regression model that describes the land price by both spatial and aspatial variables. The former consists of the

distance to three landmarks, while the latter includes lot size, building and land use regulations, and surrounding land use represented by ten variables. We compared two models, one is based on sample points selected randomly from original data, and the other utilizes the same number of sample points that is optimized in such a way that r_{max} is smaller than 0.5. Both models chose independent variables by the forward stepwise method.

Table 2 shows the result of model estimation, where ρ is the proportion of sample points selected from the original ones. Regression models in Tochigi Prefecture describe land price better than those in Fukuoka Prefecture at any ρ. The former utilize more distance variables, which implies that the distance to landmarks plays an important role in Tochigi Prefecture. Optimization of sample selection generally improves the model fitness in both regions; R-square and adjusted R-square increase in nine out of twelve cases. R-square significantly increases especially when the number of distance variables increase by optimization such as the cases of ρ=0.50, 0.60, and 0.70. Optimization of sample points decreases the correlation between distance variables, which permits regression models to evaluate more distance variables, and consequently, improves the model fitness.

The above discussion also indicates that the optimization of sample selection is not so effective when the distance to landmarks is not critical to land price. We confirm this by the result obtained in Fukuoka Prefecture, where optimization of sample selection are less effective than in Tochigi Prefecture. Optimization of sample selection is worth trying in any case, but we should keep in mind that it does not always greatly improve the accuracy of regression models.

Table 2 Summary statistics of estimated models. *F* value is significant at significance level 0.01 in all the models.

Tochigi Prefecture

ρ	Extraction	R^2	Adjusted R^2	F	Number of independent variables: Total	Distance
0.50	Random	0.747	0.734	54.906	7	1
	Optimal	0.782	0.769	57.921	8	2
0.60	Random	0.690	0.676	50.220	7	1
	Optimal	0.806	0.796	81.492	8	3
0.70	Random	0.803	0.793	83.091	8	1
	Optimal	0.822	0.813	94.152	9	2
0.80	Random	0.785	0.776	85.571	9	2
	Optimal	0.797	0.788	91.791	9	2
0.90	Random	0.774	0.767	102.995	7	1
	Optimal	0.752	0.743	80.582	9	2
0.95	Random	0.772	0.755	95.021	9	1
	Optimal	0.755	0.746	86.456	8	2

Fukuoka Prefecture

ρ	Extraction	R^2	Adjusted R^2	F	Number of independent variables: Total	Distance
0.50	Random	0.544	0.539	111.559	2	0
	Optimal	0.634	0.626	80.067	4	1
0.60	Random	0.526	0.518	61.954	2	1
	Optimal	0.534	0.530	129.067	4	1
0.70	Random	0.539	0.534	102.055	3	0
	Optimal	0.577	0.569	71.193	5	1
0.80	Random	0.523	0.518	110.025	3	1
	Optimal	0.532	0.528	114.170	3	1
0.90	Random	0.527	0.522	125.743	3	1
	Optimal	0.466	0.462	98.517	3	1
0.95	Random	0.518	0.514	128.392	3	1
	Optimal	0.519	0.515	128.426	3	1

Sample points of land price and real estate data are often located primarily in downtown areas as seen in the two examples. Such distributions inevitably contain point clusters, which is a primary source

of the correlation between distance variables. If data are already prepared, we should avoid local clusters by choosing sample points in such a way that they are distributed more uniformly than their original distribution. When we newly collect sample points, uniform distribution is a simple but effective option because it yields both positive and negative correlations between distance variables as discussed in Subsection 2.4.

2.6 Limited location of sample points

This subsection discusses the case where sample points can be located only in a limited region. This happens either when we choose sample location from a given set of locations or when we locate sample points in a given sample domain.

In the former case, we solve Problem CM in a discrete space, i.e., choose the locations that minimizes $\max\{|r_{jk}|\}$ from the given set. Since it is an integer programming problem, we employ a heuristic method to derive the optimal set of sample points.

In the latter case, we add a constraint to Problem CM that limit the location of sample points:

Problem CM' (Correlation Minimization):

$$\underset{\mathbf{p}_i, i\in\boldsymbol{N}}{\text{minimize}} \max\left\{\left|r_{jk}\right|\right\} \quad (j,k\in\boldsymbol{L}, k\neq j)$$

subject to $\forall i\in\boldsymbol{N}, \mathbf{p}_i \in H$,

where H indicates sample domain. Sample points are initially distributed according to the following density function:

$$g'(\mathbf{p}) = \frac{g(\mathbf{p})}{\int_{\mathbf{q}\in H} g(\mathbf{q})\,d\mathbf{q}} = \frac{\sum_j \sum_{k\neq j} f_{jk}(\mathbf{p})}{\sum_j \sum_{k\neq j} \int_{\mathbf{q}\in H} f_{jk}(\mathbf{q})\,d\mathbf{q}} \quad . \tag{10}$$

We move sample points from their initial location in a similar way as to that used in Subsection 2.2.

We apply this method to a numerical experiment. Three landmarks F_1, F_2, and F_3 are located at (0, 0), (3, 0), and (2, 2), respectively, in rectangular region Ψ bounded by four lines x=-5, x=8, y=-5, and y=7. We derive the optimal location of 1000 sample points in circular sample domain H_S of radius l. The center of H_S moves at lattice points of interval 0.05 in Ψ. The radius l varies from 0.5 to 2.5 at the interval of 0.5.

Figure 13 shows the relationship between $r_{\max}$ and the location of the center of H_S when l=0.5,

1.5, and 2.5. The left hand figures show max{$|r_{jk}|$} when sample points are randomly distributed in H_S, while the right hand figures indicate r_{max} obtained after optimization. The figures clearly present that the optimization of sample location drastically reduces the correlation between distance variables.

Let us focus on the right hand figures of Figure 13. The figures show that we can keep r_{max} smaller than 0.5 when sample domain H_S is close to landmarks. The maximum correlation r_{max} increases as H_S moves away either inside or outside of $\Delta F_1F_2F_3$. This is consistent with the result shown in Figure 7b where the correlation between distance variables increases with the distance from landmarks. We can also reduce r_{max} by employing larger sample domain because it increases the flexibility of sample location. A careful consideration is necessary when H_S is located around the lines F_1F_2, F_3F_3, and F_1F_3. The maximum correlation r_{max} is small only when H_S is close to landmarks. Otherwise, the neighborhood of these lines yields worse result.

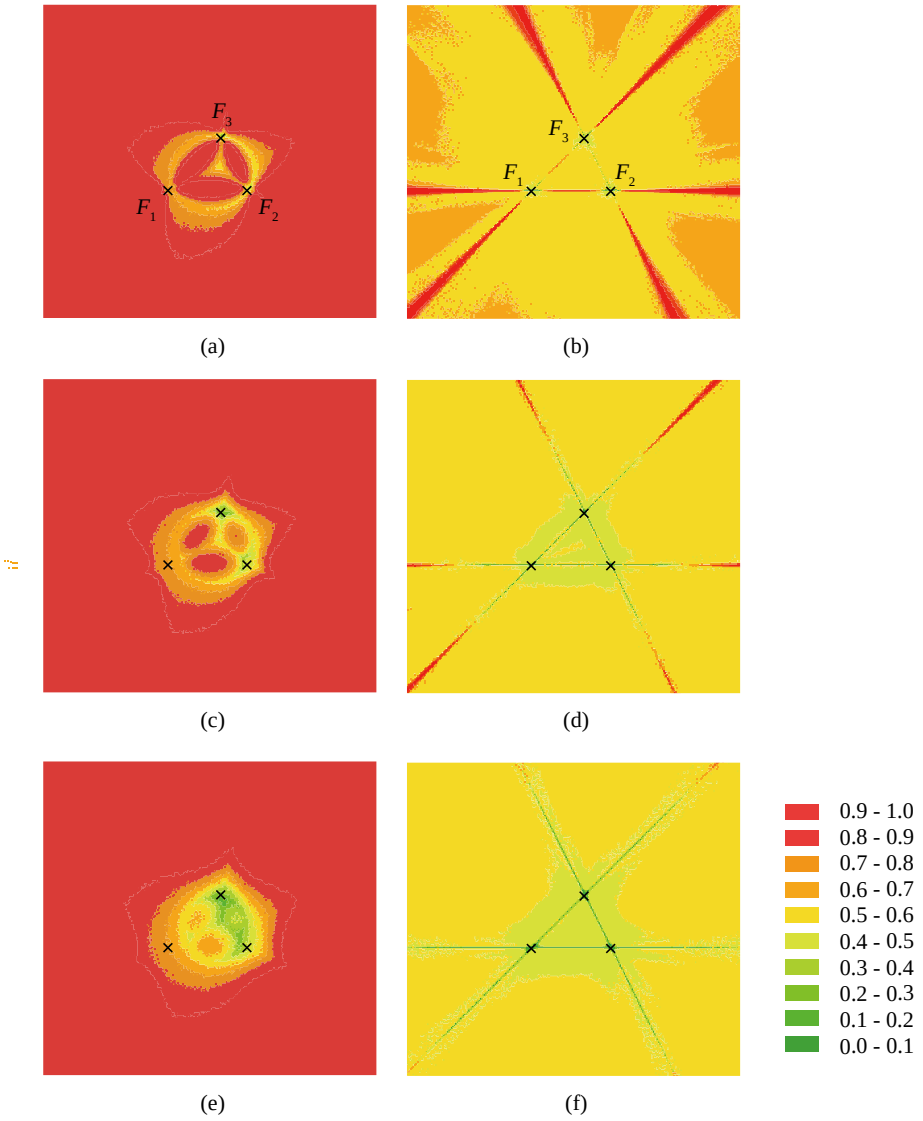


Figure 13 The relationship between $r_{\max}$ and the location of sample domain H_S. The left hand figures show $\max\{|r_{jk}|\}$ when sample points are randomly distributed in H_S, while the right hand figures indicate $r_{\max}$ obtained after optimization. The radius of H_S is (a)(b) 0.5, (c)(d) 1.5, and (e)(f) 2.5.

3. Concluding discussion

This paper developed a method for deriving the configuration of sample points that reduces the multicollinearity among distance variables in regression models. In the case of two landmarks, we derived the density function of sample points that totally vanishes the correlation between distance variables. We then proposed a mathematical procedure for locating sample points that minimizes the correlation between distances to more than two landmarks. To test the validity of the proposed method, we performed numerical experiments and empirical applications varying in the number of landmarks and sample domain. The results supported the technical soundness of the method, and provided useful implications for the design of sample location.

We finally indicate two limitations of the paper and potential extensions for future research.

First, we should discuss more thoroughly the extensions of the proposed method. Though Subsection 2.3 briefly describes some directions, we have not yet test their validity through numerical experiments and real applications. Computational cost and initial configuration of sample points in those extensions clearly need a detailed examination from both theoretical and empirical perspectives. Applications in various situations are also necessary to find an appropriate approach that is theoretically sound and practically feasible.

Second, the definition and treatment of distance variables need further discussion. This paper defines the distance as the Euclidean distance in a linear form in regression models. However, other measures such as the Manhattan distance, network distance and time distance are often used in regional science and geography. We can incorporate these distances into regression models in various forms including square root, inverse, logarithm, and exponential functions. We should extend the proposed method to treat a wider variety of distance measures.

References

Asteriou, Dimitrios and Stephen G Hall. 2007. *Applied Econometrics: a modern approach using eviews and microfit*: Palgrave Macmillan New York.

Avriel, Mordecai. 2003. *Nonlinear programming: analysis and methods*: Courier Dover Publications.

Bachi, Roberto. 1962. "Standard distance measures and related methods for spatial analysis." *Papers in Regional Science* 10, 83-132.

Belsley, David A, Edwin Kuh and Roy E Welsch. 1980. *Regression diagnostics: Identifying influential data and sources of collinearity*: John Wiley & Sons.

Bender, Bruce and Hae-Shin Hwang. 1985. "Hedonic housing price indices and secondary employment

centers." *Journal of Urban Economics* 17, 90-107.

Berry, Brian J. L. 1976. "Ghetto expansion and single-family housing prices: Chicago, 1968–1972." *Journal of Urban Economics* 3, 397-423.

Boland, Ralph P and Jorge Urrutia. 2001. "Finding the Largest Axis-Aligned Rectangle in a Polygon in O(n log n) time," paper presented at Proc. 13th Canad. Conf. Comput. Geom.

Burt, James E, Gerald M Barber and David L Rigby. 2009. *Elementary statistics for geographers*: Guilford Press.

Chatterjee, Samprit and Ali S Hadi. 2013. *Regression analysis by example*: John Wiley & Sons.

Chen, Gikuang Jeff. 2012. "A simple way to deal with multicollinearity." *Journal of Applied Statistics* 39, 1893-1909.

Curto, José Dias and José Castro Pinto. 2007. "New Multicollinearity Indicators in Linear Regression Models." *International Statistical Review* 75, 114-121.

Daniels, Karen, Victor Milenkovic and Dan Roth. 1997. "Finding the largest area axis-parallel rectangle in a polygon." *Computational Geometry* 7, 125-148.

Davis, Philip J and Philip Rabinowitz. 2007. *Methods of numerical integration*. New York: Dover.

Dewhurst, J.H.L. 1993. *Spatial Multicollinearity and Sample Selection in Models with Inverse Distance Measures*: University of Dundee, Department of Economics and Management.

Dormann, Carsten F, Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, Jaime R García Marquéz, Bernd Gruber, Bruno Lafourcade and Pedro J Leitão. 2013. "Collinearity: a review of methods to deal with it and a simulation study evaluating their performance." *Ecography* 36, 027-046.

Farrar, Donald E and Robert R Glauber. 1967. "Multicollinearity in regression analysis: the problem revisited." *The Review of Economic and Statistics*, 92-107.

Fletcher, Roger. 2013. *Practical methods of optimization*: John Wiley & Sons.

Gordon, Alan G and Eville Gorham. 1963. "Ecological aspects of air pollution from an iron-sintering plant at Wawa, Ontario." *Canadian Journal of Botany* 41, 1063-1078.

Hamacher, Horst W and Zvi Drezner. 2002. *Facility location: applications and theory*: Springer.

Harrison Jr, David and Daniel L Rubinfeld. 1978. "Hedonic housing prices and the demand for clean air." *Journal of Environmental Economics and Management* 5, 81-102.

Heikkila, Eric. 1988. "Multicollinearity in Regression Models with Multiple Distance Measures." *Journal of Regional Science* 28, 345-362.

Ihlanfeldt, Keith R and Laura O Taylor. 2004. "Externality effects of small-scale hazardous waste sites: evidence from urban commercial property markets." *Journal of Environmental Economics and Management* 47, 117-139.

Jilcott Pitts, Stephanie B, Qiang Wu, Jared T McGuirt, Thomas W Crawford, Thomas C Keyserling and Alice S Ammerman. 2013. "Associations between access to farmers' markets and supermarkets, shopping patterns, fruit and vegetable consumption and health indicators among women of reproductive age in

eastern North Carolina, USA." *Public health nutrition* 16, 1944-1952.
Karlof, John K. 2005. *Integer programming: theory and practice*: CRC Press.
Kashid, D. N. and S. R. Kulkarni. 2002. "A more genral criterion for subset selection in multiple linear regression." *Communications in Statistics - Theory and Methods* 31, 795-811.
Kovàcs, Péter, Tibor Petres and László Tóth. 2005. "A New Measure of Multicollinearity in Linear Regression Models." *International Statistical Review* 73, 405-412.
Li, Mingche M and H James Brown. 1980. "Micro-neighborhood externalities and hedonic housing prices." *Land economics*, 125-141.
Mansfield, Edward R, John T Webster and Richard F Gunst. 1977. "An analytic variable selection technique for principal component regression." *Applied statistics*, 34-40.
Miller, Alan. 2012. *Subset selection in regression*: CRC Press.
Morland, Kimberly, Ana V Diez Roux and Steve Wing. 2006. "Supermarkets, other food stores, and obesity: the atherosclerosis risk in communities study." *Am J Prev Med* 30, 333-339.
Nemhauser, George L and Laurence A Wolsey. 1988. *Integer and combinatorial optimization*: Wiley New York.
Ni, Liqiang. 2011. "Principal component regression revisited." *Statistica Sinica* 21, 741.
Noonan, Douglas S., Douglas J. Krupka and Brett M. Baden. 2007. "Neighborhood dynamics and price effects of superfund site clean-up." *Journal of Regional Science* 47, 665-692.
Pielou, E. C. 1977. *Mathematical Ecology*. New York: Wiley.
Riga-Karandinos, AN and MG Karandinos. 1998. "Assessment of air pollution from a lignite power plant in the plain of Megalopolis (Greece) using as biomonitors three species of lichens; impacts on some biochemical parameters of lichens." *Science of the total environment* 215, 167-183.
Ripley, Brian D. 2005. *Spatial statistics*: John Wiley & Sons.
Rogerson, Peter and Ikuho Yamada. 2008. *Statistical detection and surveillance of geographic clusters*: CRC Press.
Rundle, Andrew, Kathryn M Neckerman, Lance Freeman, Gina S Lovasi, Marnie Purciel, James Quinn, Catherine Richards, Neelanjan Sircar and Christopher Weiss. 2009. "Neighborhood food environment and walkability predict obesity in New York City." *Environ Health Perspect* 117, 442-447.
Sadahiro, Yukio and Yan Wang. 2015. "Configuration of sample points for the reduction of multicollinearity in regression models with distance variables," *Discussion Paper Series*. Center for Spatial Information Science, The University of Tokyo.
Snyman, Jan. 2005. *Practical mathematical optimization: an introduction to basic optimization theory and classical and new gradient-based algorithms*: Springer.
Spanos, Aris and Anya McGuirk. 2002. "The problem of near-multicollinearity revisited: erratic vs systematic volatility." *Journal of Econometrics* 108, 365-393.
Vigneau, E, MF Devaux, EM Qannari and P Robert. 1997. "Principal component regression, ridge regression and ridge principal component regression in spectroscopy calibration." *Journal of chemometrics* 11,

239-249.

Weissfeld, Lisa A and Susan M Sereika. 1991. "A multicollinearity diagnostic for generalized linear models." *Communications in Statistics-Theory and Methods* 20, 1183-1198.

Wolsey, Laurence A. 1998. *Integer programming*: Wiley New York.

Yu, KN, YP Cheung, T Cheung and Ronald C Henry. 2004. "Identifying the impact of large urban airports on local air quality by nonparametric regression." *Atmospheric Environment* 38, 4501-4507.