CSIS Discussion Paper No. **138**

Multicollinearity in regression models with distance variables: application of linear transformation

Yukio Sadahiro

May 2015

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

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 discusses the linear transformation of distance variables as a solution for multicollinearity. A general linear transformation including principal component analysis is considered.

1. Method and applications

Suppose a spatial phenomenon represented as a continuous function $f(\mathbf{x})$ defined on a two-dimensional plane Ξ. The function is measured at M sample points. Let P_i, $\mathbf{p}_i$, and f_i be ith sample point, its location, and the function value observed at P_i, respectively, where $i \in \boldsymbol{M} = \{1, 2, ..., M\}$. Function $y(\mathbf{x})$ is determined by both spatial and aspatial factors, the former of which include the distance to urban facilities such as schools, railway stations, and urban parks, which we call *landmarks*. Let L_j and $\mathbf{z}_j$ be jth landmark and its location, respectively ($j \in \boldsymbol{N} = \{1, 2, ..., N\}$). The coordinates of P_j and L_j are represented as (x_j, y_j) and (u_j, v_j), respectively. The distance between location $\mathbf{p}$ and $\mathbf{z}_j$ is denoted by $d(\mathbf{p}, \mathbf{z}_j)$.

We build a regression model that explains $f(\mathbf{x})$ by its determinants based on the data observed at sample points. We omit aspatial factors in the model for the present to focus on the multicollinearity among distance variables:

$$f(\mathbf{p}_i) = \beta_0 + \beta_1 d(\mathbf{p}_i, \mathbf{z}_1) + \beta_2 d(\mathbf{p}_i, \mathbf{z}_2) + ... + \beta_N d(\mathbf{p}_i, \mathbf{z}_N), \quad (1)$$

where β_k's are the parameters to be estimated ($k \in \boldsymbol{N}$). The multicollinearity among distance variables is unavoidable since the landmarks are distributed on the same two-dimensional space. Let r_{jk} be the absolute correlation coefficient between D_j' and D_k' ($j \neq k$). The correlation coefficients are an indicator of multicollinearity that can be easily calculated and interpreted.

2. Two landmarks

This section considers the case of two landmarks. We employ a simplified representation of Equation (17):

$$f = \beta_0 + \beta_1 d_1 + \beta_2 d_2. \quad (2)$$

To consider the relationship between d_1 and d_2, let us suppose XY-coordinate system as shown in Figure 1a, where the location of two landmarks L_1 and L_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_1 and d_2 as shown in Figure 1b. We call the latter *d_1d_2-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 we call *possible sample region*. It is bounded by three lines $d_1-d_2=-2c$, $d_1-d_2=2c$, and $d_1+d_2=2c$.

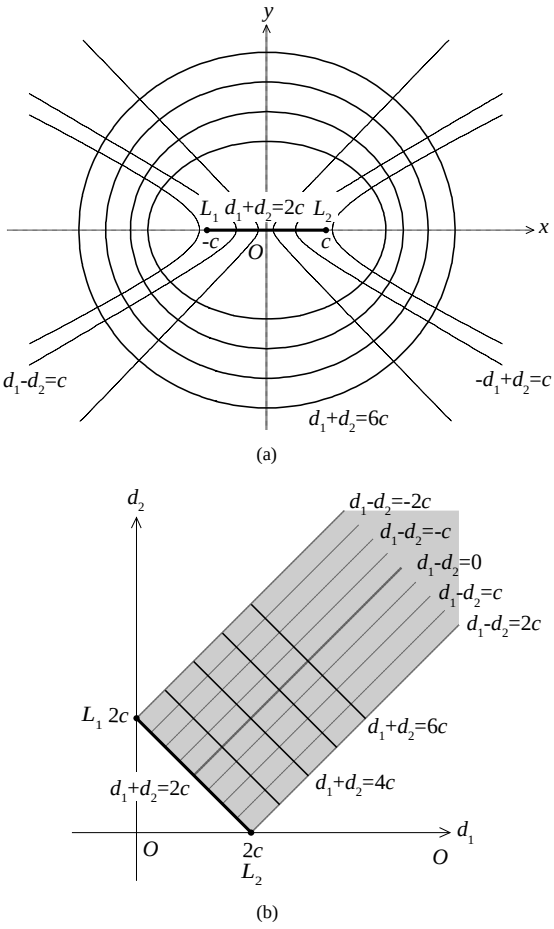


Figure 1 The relationship between the location of sample points in the real space and that in the distance spaces. The same symbols in the two figures indicate the same objects. (a) The real space. Landmarks L_1 and L_2, ellipses and hyperbolas whose foci are L_1 and L_2 are drawn. (b) The possible sample region in the d_1d_2-space.

The lines shown in Figure 1b are represented mathematically as

$$d_1+d_2=\begin{cases} 2c & \Rightarrow \quad y=0\left(-c\le x\le c\right) \\ K\left(2c<K\right) & \Rightarrow \quad \dfrac{x^2}{\dfrac{K^4}{4\left(K^2-4\right)}}+\dfrac{y^2}{\dfrac{K^2}{4}}=1 \end{cases} \tag{3}$$

and

$$d_1 - d_2 = \begin{cases} 2c & \Rightarrow \quad y = 0\left(c < x\right) \\ K\left(0 < K < 2c\right) & \Rightarrow \quad \dfrac{x^2}{\dfrac{K^4}{4\left(K^2-4\right)}} - \dfrac{y^2}{\dfrac{K^2}{4}} = 1 \\ 0 & \Rightarrow \quad x = 0 \\ K\left(-2c < K < 0\right) & \Rightarrow \quad -\dfrac{x^2}{\dfrac{K^4}{4\left(K^2-4\right)}} + \dfrac{y^2}{\dfrac{K^2}{4}} = 1 \\ -2c & \Rightarrow \quad y = 0\left(x < -c\right) \end{cases}.$$

(4)

Figure 1b clearly indicates that a serious correlation occurs when sample points are distributed randomly over plane Ξ. They are densely clustered in the gray-shaded in the figure, which leads to a high positive correlation between d_1 and d_2. One method to avoid this is to locate sample points are distributed symmetrically in either vertical or horizontal direction in the distance space (Sadahiro and Wang 2015). This, however, limits the sample region to a relatively small area. For instance, if we locate sample points in rectangle bounded by

$$\begin{aligned} d_1 + d_2 &= 2c \\ d_1 + d_2 &= 6c \\ d_1 - d_2 &= 2c \\ d_1 - d_2 &= -2c \end{aligned},$$

(5)

the corresponding sample region in the real space is the outermost ellipse in Figure 1a defined by

$$\frac{x^2}{\dfrac{81c^4}{9c^2-1}} + \frac{y^2}{9c^2} = 1.$$

(6)

We can avoid this problem is to apply a linear transformation to the distance variables in such a way that the possible sample region extends in parallel with either the vertical or horizontal axis. Let **A** be a matrix that represents a linear transformation:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{21} \end{pmatrix}.$$

(7)

The distance variables are transformed as

$$\begin{pmatrix} D_1 \\ D_2 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{21} \end{pmatrix} \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} = \begin{pmatrix} a_{11}d_1 + a_{12}d_2 \\ a_{21}d_1 + a_{22}d_2 \end{pmatrix}.$$

(8)

Since-$2c \leq d_1 - d_2 \leq 2c$, necessary condition that the possible sample region extends in parallel with the vertical axis is

$$a_{11} + a_{12} = 0,$$

(9)

which implies

$$D_1 = k\left(d_1 + d_2\right).$$

(10)

Condition for the horizontal direction is

$$a_{21} + a_{22} = 0.$$

(11)

Both cases consider the difference between d_1 and d_2.

Numerous transformations satisfy the above conditions. We thus should choose an appropriate one that yields an interpretable result. This paper proposes two methods, one is we call *average distance* method, and the other is *incremental distance* method developed by Partridge, Olfert, and Alasia (2007), Partridge et al. (2008), and Partridge et al. (2008).

The average distance method rotates d_1 and d_2 by $\pi/4$ around the origin either clockwise or counterclockwise so that the possible sample region is in parallel with either the horizontal or vertical axis. The clockwise rotation is

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{21} \end{pmatrix} = \frac{\sqrt{2}}{2} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix},$$

(12)

which yields

$$\begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix} = \frac{\sqrt{2}}{2} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} = \frac{\sqrt{2}}{2} \begin{pmatrix} d_1 + d_2 \\ d_1 - d_2 \end{pmatrix}.$$

(13)

It is an orthogonal transformation that keeps the vector length in the d_1d_2-space. Figure 2a shows the result of this transformation. We can easily locate sample points symmetrically with respect to the δ_1-axis in the possible sample region. This implies that sample points can be located far away from the landmarks without high correlation. The figure even suggests that no serious correlation would occur even if we randomly distribute sample points over plane Ξ.

The incremental distance method assumes a regression model incorporating the distance to urban facilities that provide different levels of services. Assume that the ith level tier provides services of levels from 1 to i. The incremental distance first calculate the distance to the lowest tier that provides and then calculate the additional distance to the tire of the next higher level. This process is repeated until the incremental distance to the highest tier is obtained. In the case of two facilities, distance variables are defined as

$$\begin{pmatrix} \varphi_1 \\ \varphi_2 \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 - d_1 \end{pmatrix} \tag{14}$$

and thus

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{21} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}. \tag{15}$$

Unlike the average distance, the incremental distance method is not an orthogonal transformation. It changes the shape of possible sample region as shown in Figure 2b. However, we can still easily locate sample points symmetrically with respect to line $\varphi_2=c$ by avoiding only the close neighborhood of F_1. We should note, however, that random distribution of sample points is not symmetrical in the φ_1-axis. Figure 2b shows the distribution of 2000 sample points in circle of radius $20c$ centered at the origin in $\varphi_1\varphi_2$-space. Point density is higher on the positive side of φ_2 than its negative side at each section of φ_1. This result in the correlation coefficient -0.118 while that of δ_1 and δ_2 of the same distribution is -0.012.

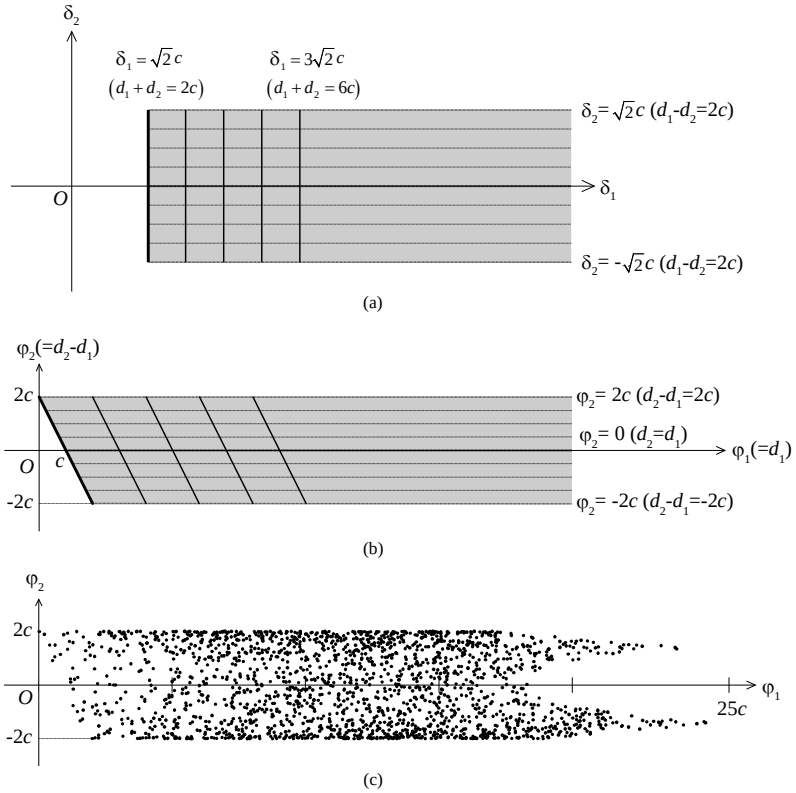


Figure 2 The average distance and incremental distance methods. (a) The possible sample region in the $\delta_1\delta_2$-space. (a) The possible sample region in the $\varphi_1\varphi_2$-space. (c) Sample points randomly distributed of radius 20c centered at the origin.

Each method has its own strengths. The average distance method has no limitation in the location of sample points while the incremental distance avoids the neighborhood of one of the two facilities. Random distribution of sample points cause a slight correlation between distance variables in the incremental distance method, though it is not serious as seen in the above example. One important strength of the incremental distance method is an integrated framework for the interpretation of results that is based on econometrics. Though each distance in the average distance method is interpretable separately, the significance of using them simultaneously is not clear. Distance δ_1 represents the average distance to the two facilities, which is interpretable as it is, while the meaning of δ_2 in relation to δ_1 is rather ambiguous.

3. Three landmarks

This section considers the case of three landmarks. We employ a simplified representation of Equation (17):

$$f = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3 .$$

(16)

3.1 Relationship between distance variables

This subsection discusses the relationship between distance variables. The distance from sample point P at (x, y) to L_j is

$$d_i = \sqrt{\left(x - u_i\right)^2 + \left(y - v_i\right)^2} .$$

(17)

Figure 3 illustrates the function representing the relationship between distance variables d_1, d_2 and d_3. As seen in the figure, a high correlation often exists between distance variables. It appears especially when landmarks are located away from sample points.

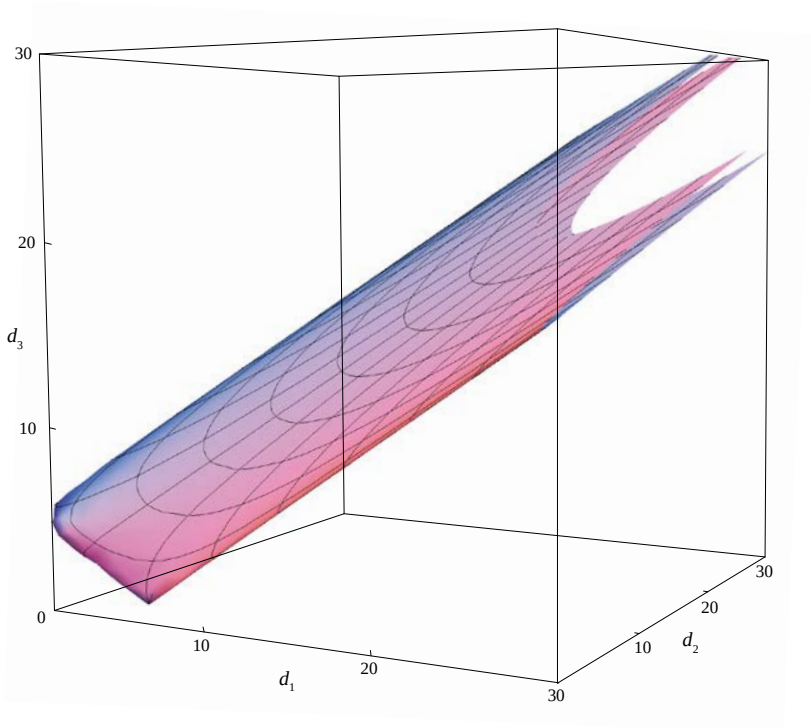


Figure 3 The relationship between distance variables d_1, d_2, and d_3.

One method to reduce the correlation is to apply a linear transformation including principal component analysis to distance variables. Let **A** be a transformation matrix defined by

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}.$$

(18)

This yields

$$\begin{pmatrix} D_1 \\ D_2 \\ D_3 \end{pmatrix} = \mathbf{A} \begin{pmatrix} d_1 \\ d_2 \\ d_3 \end{pmatrix} = \begin{pmatrix} a_{11}d_1 + a_{12}d_2 + a_{13}d_3 \\ a_{21}d_1 + a_{22}d_2 + a_{23}d_3 \\ a_{31}d_1 + a_{32}d_2 + a_{33}d_3 \end{pmatrix}.$$

(19)

We assume that sample point P is located far from the landmarks permits us to approximate distances d_i as

$$\begin{aligned} d_i &= \sqrt{\left(x - u_i\right)^2 + \left(y - v_i\right)^2} \\ &= \sqrt{x^2 + y^2}\sqrt{1 + \frac{-2u_i x - 2v_i y + u_i^2 + v_i^2}{x^2 + y^2}} \\ &\approx \sqrt{x^2 + y^2}\left(1 - \frac{u_i x + v_i y}{x^2 + y^2}\right) \\ &= \sqrt{x^2 + y^2} - \frac{u_i x + v_i y}{\sqrt{x^2 + y^2}} \end{aligned}.$$

(20)

Then we have

$$\begin{pmatrix} D_1 \\ D_2 \\ D_3 \end{pmatrix} = \begin{pmatrix} a_{11}d_1 + a_{12}d_2 + a_{13}d_3 \\ a_{21}d_1 + a_{22}d_2 + a_{23}d_3 \\ a_{31}d_1 + a_{32}d_2 + a_{33}d_3 \end{pmatrix}$$

$$\approx \begin{pmatrix} \left(a_{11} + a_{12} + a_{13}\right)\sqrt{x^2 + y^2} - \frac{1}{\sqrt{x^2 + y^2}}\left\{\left(a_{11}u_1 + a_{12}u_2 + a_{13}u_3\right)x + \left(a_{11}v_1 + a_{12}v_2 + a_{13}v_3\right)y\right\} \\ \left(a_{21} + a_{22} + a_{23}\right)\sqrt{x^2 + y^2} - \frac{1}{\sqrt{x^2 + y^2}}\left\{\left(a_{21}u_1 + a_{22}u_2 + a_{23}u_3\right)x + \left(a_{21}v_1 + a_{22}v_2 + a_{23}v_3\right)y\right\} \\ \left(a_{31} + a_{32} + a_{33}\right)\sqrt{x^2 + y^2} - \frac{1}{\sqrt{x^2 + y^2}}\left\{\left(a_{31}u_1 + a_{32}u_2 + a_{33}u_3\right)x + \left(a_{31}v_1 + a_{32}v_2 + a_{33}v_3\right)y\right\} \end{pmatrix}.$$

(21)

Let R be the distance from the origin to P, i.e.,

$$R = \sqrt{x^2 + y^2} \, .$$

(22)

Equation (21) then becomes

$$\begin{pmatrix} D_1 \\ D_2 \\ D_3 \end{pmatrix} \approx \begin{pmatrix} \left(a_{11}+a_{12}+a_{13}\right)R - \dfrac{1}{R}\left\{\left(a_{11}u_1+a_{12}u_2+a_{13}u_3\right)x + \left(a_{11}v_1+a_{12}v_2+a_{13}v_3\right)y\right\} \\ \left(a_{21}+a_{22}+a_{23}\right)R - \dfrac{1}{R}\left\{\left(a_{21}u_1+a_{22}u_2+a_{23}u_3\right)x + \left(a_{21}v_1+a_{22}v_2+a_{23}v_3\right)y\right\} \\ \left(a_{31}+a_{32}+a_{33}\right)R - \dfrac{1}{R}\left\{\left(a_{31}u_1+a_{32}u_2+a_{33}u_3\right)x + \left(a_{31}v_1+a_{32}v_2+a_{33}v_3\right)y\right\} \end{pmatrix} .$$

(23)

The above equation can be rewritten as

$$\begin{pmatrix} D_1 \\ D_2 \\ D_3 \end{pmatrix} \approx \begin{pmatrix} K_1R + b_1x + c_1y \\ K_2R + \dfrac{b_2}{R}x + c_2y \\ K_3R + \dfrac{b_3}{R}x + c_3y \end{pmatrix},$$

(24)

where

$$\begin{aligned} K_i &= a_{i1} + a_{i2} + a_{i3} \\ b_i &= a_{i1}u_1 + a_{i2}u_2 + a_{i3}u_3 \, . \\ c_i &= a_{i1}v_1 + a_{i2}v_2 + a_{i3}v_3 \end{aligned}$$

(25)

Let us consider the conditions to avoid a serious correlation between the distance variables. Figure 3 suggests that if the location of sample points is not limited to a small region, a correlation occurs if two or more variables increases similarly with R. We can avoid this problem by transforming the distances in such a way that the possible sample region is parallel with one of the distance axes. This implies that only one distance increases with R while the other two are approximately stable against R. Assuming that only D_1 increase with R, a condition to avoid correlation is

$$a_{11} + a_{12} + a_{13} \neq 0$$

(26)

and

$$a_{21}+a_{22}+a_{23}=a_{31}+a_{32}+a_{33}=0.$$

(27)

In this case, since the second and third terms in Equation (24) are negligible compared with the first term, we obtain

$$\begin{pmatrix} D_1 \\ D_2 \\ D_3 \end{pmatrix} \approx \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} d_1 \\ d_2 \\ d_3 \end{pmatrix}, \\ = R \begin{pmatrix} K \\ b_2 x + c_2 y \\ b_3 x + c_3 y \end{pmatrix}$$

(28)

where

$$K = \frac{a_{11}+a_{12}+a_{13}}{R}.$$

(29)

Unlike the two-facility case discussed in the previous section, the three-facility case requires us to consider the three correlations between distance variables, i.e., r_{12}, r_{13}, and r_{23}. Since the above condition provides us a means of reducing r_{12} and r_{13}, we now consider the reduction of r_{23}. To this end, let us consider the section of possible sample region along the D_1-axis. Assuming b_2C_3-$b_3C_2 \neq 0$, we obtain

$$\begin{pmatrix} x \\ y \end{pmatrix} = \frac{1}{b_2c_3 - b_3c_2} \begin{pmatrix} c_3 & -c_2 \\ -b_3 & b_2 \end{pmatrix} \begin{pmatrix} D_2 \\ D_3 \end{pmatrix} \\ = \frac{1}{b_2c_3 - b_3c_2} \begin{pmatrix} c_3D_2 - c_3D_3 \\ -b_3D_2 + b_2D_3 \end{pmatrix}.$$

(30)

Substitution of the above equation into Equation (22) yields

$$x^2 + y^2 = 1 \\ = \frac{\left(b_3^2 + c_3^2\right)D_2^2 + \left(b_2^2 + c_2^2\right)D_3^2 - 2\left(b_2b_3 + c_2c_3\right)D_2D_3}{\left(b_2c_3 - b_3c_2\right)^2},$$

(31)

and consequently,

$$\frac{\left(b_3^2+c_3^2\right)D_2^2+\left(b_2^2+c_2^2\right)D_3^2-2\left(b_2b_3+c_2c_3\right)D_2D_3}{\left(b_2c_3-b_3c_2\right)^2}=1.$$

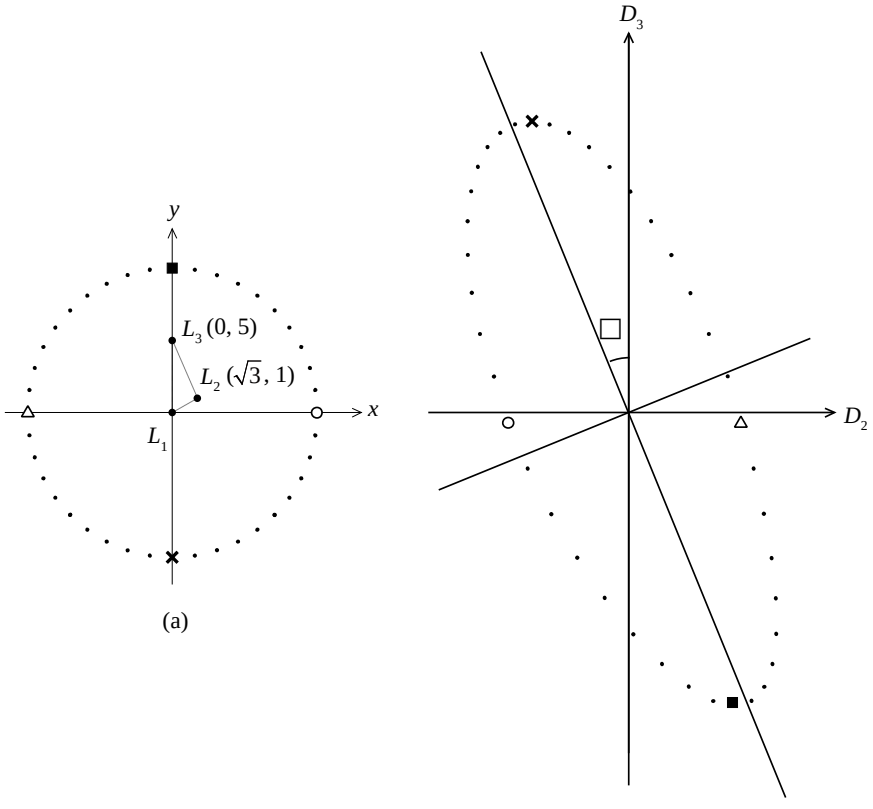
(32)

The above equation indicates that D_2 and D_3 form an ellipse centered at the origin. This paper calls this function D_2D_3-*function* hereafter. It depends only on the location of landmarks and matrix **A**, i.e., is independent the distance from the origin. Figure 4 shows the relationship between the sample points in the real space and those in the D_2D_3-space, where

$$\begin{pmatrix} u_1 & v_1 \\ u_2 & v_2 \\ u_3 & v_3 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ \sqrt{3} & 1 \\ 0 & 5 \end{pmatrix}.$$

(33)

The figure clearly indicates that the location of sample points in the D_2D_3-space only depends on the direction in the real space. A one-to-one correspondence exists between the direction in the real space and that in the D_2D_3-space.



(a) (b)

Figure 4 The relationship between the sample points in the real space and those in the distance space. (a) The real space, (b) the D_2D_3-space.

The length of the major and minor axes of the ellipse is represented as

$$\begin{pmatrix} E_1 \\ E_2 \end{pmatrix} = \sqrt{2}\left|b_2c_3 - b_3c_2\right| R \begin{pmatrix} \dfrac{1}{\sqrt{b_2^2+b_3^2+c_2^2+c_3^2-\sqrt{\left(b_2^2-b_3^2+c_2^2-c_3^2\right)^2+4\left(b_2b_3+c_2c_3\right)^2}}} \\ \dfrac{1}{\sqrt{b_2^2+b_3^2+c_2^2+c_3^2+\sqrt{\left(b_2^2-b_3^2+c_2^2-c_3^2\right)^2+4\left(b_2b_3+c_2c_3\right)^2}}} \end{pmatrix}. \tag{34}$$

The rotation angle is

$$\begin{aligned} \vartheta &= \frac{1}{2}\arctan\left(\frac{-\left(2c_2c_3+2b_2b_3\right)}{b_2^2-b_3^2+c_2^2-c_3^2}\right) \\ &= -\frac{1}{2}\arctan\left(\frac{2\left(b_2b_3+c_2c_3\right)}{b_2^2-b_3^2+c_2^2-c_3^2}\right) \end{aligned}. \tag{35}$$

Figure 5 illustrates the relationship between D_1, D_2, and D_3 in the distance space.

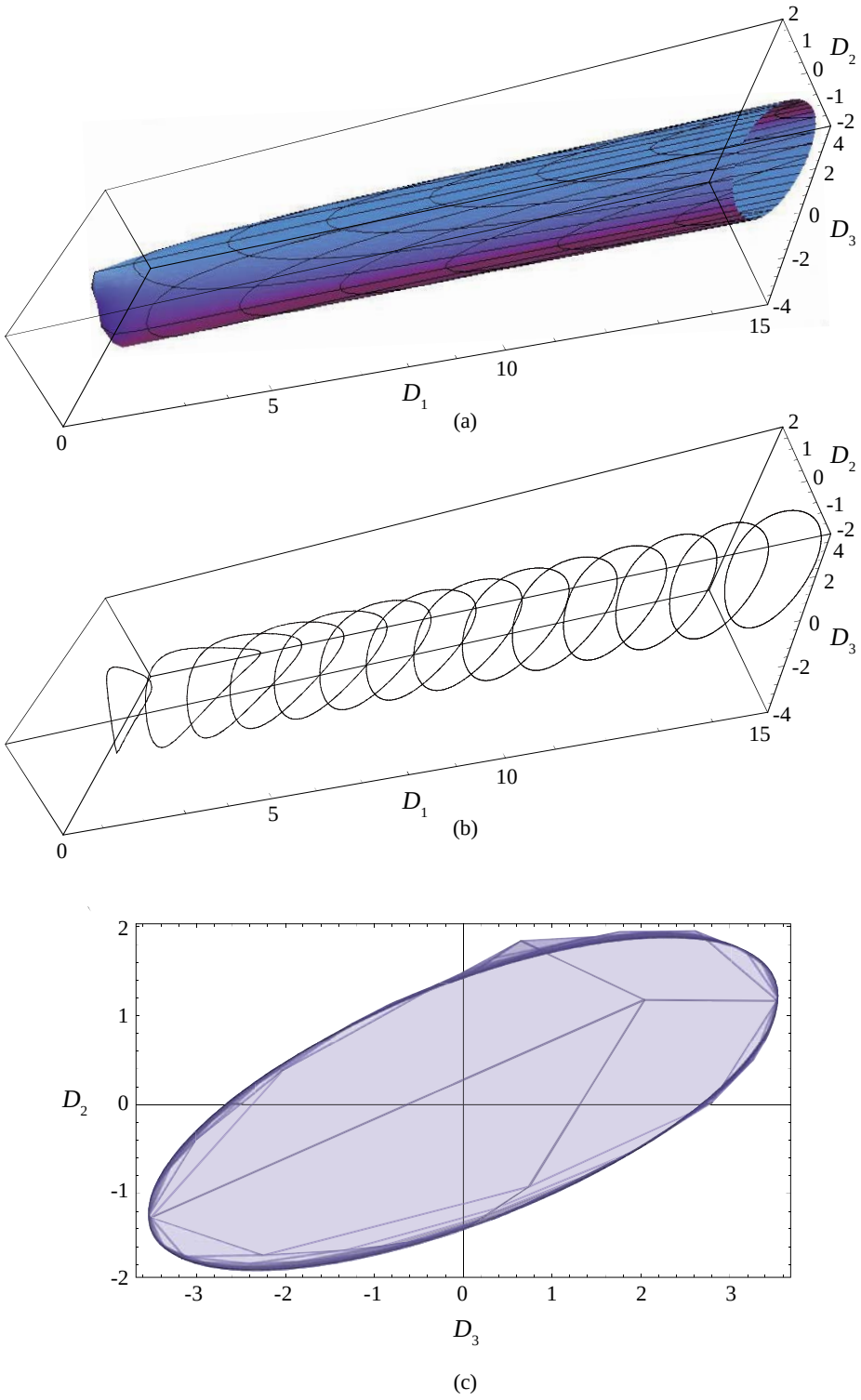


Figure 5 The relationship between D_1, D_2, and D_3 in the distance space. (a) Three dimensional representation, (b) sections along the D_1-axis, (c) Projection onto the D_2-D_3 plane.

3.2 Transformation methods

This subsection discusses the properties of the two transformations mentioned in the previous sections.

The average distance method transforms the distance variables in such a way that D_1 represents the average distance to the three facilities while the D_2D_3-function is symmetrical with respect to the D_2- and D_3-axes. This allows us to locate sample points symmetrical with respect to the D_2- and D_3-axes, which does not cause a serious correlation. To derive such a transformation, we first consider the rotation that makes the line $d_1=d_2=d_3$ coincide with the D_1-axis. The rotation matrices are given by

$$R_2 = \begin{pmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{pmatrix} \tag{36}$$

and

$$R_3 = \begin{pmatrix} \cos\varphi & -\sin\varphi & 0 \\ \sin\varphi & \cos\varphi & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{37}$$

respectively. The vector (1, 1, 1) is converted into

$$R_3R_2\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} \cos\varphi\left(\sin\theta+\cos\theta\right)-\sin\varphi \\ \sin\varphi\left(\sin\theta+\cos\theta\right)+\cos\varphi \\ -\sin\theta+\cos\theta \end{pmatrix}.$$

To meet

$$\begin{pmatrix} \cos\varphi\left(\sin\theta+\cos\theta\right)-\sin\varphi \\ \sin\varphi\left(\sin\theta+\cos\theta\right)+\cos\varphi \\ -\sin\theta+\cos\theta \end{pmatrix} = \begin{pmatrix} \sqrt{3} \\ 0 \\ 0 \end{pmatrix}, \tag{38}$$

angles need to satisfy

$$\theta = \frac{\pi}{4}$$
$$\tan\varphi = -\frac{\sqrt{2}}{2}, \cos\varphi = \frac{\sqrt{6}}{3}, \sin\varphi = -\frac{\sqrt{3}}{3}. \tag{39}$$

We can confirm

$$\begin{pmatrix} \cos\varphi\left(\sin\theta+\cos\theta\right)-\sin\varphi \\ \sin\varphi\left(\sin\theta+\cos\theta\right)+\cos\varphi \\ -\sin\theta+\cos\theta \end{pmatrix} = \begin{pmatrix} \sqrt{2}\cos\varphi-\sin\varphi \\ \sqrt{2}\sin\varphi+\cos\varphi \\ 0 \end{pmatrix} = \begin{pmatrix} \sqrt{3} \\ 0 \\ 0 \end{pmatrix}. \tag{40}$$

Point at (d_1, d_2, d_3) is then converted into

$$\begin{aligned} \begin{pmatrix} D_1' \\ D_2' \\ D_3' \end{pmatrix} &= R_3 R_2 \begin{pmatrix} d_1 \\ d_2 \\ d_3 \end{pmatrix} \\ &= \begin{pmatrix} \frac{\sqrt{3}}{3}d_1+\frac{\sqrt{3}}{3}d_2+\frac{\sqrt{3}}{3}d_3 \\ -\frac{\sqrt{6}}{6}d_1+\frac{\sqrt{6}}{3}d_2-\frac{\sqrt{6}}{6}d_3 \\ -\frac{\sqrt{2}}{2}d_1+\frac{\sqrt{2}}{2}d_3 \end{pmatrix}. \end{aligned} \tag{41}$$

The final result is thus given by

$$\begin{aligned} \begin{pmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{pmatrix} &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\vartheta & \sin\vartheta \\ 0 & -\sin\vartheta & \cos\vartheta \end{pmatrix} \begin{pmatrix} \frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} \\ -\frac{\sqrt{6}}{6} & \frac{\sqrt{6}}{3} & -\frac{\sqrt{6}}{6} \\ -\frac{\sqrt{2}}{2} & 0 & \frac{\sqrt{2}}{2} \end{pmatrix} \begin{pmatrix} d_1 \\ d_2 \\ d_3 \end{pmatrix} \\ &= \frac{1}{6} \begin{pmatrix} 2\sqrt{6} & 2\sqrt{6} & 2\sqrt{6} \\ -\sqrt{6}\cos\vartheta-3\sqrt{2}\sin\vartheta & 2\sqrt{6}\cos\vartheta & -\sqrt{6}\cos\vartheta+3\sqrt{2}\sin\vartheta \\ -\sqrt{6}\sin\vartheta-3\sqrt{2}\cos\vartheta & -2\sqrt{6}\sin\vartheta & \sqrt{6}\sin\vartheta+3\sqrt{2}\cos\vartheta \end{pmatrix} \begin{pmatrix} d_1 \\ d_2 \\ d_3 \end{pmatrix} \end{aligned}. \tag{42}$$

Since

$$
\begin{aligned}
b_2 &= \frac{a_{21}u_1 + a_{22}u_2 + a_{23}u_3}{R} \\
&= \frac{1}{R}\left(-\frac{\sqrt{6}}{6}u_1 + \frac{\sqrt{6}}{3}u_2 - \frac{\sqrt{6}}{6}u_3 \right) \\
b_3 &= \frac{a_{21}u_1 + a_{22}u_2 + a_{23}u_3}{R} \\
&= \frac{1}{R}\left(-\frac{\sqrt{2}}{2}u_1 + \frac{\sqrt{2}}{2}u_3 \right) \qquad , \\
c_2 &= \frac{1}{R}\left(-\frac{\sqrt{6}}{6}v_1 + \frac{\sqrt{6}}{3}v_2 - \frac{\sqrt{6}}{6}v_3 \right) \\
c_3 &= \frac{1}{R}\left(-\frac{\sqrt{2}}{2}v_1 + \frac{\sqrt{2}}{2}v_3 \right)
\end{aligned}
$$

(43)

we substitute these equations into Equation (35) to calculate the rotation angle ϑ. We then substitute it into Equation (48), we can calculate the transformation matrix of the average distance method $\mathbf{A}_A$ that makes the D_2D_3-function is symmetrical with respect to the D_2- and D_3-axes.

We then discuss the incremental distance method. Since it is defined as

$$
\begin{pmatrix} \varphi_1 \\ \varphi_2 \\ \varphi_3 \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 - d_1 \\ d_3 - d_2 \end{pmatrix},
$$

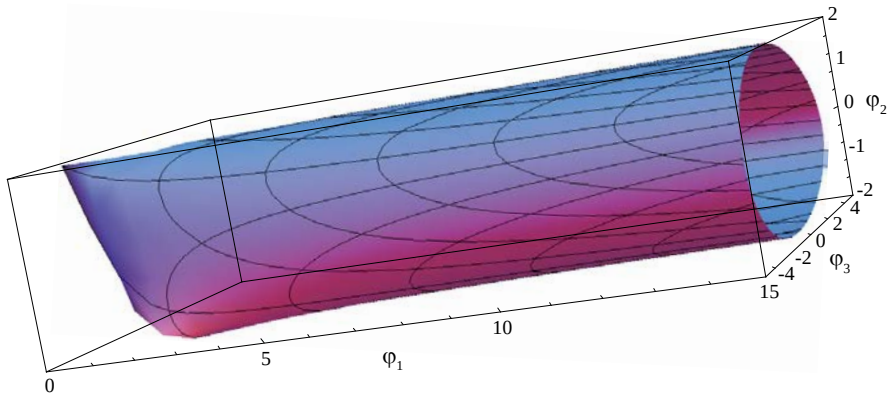
(44)

transformation matrix is represented as

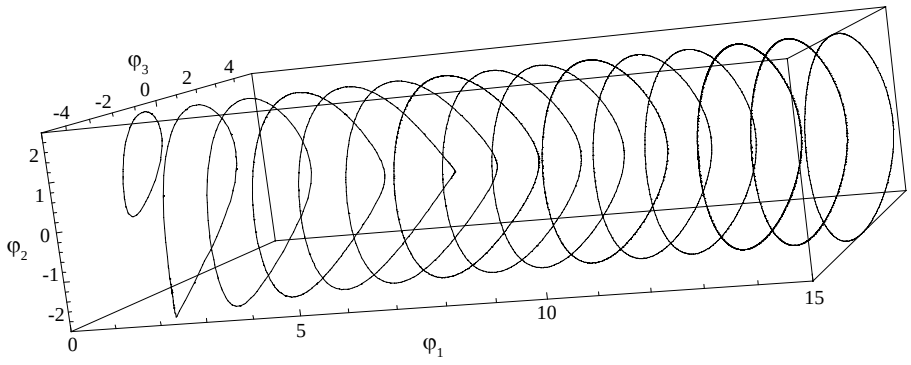
$$
\mathbf{A}_I = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix}.
$$

(45)

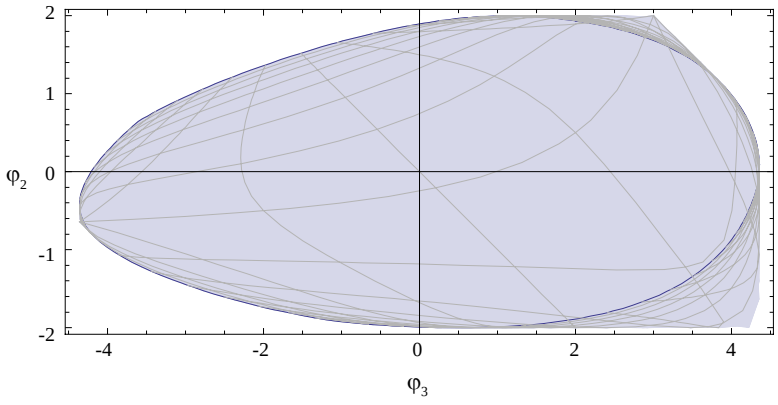
Figure 6 illustrates the relationship between φ_1, φ_2, and φ_3 in the distance space when the location of landmarks is given by Equation (33).



(a)



(b)



(c)

Figure 6 The relationship between φ_1, φ_2, and φ_3 in the distance space. (a) Three dimensional representation, (b) sections along the φ_1-axis, (c) Projection onto the $\varphi_2\varphi_3$-plane.

In this case,

$$
\begin{aligned}
b_2 &= \frac{-u_1 + u_2}{R} \\
b_3 &= \frac{-u_2 + u_3}{R} \\
c_2 &= \frac{-v_1 + v_2}{R} \\
c_3 &= \frac{-v_2 + v_3}{R}
\end{aligned}
\quad . \tag{46}
$$

Since

$$
b_2 b_3 + c_2 c_3 = \frac{(u_1 - u_2)(u_2 - u_3) + (v_1 - v_2)(v_2 - v_3)}{R^2} . \tag{47}
$$

is not always equal to zero, the obtained D_2D_3-function is not symmetrical with respect to the D_2- and D_3-axes. Random distribution of sample points over plane Ξ causes a correlation between φ_2 and φ_3 due to the concentration of sample points along the direction of the major axis of D_2D_3-function. A careful consideration is necessary to choose the location of sample points in the incremental distance method.

3.3 Location of sample points

This subsection discusses the location of sample points in the two transformation methods. We consider two situations, one where we can locate sample points anywhere over plane Ξ, and the other where the location of sample points is limited to a certain sample region.

1) Sample points without locational limitation

The average distance method permits us to locate sample points easily in which distance variables do not highly correlated with each other. For instance, a random distribution is expected to cause no serious correlation as mentioned in the previous section. In reality, however, this ignores the approximation error in the D_2D_3-function, which might be not negligibly small especially in sample points close around the landmarks. We need to calculate the correlation coefficient between δ_1, δ_2, and δ_3 in applications.

To test the validity of the above procedure, we performed a numerical experiment. Three landmarks L_1, L_2, and L_3 are located as defined by Equation (33). We locate 1000 sample points randomly within the distance of radius 20 from the origin, and calculate the absolute correlation coefficients between the distance variables. The result is r_{12}=0.112, r_{23}=0.005, and r_{31}=0.192. Though they are not very close to zero due to the approximation error mentioned earlier, the values are small enough to avoid a serious multicollinearity.

In the incremental distance method, it is not straightforward to avoid the correlation between φ_1, φ_2, and φ_3, especially r_{23} because theD_2D_3-function is not symmetrical with respect to the φ_2- and φ_3-axes. As seen in Figure 4, if we locate sample points randomly, they are clustered in the direction of the major axis of the D_2D_3-function. It causes a correlation between D_2 and D_3 in the direction of the major axis.

To avoid this problem, we locate sample points more densely in the direction that corresponds to the minor axis of the D_2D_3-function than in the direction for the major axis (rotation angle of the the D_2D_3-function is 0.3038). More precisely, the point density of a certain direction is in proportion with the length of its corresponding radius of the D_2D_3-function. We locate 1000 points within the distance of radius 20 from the origin as seen in Figure 7. The result is r_{12}=0.036, r_{23}=0.012, and r_{31}=0.167, which are all again small enough to avoid a serious multicollinearity.

We should note that the above methods does not work successfully when the three landmarks form a very thin triangle, i.e., one of the interior angles is close to π. This generates a thin elliptical D_2D_3-function, where reduction of r_{23} is quite difficult. In such a case, we should omit one of the three distance variables in regression analysis.

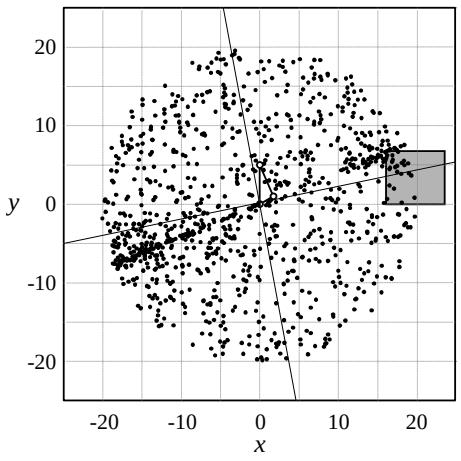


Figure 7 Location of sample points.

2) Sample points within a sample region

We then consider the case where we can locate sample points only in a given sample region. Whether we can avoid the correlation between distance variables depends on the location of sample region. A correspondence exists between the direction of sample points in the real space and that in the D_2D_3-space as seen in Figure 4. If the sample region is situated on either the major or minor axis, we can locate sample points almost symmetrically with respect to the axis, which permits to reduce the correlation by rotating the D_2D_3-function. We cannot locate sample points symmetrically if the sample region is away

from the axes due to the lack of symmetry in the D_2D_3-function.

To evaluate the possible location of sample region, we perform a numerical experiment. Let r_{max} be the maximum correlation given by $r_{max}=\max\{r_{12}, r_{23}, r_{31}\}$. The experiment derives the minimum size of sample region that permits us to reduce r_{max} to a desirable level, as well as the minimum r_{max} that can be obtained by a sample region of appropriate size. We first locate a circular sample region of radius 5.0 on the circle of radius 20 centered at the origin. The sample region contains 1000 sample points distributed randomly. We calculate r_{12}, r_{23}, and r_{31}, and move the points in such a way that reduces $r_{max}=\max\{r_{12}, r_{23}, r_{31}\}$ by the steepest descent method (Hamacher and Drezner (2002); Avriel (2003); Snyman (2005); Fletcher (2013)). When r_{max} becomes smaller than 0.5, we then shrink the sample region with 1000 points and move them until $r_{max}<0.5$. We repeat the same procedure for sample regions centered on the circle of radius 20 at $\pi/90$ interval from angle $-\pi/18$ to $17\pi/18$ measured from the x-axis.

Figure 8 shows the result of experiment of the average distance method. As clearly indicated in the figure, the method works successfully only around the area that corresponds to the major and minor axes of the D_2D_3-function. It is almost impossible to control r_{max} smaller than 0.5 in other area even if we accept sample points distributed in a circle of radius 5. In such areas r_{max} is close to 1.0 as seen in Figure 8.

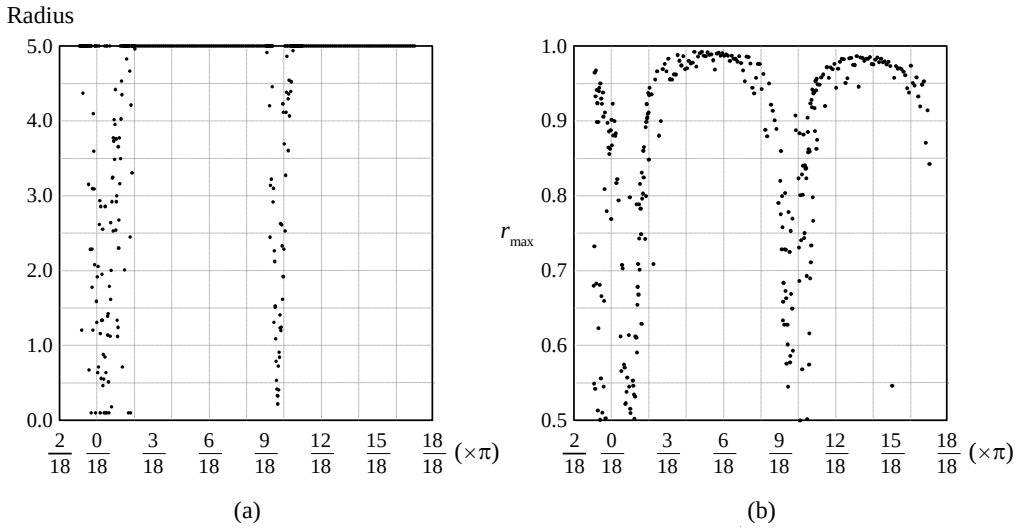


Figure 8 Sample points in a limited region. (a) The minimum radius of sample region to attain $r_{max}<0.5$. (b) The minimum r_{max} $0.5<r_{max}$ by sample region of radius 5.

A similar result is obtained for the incremental distance method. On the other hand, if we rotate the D_2D_3-function to minimize the correlation in every arrangement of sample points, r_{max} becomes smaller than 0.5 in any case. This indicates that it is inappropriate to specify the rotation of the D_2D_3-function when we can locate sample points only in a limited region except when it is close to the major

or minor axis of the D_2D_3-function.

If the sample region is on one of the axis of the D_2D_3-function in the D_2D_3-space, the average distance method works successfully. We performed a numerical simulation to examine this in detail. For the average distance method, we prepare a fan-shaped sample region bounded by two circles of radius 18 and 22 centered at the origin, and two lines of angles $-\pi/25.0$ and $\pi/25.0$ measured from the line that corresponds to the minor axis of elliptical function. We initially locate sample points on concentric circles of interval 0.5 to obtain 116 points on a curved lattice as shown in Figure 7a. The absolute correlations are r_{12}=0.650, r_{23}=0.286, and r_{31}=0.020. Since r_{12} is not small enough, we move sample points by the optimization method mentioned above to reduce the correlations smaller than 0.5. We limited the moving range of each point within distance 1.0 from its initial location to keep the uniformity of sample points. Figure 7b shows the result, where all the correlations are smaller than 0.5. As seen in the figure, we could reduce the correlations with only a slight modification of sample distribution. The incremental distance method also yields a similar result.

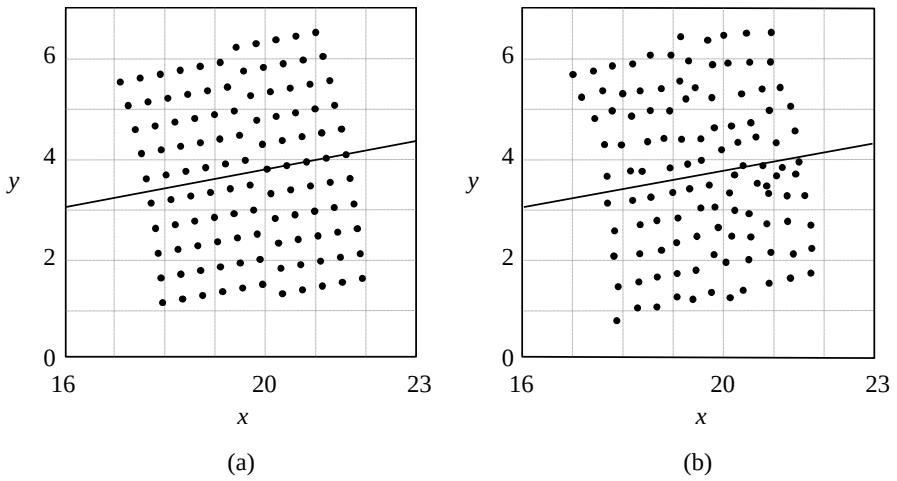


Figure 9 Calculation of sample points. (a) The initial location of sample points in the limited sample region. (b) The final location of sample points obtained after optimization.

References

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

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

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

Partridge, Mark D, Dan S Rickman, Kamar Ali and M Rose Olfert. 2008. "Employment growth in the American urban hierarchy: long live distance." *The BE Journal of Macroeconomics* 8.

−−−. 2008. "Lost in space: population growth in the American hinterlands and small cities." *Journal of Economic Geography*, lbn038.

Partridge, Mark, M. Rose Olfert and Alessandro Alasia. 2007. "Canadian cities as regional engines of growth: agglomeration and amenities

Les villes canadiennes en tant qu'engins de croissance: agglomération ou commodités?" *Canadian Journal of Economics/Revue canadienne d'économique* 40, 39-68.

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.

Appendix A1

The relationship between distance variables D_2, D_3, δ_2, and δ_3 and coordinates of sample points x and y is represented as a linear transformation:

$$\begin{aligned}\begin{pmatrix} D_2 \\ D_3 \end{pmatrix} &= \begin{pmatrix} b_2 & c_2 \\ b_3 & c_3 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} \\ &= \begin{pmatrix} b_2 x + c_2 y \\ b_3 x + c_3 y \end{pmatrix}\end{aligned} .$$

(48)

Let us assume that $ad\text{-}bc \neq 0$. This leads to

$$\begin{aligned}\begin{pmatrix} x \\ y \end{pmatrix} &= \frac{1}{b_2 c_3 - b_3 c_2}\begin{pmatrix} c_3 & -c_2 \\ -b_3 & b_2 \end{pmatrix}\begin{pmatrix} D_2 \\ D_3 \end{pmatrix} \\ &= \frac{1}{b_2 c_3 - b_3 c_2}\begin{pmatrix} c_3 D_2 - c_2 D_3 \\ -b_3 D_2 + b_2 D_3 \end{pmatrix} \\ x^2 + y^2 &= \frac{1}{\left(b_2 c_3 - b_3 c_2\right)^2}\left\{\left(c_3 D_2 - c_2 D_3\right)^2 + \left(-b_3 D_2 + b_2 D_3\right)^2\right\} \\ &= \frac{1}{\left(b_2 c_3 - b_3 c_2\right)^2}\left\{\left(b_3^2 + c_3^2\right)D_2^2 + \left(b_2^2 + c_2^2\right)D_3^2 - \left(2c_2 c_3 + 2b_2 b_3\right)D_2 D_3\right\} \\ &= R^2\end{aligned} .$$

(49)

This yields

$$\frac{\left(b_3^2 + c_3^2\right)D_2^2 + \left(b_2^2 + c_2^2\right)D_3^2 - 2\left(c_2 c_3 + b_2 b_3\right)D_2 D_3}{\left(b_2 c_3 - b_3 c_2\right)^2 R^2} = 1 .$$

(50)

Distance variables D_2 and D_3 form a quadratic function without linear terms. They form an ellipse centered at the origin when

$$\left(b_2^2 + c_2^2\right)\left(b_3^2 + c_3^2\right) - \left(b_2 b_3 + c_2 c_3\right)^2 > 0 .$$

(51)

We can prove this as follows:

$$\begin{aligned}\left(b_2^2 + c_2^2\right)\left(b_3^2 + c_3^2\right) - \left(b_2 b_3 + c_2 c_3\right)^2 &= b_2^2 b_3^2 + c_2^2 c_3^2 + b_2^2 c_3^2 + c_2^2 c_3^2 - b_2^2 b_3^2 - c_2^2 c_3^2 - 2b_2 b_3 c_2 c_3 \\ &= b_2^2 c_3^2 + c_2^2 c_3^2 - 2b_2 b_3 c_2 c_3 \\ &= \left(b_2 c_3 - b_3 c_2\right)^2 > 0\end{aligned}$$

(52)

if $ad\text{-}bc\neq 0$.

The rotation angle is

$$\begin{aligned}\vartheta &= \frac{1}{2}\arctan\left(\frac{-\left(2b_2b_3+2c_2c_3\right)}{b_2^2-b_3^2+c_2^2-c_3^2}\right)\\ &= -\frac{1}{2}\arctan\left(\frac{2\left(b_2b_3+c_2c_3\right)}{b_2^2-b_3^2+c_2^2-c_3^2}\right)\end{aligned}.$$

(53)

Rotating the ellipse ϑ counterclockwise, we obtain an ellipse that is symmetrical with respect to both the D_2- and D_3- axes:

$$\begin{aligned}\begin{pmatrix}D_2'\\D_3'\end{pmatrix} &= \begin{pmatrix}\cos\vartheta & -\sin\vartheta\\ \sin\vartheta & \cos\vartheta\end{pmatrix}\begin{pmatrix}D_2\\D_3\end{pmatrix}\\ &= \begin{pmatrix}\left(b_2\cos\vartheta-b_3\sin\vartheta\right)x+\left(c_2\cos\vartheta-c_3\sin\vartheta\right)y\\ \left(b_2\sin\vartheta+b_3\cos\vartheta\right)x+\left(c_2\sin\vartheta+c_3\cos\vartheta\right)y\end{pmatrix}\end{aligned}.$$

(54)

Distances D_2 and D_3 are thus represented as

$$\begin{aligned}\begin{pmatrix}D_2\\D_3\end{pmatrix} &= \begin{pmatrix}\cos\vartheta & \sin\vartheta\\ -\sin\vartheta & \cos\vartheta\end{pmatrix}\begin{pmatrix}D_2'\\D_3'\end{pmatrix}\\ &= \begin{pmatrix}D_2'\cos\vartheta+D_3'\sin\vartheta\\ D_2'\sin\vartheta+D_3'\cos\vartheta_3\end{pmatrix}\end{aligned}.$$

(55)

Substituting the above equation into Equation (49), we obtain

$$\begin{aligned}&\left\{\left(b_3^2+c_3^2\right)\left(D_2'\cos\vartheta+D_3'\sin\vartheta\right)^2+\left(b_2^2+c_2^2\right)\left(D_2'\sin\vartheta+D_3'\cos\vartheta\right)^2\right.\\ &\left.-2\left(c_2c_3+b_2b_3\right)\left(D_2'\cos\vartheta+D_3'\sin\vartheta\right)\left(D_2'\sin\vartheta+D_3'\cos\vartheta\right)\right\}\frac{1}{\left(b_2c_3-b_3c_2\right)^2R^2}\\ &=\left[\left\{\left(b_3^2+c_3^2\right)\cos^2\vartheta+\left(b_2^2+c_2^2\right)\sin^2\vartheta-2\left(c_2c_3+b_2b_3\right)\sin\vartheta\cos\vartheta\right\}D_2'^2\right.\\ &+\left\{\left(b_3^2+c_3^2\right)\sin^2\vartheta+\left(b_2^2+c_2^2\right)\cos^2\vartheta-2\left(c_2c_3+b_2b_3\right)\sin\vartheta\cos\vartheta\right\}D_3'^2\\ &\left.+2\left\{b_2^2+b_3^2+c_2^2+c_3^2-2\left(c_2c_3+b_2b_3\right)\right\}\sin\vartheta\cos\vartheta D_2'D_3'\right]\frac{1}{\left(b_2c_3-b_3c_2\right)^2R^2}\\ &=1\end{aligned}.$$

(56)

The length of axes of the ellipse is represented as

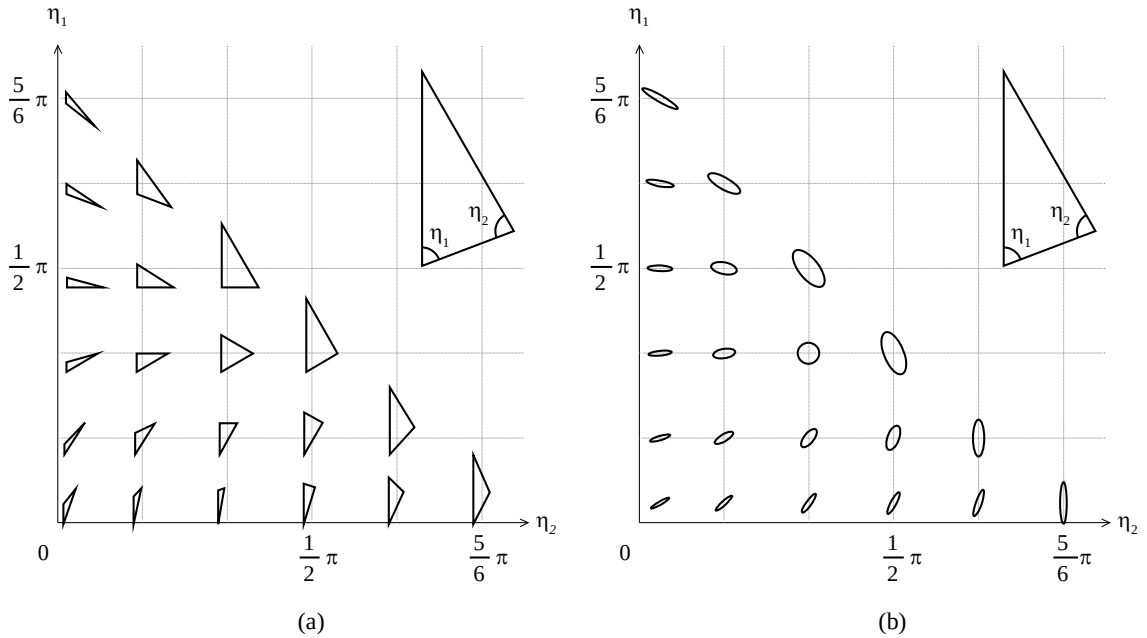
$$\begin{pmatrix} E_1 \\ E_2 \end{pmatrix} = \sqrt{2}\left|b_2c_3 - b_3c_2\right| R \begin{pmatrix} \dfrac{1}{\sqrt{b_2^2+b_3^2+c_2^2+c_3^2-\sqrt{\left(b_2^2-b_3^2+c_2^2-c_3^2\right)^2+4\left(b_2b_3+c_2c_3\right)^2}}} \\ \dfrac{1}{\sqrt{b_2^2+b_3^2+c_2^2+c_3^2+\sqrt{\left(b_2^2-b_3^2+c_2^2-c_3^2\right)^2+4\left(b_2b_3+c_2c_3\right)^2}}} \end{pmatrix}. \tag{57}$$

Appendix A2

This appendix examines the relationship between the arrangement of landmarks and the distance function in detail. Let η_i be the internal angle of triangle $L_1L_2L_3$ at L_i. We assume that L_1L_2=1 since the elliptical function is independent of the absolute size of $L_1L_2L_3$. The coordinates of the landmarks then become

$$\begin{aligned} \left(u_1, v_1\right) &= \left(0, 0\right) \\ \left(u_2, v_2\right) &= \left(\sin\eta_1, \cos\eta_1\right) \\ \left(u_3, v_3\right) &= \left(0, \cos\eta_1 - \sin\eta_1 \cot\left(\eta_1+\eta_2\right)\right) \end{aligned} . \tag{58}$$

Figure 10a and b show the arrangement of landmarks and the distance function, respectively.



(a) (b)

Figure 10 The relationship between the arrangement of landmarks and the D_2-D_3 function. (a) The arrangement of landmarks. (b) The D_2-D_3 function.

As seen in Figure 10, the distance function becomes elongated as the smallest interior angle of $L_1L_2L_3$ shrinks. The orientation of the function also depends on the shape of $L_1L_2L_3$. It seems closely related to the orientation of triangle $L_1L_2L_3$.

To clarify this relationship in more detail, we examine the correspondence between the location of sample points on plane Ξ and that on the distance function at a certain section along the D_1'-axis. When sample points are distant from the landmarks, the location of sample points on the D_2D_3 plane depends only on the direction with respect to the landmarks as mentioned earlier. We thus consider sample points located on a large circle centered at the gravity center of the landmarks. Landmarks are arranged in ten among twenty-one patterns shown in Figure 10a, where

$$\left(\eta_1,\eta_2\right)=\left(\frac{p}{6}\pi,\frac{q}{6}\pi\right)\quad\left(p,q=1,...,5,\,p>q\right). \tag{59}$$

We calculate the location of end points of the major and minor axes of the ellipses in Figure 10b on plane Ξ. Figure 11a shows the distance function rotated in such a way that the major axis is horizontal. Figure 11b indicates the spatial relationship between the landmarks and sample points that correspond to the end points of distance function. Both the landmarks and sample points are rotated in such a way that the diameter connecting the sample points corresponding to the end points of the major axis is horizontal.

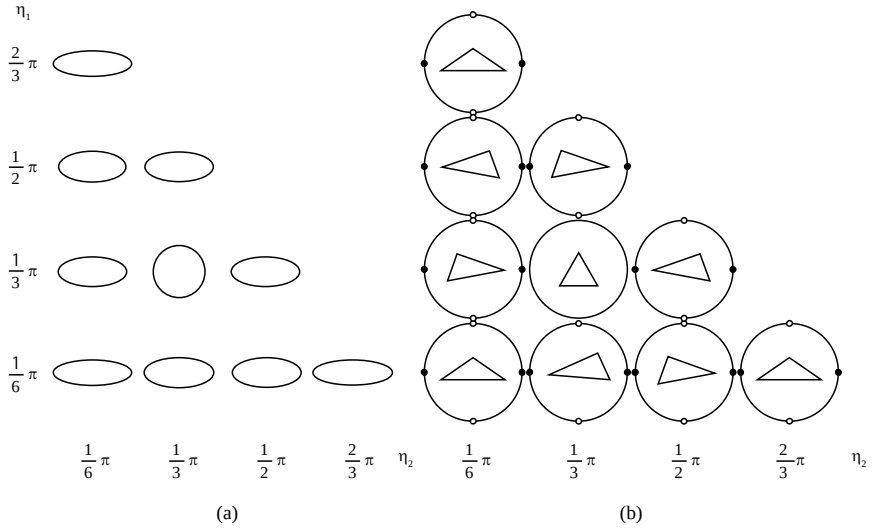


Figure 11 The distance function and the spatial relationship between the landmarks and the end points

of the distance function. (a) The distance function rotated in such a way that the major axis is horizontal. The elliptical function. (b) The spatial relationship between the landmarks and the end points of the distance function. Both the landmarks and sample points are rotated in such a way that the diameter connecting the sample points corresponding to the end points of the major axis is horizontal.