International Journal of Architectural, Civil and Construction Sciences
ISSN: 2415-1734
Vol:9, No:4, 2015

A Research on Inference from Multiple Distance Variables in Hedonic Regression – Focus on Three Variables

Yan Wang, Yasushi Asami, Yukio Sadahiro

***Abstract*—**In urban context, urban nodes such as amenity or hazard will certainly affect house price, while classic hedonic analysis will employ distance variables measured from each urban nodes. However, effects from distances to facilities on house prices generally do not represent the true price of the property. Distance variables measured on the same surface are suffering a problem called multicollinearity, which is usually presented as magnitude variance and mean value in regression, errors caused by instability. In this paper, we provided a theoretical framework to identify and gather the data with less bias, and also provided specific sampling method on locating the sample region to avoid the spatial multicollinerity problem in three distance variable's case.

***Keywords*—**Hedonic regression, urban node, distance variables, multicollinerity, collinearity.

I. Introduction

SINCE Lancaster (1966) with consumer theory [1] and Rosen (1974) with the theoretical model [2] in urban spatial structure proposed, the hedonic pricing theory in urban spatial structure has become widely accepted by researchers and economists. Classic linear functional form of Hedonic pricing model can be written as:

$$P = a + b_1 d_1 + b_2 d_2 + b_3 d_3 + \ldots + e$$

where P denotes the house price, a denotes constant term, d_1 and d_2 denote the distances measured from the landmarks, b_1 and b_2 indicate the parameter of distance variables, which are usually called as regression coefficients, and e denotes the independent and identically distributed random variable with the mean value equal to zero and constant variance. Hedonic price theory elaborated that certain urban nodes often influence the market price. Hedonic regression provided a method of estimating the house price in urban area. This method is often performed in order to assess the economic values of urban nodes such as landmarks, hospitals, stations or chemistry factories, medical waste facilities etc. However, in regression, problems such as magnitude variance and unacceptable coefficient fluctuation will present when using multiple distance variables in hedonic analysis. As an obvious research that multivariate statistics on geography which is generally used the multiple linear regression [3]. Recently, the application of R, SPSS, GIS and other tools in urban engineering showed the powerful ability in analysis. In early studies, Cook (1977) developed regression in diagnostics method [4]. In the ordinary least squares (OLS) method, Hoerl and Kennard (1970) pointed out that if some of the independent predictors are correlated strongly, the regression will be led to ill circumstance. They introduced ridge regression to isolate the facilities that didn't produce effect on the regression [5]. In some cases, as the multiple variables were employed, the multiple collinearity problems will rise. To solve this problem, various functional forms were introduced in previous research. Hoaglin and Welsch (1978) provided a hat matrix to delete residuals [6]. So Neter et al. (1996) introduced another technique by using the index of variance inflation factor (VIF) to monitor the multicollinearity between household attributes and instability of assessment [7].

II. Literature Review

Some scholars turned to the research on the sampling method instead of statistics. Griffith (1981) showed that the third distance variable may be needless in a two nodes model [8]. However, instead of affecting the household, it does affect rents. Eric Heikkila (1988) pointed out the possibility to avoid or reduce potential spatial collinearity by selecting the location range. The sample area should be carefully located between the area when some of the urban nodes are located on the edge of the research region and others located on the center [9]. M. Tiefelsodorf (2003) estimated the spatial structure effects of general interaction model by applying distance variables in hedonic regression. But the multicollinearities improved the difficulties on interpreting the coefficient properly [10]. Jaewoo Lee and Keemin Sohn (2003) used dummy variable and Box-Cox transformation in hedonic regression to identify the influence on land prices by replacing at-grade or elevated railways to subways in Seoul metropolitan area [11]. Instead of using Cartesian coordinates, Sander et al. (2010) did a research on the effects of different distance calculations in spatial modeling by focusing on the measurement of spatial distance from open space [12]. J. M. Ross et al. (2011) gave the method on identifying the true landmarks which affect the house price and made a suggestion on the second landmark; also they mentioned the multicollinearity will cause unstable error in analysis when the third distance variable was involved in

Yan Wang, Ph.D. Candidate of the Graduate School of Urban Engineering, the University of Tokyo, Japan (HP: 81-80-4926-9090; fax: 03-5841-8521; e-mail: wangyan@ua.t.u-tokyo.ac.jp / brucio99@gmail.com).
Yasushi Asami, Professor of Department of Urban Engineering, the University of Tokyo, Japan (e-mail: asami@csis.u-tokyo.ac.jp).
Yukio Sadahiro, Professor of Center for Spatial Information Science, the University of Tokyo, Japan (e-mail: sada@csis.u-tokyo.ac.jp).

hedonic regression [13]. Hao Huang and Li Yin (2014) used hedonic price models to examine a comprehensive set of environmental sustainability elements in Wuhan, China. They mentioned that the distance from transit systems and central business districts will affect the house price both positively and negatively [14].

The method outlined in this paper focuses more on the location and collinearity attributes of the data itself rather than using mathematical operation and statistical properties. This paper shared similar purpose with Eric Heikkila (1988)'s work [9]. But it used diverse simulations in analysis.

A lot of researches pointed out the limitations of identifying the correlation among the variables on the effect from facilities and estimation of house price. It cannot be solved by the statistic and relative economic analysis easily. But when we are facing some research regions with several specific facilities such as stations and schools involved, figuring out the effective data and selecting the samples with less or minimum bias in hedonic regression is not mentioned. Three nodes in space, which are considered as weighted vectors to some specific research regions, the distance variables are measured on the same two-dimensional surface and the correlation between three facilities will cause collinearity attributes on the regression coefficients. This paper will provide several solutions by reducing the data with bias. And we will offer sampling method in three distance variables case and assess the effect of the specific selected facilities to the observation area. In this study we will focus on the theoretical model with three nodes in urban context.

To demonstrate the three landmarks sampling methods in hedonic regression, we have built a set of simulations in an ideal model. This paper is arranged into 6 sections. After the introduction (Section I) and the literature review (Section II), we will run the simulation based on correlation analysis both in a tiny region and wide region context (Section III). Then we will analyze the distance variables' regression coefficients' fluctuation by these simulations. The second set of simulations is under the condition of a tiny region (Section IV). The third set of simulations is under the condition of a wide region (Section V). To analyze the sampling area and the influence of positioning on distance variables, we took an analysis in a wide region. After that, we will show the advantage of the recommended sampling area by illustrating the variance, coefficient mean and regressor correlation, also the method to locate the sample region with less bias. Finally, offer the conclusions and suggestions in hedonic analysis with three specific urban nodes (Section VI).

III. Correlation Analysis

Correlation Coefficient Simulation

In three landmarks case, without considering other influence but distance as regressor, the basic hedonic linear functional form could be written as:

$$P_i = a + b_1 d_{1i} + b_2 d_{2i} + b_3 d_{3i} + e_i \quad (1)$$

$$\hat{P}_i = \hat{a} + \hat{b}_1 d_{1i} + \hat{b}_2 d_{2i} + \hat{b}_3 d_{3i} + \hat{e}_i \quad (2)$$

where i denotes each random sample, P_i denotes the house price of the sample point S_i, a denotes the constant term in regression, d_{1i}, d_{2i} denotes the distances that measured from the two facilities Q_1 and Q_2 to sample point S_i, b_1, b_2 denotes the coefficient of distance variables, and e_i denotes the error term. The coefficients are usually unknown and the error term is unobserved. The statistical model depends on the estimation of b_1, b_2, b_3 and the variance of P. In Simulation-1, we have set the observation area as a 10×10 square region which is centered on origin $O(0\,0)$ with three specific urban nodes that located on $Q_1(-1.5,0)$, $Q_2(1.5,0)$ and $Q_3(0.5,2)$ as in Fig. 1. Take the sample points as uniform distribution interval defined by n. Calculate the distance from each point $S_{i(xi,\,yi)}$ to the three nodes in the square region, mark them as d_{1i}, d_{2i} and d_{3i}. Then the price of S_i could be calculated by following the P_i equation, (1). Taking a set of observations (for example 500 samples) and ran the regression with P_i and d_{1i}, d_{2i}, d_{3i} for 1000 trails. Analyzing the fluctuation of parameter $\hat{b}_1$ (coefficient of d_{1i}), $\hat{b}_2$ (coefficient of d_{2i}) and $\hat{b}_3$ (coefficient of b_3),we can estimate the model and sampling region from the bias of $\hat{b}_1 - b_1$, $\hat{b}_2 - b_2$ and $\hat{b}_3 - b_3$.

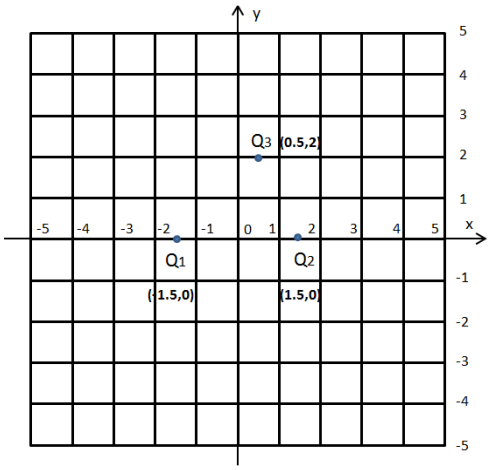

Fig. 1 Observe region of Simulation-1

Since the relationship of correlation and collinearity are quite inseparable, we calculated the correlation coefficient of the square region first (Simulation-1) before running the regression. Hence the three distance variables on same surface with longitude and latitude coordinates will form problems such as multicollinearity that proved by lots of other researchers in previous studies (see [8] and [12]).

In the performance of correlation calculation, the observation region was set with interval step as t'=0.02, the sampling area is formed as a circle region C (radius $r_c = 0.5$). Take 1000 observations randomly in this circle region. When C's center is moving in the whole area as interval step t_1=0.05,

both on x and y axes, the distance from sampling point S_i to node $Q_1(-1.5,0)$, $Q_2(1.5,0)$ and the third node $Q_3(0.5,2)$ can be measured. Then calculating each distance d_{1i}, d_{2i} and d_{3i} correlation coefficient R_{12}, R_{13} and R_{23}. By plotting the results, we can get the contour line map and the surface map of the entire observation region's correlation coefficient from Figs. 2-7.

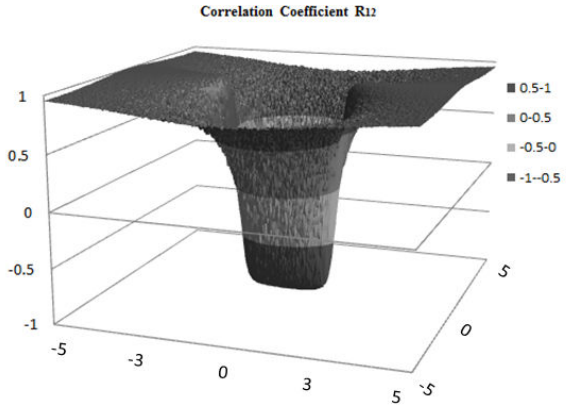

Fig. 2 Surface map of correlation coefficient R_{12}

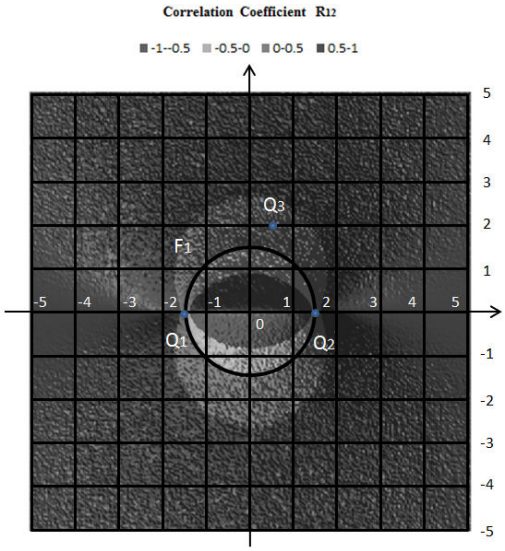

Fig. 3 Contour line map of correlation coefficient R_{12}

From Figs. 2, 4 and 6, we could find out that the correlation coefficient changed sharply from nearly 1 to -1, but tend to be zero when the location is close to the circle which passed through each two facilities, which were marked as circle F_1, F_2 and F_3. The locations, which are outside of circle F_1, F_2 and F_3, produced magnitude positive correlations which are close to 1, meanwhile the points around the circle F_1, F_2 and F_3's center produced huge negative correlations which are close to -1. The correlation started to fall when the sample region C is getting close to the circles. Sample points on or close to circle F_1, F_2 and F_3 showed quite tiny correlation value which is close to 0. They are more effective than others in the observation region.

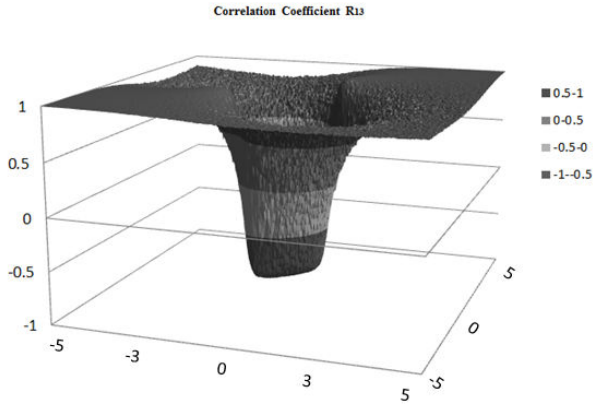

Fig. 4 Surface map of correlation coefficient R_{13}

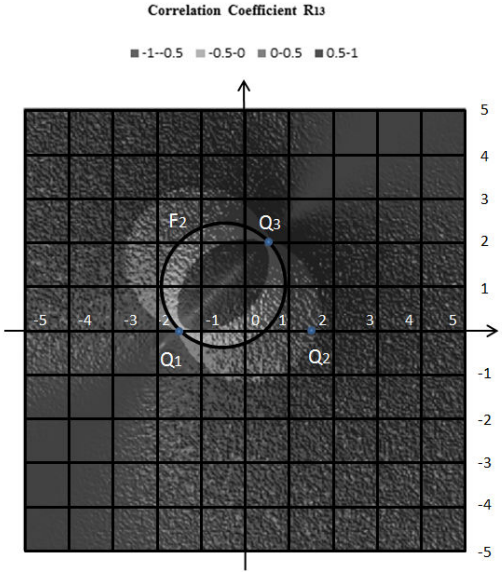

Fig. 5 Contour line map of correlation coefficient R_{13}

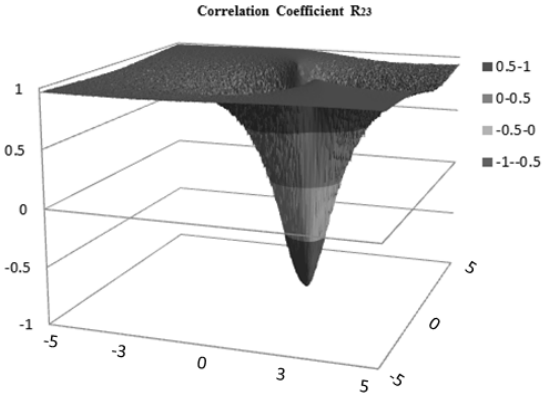

Fig. 6 Surface map of correlation coefficient R_{23}

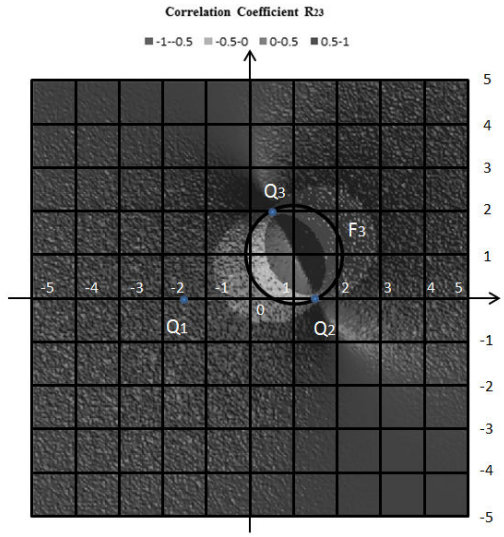

Fig. 7 Contour line map of correlation coefficient R_{23}

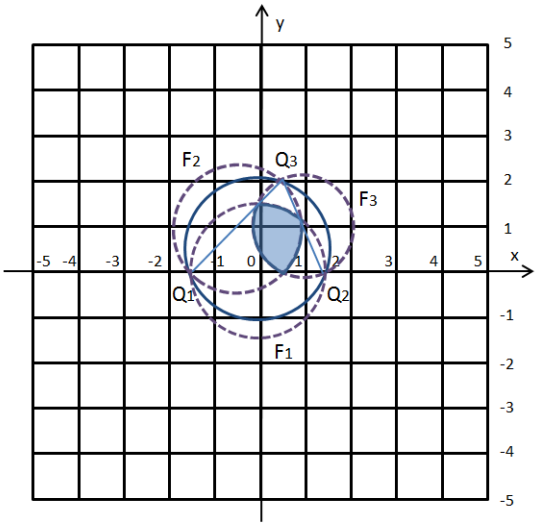

Fig. 8 Correlation Coefficient Analysis-2

Based on this result above, in regression, we should choose the samples in the area where the absolute value of correlation coefficient |R| is small. Usually, samples with high correlation value will produce more bias in regression. Obviously, it will not be the region inside or outside the circles marked in the contour line maps. It depends on the results showed in above figures. To identify the valid samples, we overlapped $F1$, $F2$ and $F3$ circle on the same coordinate illustration as Fig. 8. The minimum coefficient circles $F1$, $F2$ and $F3$ were marked as dotted circles, they are overlapped on the shadow region, where shares the comparatively small correlation. Additionally, the center of the shadow region is tending to be the best sampling region based on the analysis result. However, it will be difficult to locate observations exactly on the center of the shadow region. So samples should be selected around or close to the center. Furthermore, we should avoid the locations where |R| increase rapidly such as the neighborhood around three node $Q1(-1.5,0)$, $Q2(1.5,0)$, $Q3(0.5,2)$ and the center of circle $F1$, $F2$ and $F3$ that have shown in the figures.

Here we had built a regression model to confirm our hypothesis started from tinny region context.

As Fig. 8 showed, it will be hard to locate the center of the shadow region; for convenience, we have located the origin on the midpoint of segment $Q1Q2$. So if we follow this construction of coordination, the third landmark's position could be considered as a position that is changing. And it's getting more difficult to identify the center area of the overlapping region.

Since three nodes in space will remind us triangle and trigonometry easily. The shadow region is just located inside of the $\triangle Q1Q2Q3$. And Schmeidler (1969) made a suggestion that taking the sample area in the triangle of 3 nodes may have positive effect on the analysis when three regression variables are singular matrices [15]. Though, the incenter and barycenter of the triangle could be our choice; the calculation of barycenter is much easier. If we set the base of the triangle on the x axis and let the origin just fit the midpoint of the base side, barycenter will be on the 1/3 Apollonius line which is close to origin, which means $Q3$'s coordinate is related with our sample region A's coordinate. It's quite convenient for the mathematical operation.

Now we are going to confirm this barycenter sampling theory with the correlation simulation's data set. Urban nodes $Q1(-1.5,0)$, $Q2(1.5,0)$ and $Q3(0.5,2)$ (also marked as $Q3(x3, y3)$) located on the observation region are described as Fig. 9. Following the correlation analysis, we should locate the sampling region right in the shadow area. According to the previous study above, we calculated the $\angle SOQ2$'s angle θ' by trigonometry as (3), and the result is 75.96376 degree.

$$\theta' = arc\cos\left[x3 / \sqrt{(x3^2 + y3^2)} \right] \tag{3}$$

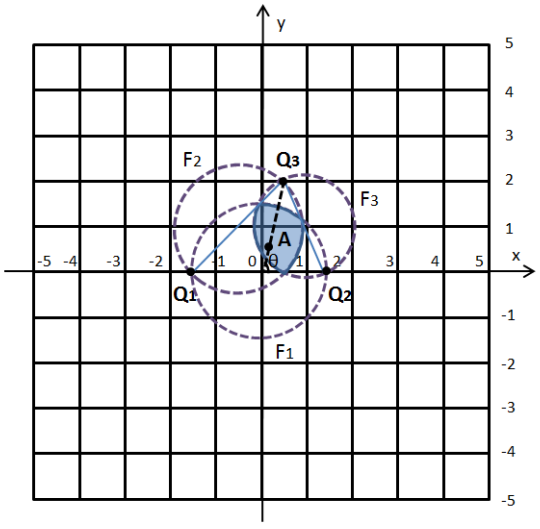

Fig. 9 Illustration of observation region of Simulation-5

The coordinate of the barycenter could be written as: A ($x_i = x_3 \cos\theta'/3$, $y_i = y_3 \sin\theta'/3$). House price P_i could be calculated by (1), d_{1i} , d_{2i} could be calculated by (4) and (5). d_{3i} can be achieved by (6).

$$d_{1i} = \left[\left(x_i + 1.5 \right)^2 + y_i^2 \right]^{1/2} \quad (4)$$

$$d_{2i} = \left[\left(x_i - 1.5 \right)^2 + y_i^2 \right]^{1/2} \quad (5)$$

$$d_{3i} = \left[\left(x_i - x_3 \right)^2 + \left(y_i - y_3 \right)^2 \right]^{1/2} \quad (6)$$

The parameters were set as: $a = 2500$, $b_1 = -130$, $b_2 = -100$, $b_3 = -80$, $e = 0.001$.

When $\delta = 0.01$ (divided by $m = 50$), the sampling region is a tiny square region with length and width both as $j = 0.02$. We run the regression with the variable d_{1i} , d_{2i} , d_{3i} and P_i for 1000 trails, marked this regression as Simulation-2. The result listed in Table I (θ' is equal to 75.9638 degree's situation).

TABLE I
RESULTS OF SIMULATION-2

Area	Value	$\hat{b}_1 - b_1$	$\hat{b}_2 - b_2$	$\hat{b}_3 - b_3$
θ'	Variance	6.592676411	4.915824629	2.299512033
	Mean	-0.053611841	-0.046556987	-0.031879241
	SD/N	0.005135242	0.004434332	0.003032828

Simulation based on 1,000 trials on sample size of N=10,201. House price has three distance variables in data producing process. SD/N means Standard Deviation divided by N.

Further, to confirm the relationship of correlation coefficient R_{12} , R_{13} and R_{23} , we built a regression with the distance d_3 and the other two distance d_1 and d_2. If the coefficient of d_1 and d_2 are β_1 and β_2 , β_0 denotes the constant term, then d_3 can be written as (7):

$$d_3 = \beta_0 + \beta_1 d_1 + \beta_2 d_2 \quad (7)$$

$$\hat{d}_3 = \hat{\beta}_0 + \hat{\beta}_1 d_1 + \hat{\beta}_2 d_2 \quad (8)$$

Take 500 observations of d_1 and d_2 that follow the uniform distribution (tiny region method) randomly. Running the regression, we get the result in Table II.

TABLE II
RESULTS OF REGRESSION -1

	Coefficients	Standard Error	t Stat	P-value
Intercept	6.563985	0.001141	5751.464	0
X Variable 1	-1.69294	0.000379	-4465.07	0
X Variable 2	-1.46244	0.000383	-3823.19	0

Regression based on sample size of N=10,201. House price has three distance variables in data producing process.

Adjusting the parameters as $\beta_0 = 6.6$, $\beta_1 = -1.7$, $\beta_2 = -1.5$. d_3 could be calculated by (7). Running the regression with d_1 , d_2 and d_3 for 1000 trails. We can get the variance, mean value and standard deviation divided by N of $\hat{\beta}_0 - \beta_0$, $\hat{\beta}_1 - \beta_1$ and $\hat{\beta}_2 - \beta_2$ in Table III.

TABLE III
RESULTS OF SIMULATION -3

Area	Value	$\hat{\beta}_0 - \beta_0$	$\hat{\beta}_1 - \beta_1$	$\hat{\beta}_2 - \beta_2$
θ'	Variance	7.82833E-29	8.92178E-30	8.49736E-30
	Mean	7.63833E-17	2.16049E-16	1.51656E-16
	SD/N	1.76956E-17	5.97387E-18	5.83004E-18

Simulation based on 1,000 trials on sample size of N=10,201. House price has three distance variables in data producing process. SD/N means Standard Deviation divided by N.

Based on this result, we could figure out that this Simulation-3 is stable, because the variance and the mean value are very tiny. This fits the prediction that distance variable d_3 has linear relationship with d_1 , d_2 , and it can be described by (9) in a tiny study region context. Basically, the influence from d_1 and d_2 are similar, with a constant term 6.6.

$$d_3 = 6.6 - 1.7 d_1 - 1.5 d_2 \quad (9)$$

When $\delta = 0.5$ (divided by $m = 50$), the sampling region is a wide square region with length and width both as $j = 1$. Run the regression with the three distance variable d_{1i} , d_{2i} , d_{3i} and P_i for 1000 trails, marked this regression as Simulation-4. The result is listed in Table IV (θ' is equal to 75.9638 degree's situation).

TABLE IV
RESULTS OF SIMULATION -4

Area	Value	$\hat{b}_1 - b_1$	$\hat{b}_2 - b_2$	$\hat{b}_3 - b_3$
θ'	Variance	9.54388E-07	7.15617E-07	3.52211E-07
	Mean	2.13723E-07	-1.74182E-07	3.17814E-06
	SD/N	1.95386E-06	1.69188E-06	1.18695E-06

Simulation based on 1,000 trials on sample size of N=10,201. House price has three distance variables in data producing process. SD/N means Standard Deviation divided by N.

We also built a regression with the distance d_3 and the other two distance d_1 and d_2. The coefficients of d_1 and d_2 are designated as β_1' and β_2' , β_0' denotes the constant number. d_3 can be calculated as (7).

Take each d_1 and d_2 as 500 samples randomly following the distribution in Simulation-5(following wide region method). Running the regression, we can get the result of $\hat{\beta}_0'$, $\hat{\beta}_1'$ and $\hat{\beta}_2'$ in Table V.

Adjusting the parameters to $\beta_0 = 6.6$, $\beta_1 = -1.7$, $\beta_2 = -1.5$. Calculate d_3 by (7). Running the regression with d_1 , d_2 and d_3 for 1000 trails. The variance, mean value and

standard deviation divided by N of $\hat{\beta}_0 - \beta_0$, $\hat{\beta}_1 - \beta_1$ and $\hat{\beta}_2 - \beta_2$ are listed in Table VI.

TABLE V
RESULTS OF REGRESSION -2

	Coefficients	Standard Error	t Stat	P-value
Intercept	6.563985079	0.001141272	5751.464344	0
X Variable 1	-1.692939383	0.000379152	-4465.067149	0
X Variable 2	-1.462439944	0.000382518	-3823.188368	0

Regression based on sample size of N=10,201. House price has three distance variables in data producing process.

TABLE VI
RESULTS OF SIMULATION -5

Area	Value	$\hat{\beta}_0 - \beta_0$	$\hat{\beta}_1 - \beta_1$	$\hat{\beta}_2 - \beta_2$
θ'	Variance	7.82833E-29	8.49736E-30	8.92178E-30
	Mean	2.16049E-16	1.51656E-16	7.63833E-17
	SD/N	5.97387E-18	5.83004E-18	1.76956E-17

Simulation based on 1,000 trials on sample size of N=10,201. House price has three distance variables in data producing process. SD/N means Standard Deviation divided by N.

Based on this result, we could figure out this Simulation-5 is stable, for the variance and the mean value are quite small. This fits the theory that distance variable $d3$ has linear relationship with $d1, d2$, and it can be described by (10) in a wide region case. As same as the tiny region case, the influence from $d1$ and $d2$ on $d3$ are sated as -1.7 and -1.5, with a constant number 6.6.

$$d_3 = 6.6 - 1.7d_1 - 1.5d_2 \tag{10}$$

Based on the results, when the sampling area centered on the barycenter of the triangle (that distributed by the three urban nodes), both the tiny region regression and the wide region regression achieved the acceptable variance and the mean value of $\hat{b} - b$ (the regression coefficient of three distance variables). At the same settings of parameters, when $\delta = 0.5$ regression variance: $\hat{b}1 - b1$ is 9.54388E-07, $\hat{b}2 - b2$ is 7.15617E-07, $\hat{b}3 - b3$ is 3.52211E-07. They are much smaller than the variance when $\delta = 0.01$. The mean value showed the similar result, when the sampling region is getting wider, the mean value will tend to be smaller. This may because when the δ increases to a comparatively larger value, the sampling region becomes a wide region, including the data located on both positive and negative correlation. They will neutralize the collinearity effect and show the better result. In Simulation-5 context, the distance variable $d3$ has a linear relationship with $d1$ and $d2$, the effect from these two variables are similar. All these results showed: in the same observation region, whether the sampling area is taken as tiny context or wide context, the linear relationship between $d3$ and $d1, d2$ are the same, and they shared one linear equation.

Now we will do more simulations to confirm this barycenter sampling theory.

IV. TINY REGION SIMULATION

Three Variables Tiny Region Simulation

In Simulation-6, two circles centered on origin were drawn in Cartesian coordinate as Fig. 10, their radius were taken as r and $3r$. For calculation convenience, r was taken as equal to 1. Observation region is set as uniform distribution. Q_1, Q_2 are the first and second urban nodes located on points $(-1,0)$ and $(1,0)$, r is the half of segment $Q_1 Q_2$. Q_3 denotes the third node. A is the center of our sampling region. Point A was taken just on the radius $r = 1$ circle. If the angle of $\angle SOQ_2$ is taken as θ, the coordinate of our uniform distributed sample point $A(x, y)$ could be written as $A(\cos\theta, \sin\theta)$. Then, Q_3's coordinate can be written as $Q_3(3\cos\theta, 3\sin\theta)$.

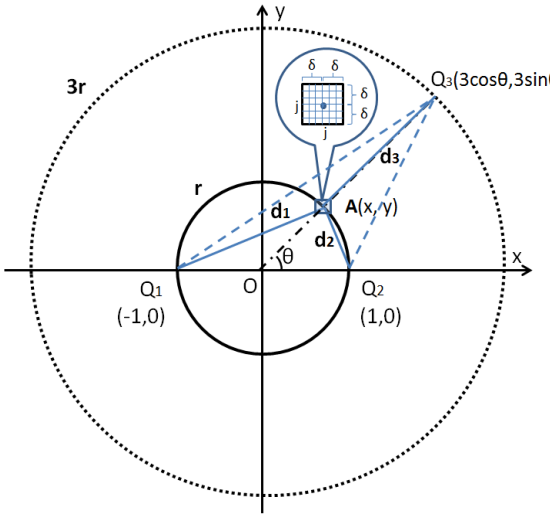

Fig. 10 Illustration of observation region of Simulation-6

The coordinate of sample point A will be set as: $x \in (\cos\theta - \delta, \cos\theta + \delta)$, $y \in (\sin\theta - \delta, \sin\theta + \delta)$, each observation's coordinate could be calculated as (11) and (12):

$$x_i = \cos\theta - \delta + \delta\frac{j}{m} \tag{11}$$

$$y_i = \sin\theta - \delta + \delta\frac{j}{m} \tag{12}$$

where δ denotes the range of the sampling area centered on A, δ is divided into m, j denotes the length and width of the tiny region. Here, in tiny region analysis, δ was taken equal to 0.01. When m was taken equal to 50, the total number N of samples in the tinny sampling region can be calculated by (13):

$$N = (2m+1)(2m+1) \tag{13}$$

In Simulation-6, the parameters were set as: $a = 2500$, $b_1 = -130$, $b_2 = -100$, $b_3 = -80$, $e = 0.001$. If we take

the coordinate of the sample point as (x_i , y_i), then each distance measured from our observation to the three nodes' distance d_{1i}, d_{2i} and d_{3i} could be calculated by (14), (15) and (16) as

$$d_{1i} = \left[(x_i + 1)^2 + y_i^2 \right]^{1/2} \quad (14)$$

$$d_{2i} = \left[(x_i - 1)^2 + y_i^2 \right]^{1/2} \quad (15)$$

$$d_{3i} = \left[(x_i - 3\cos\theta)^2 + (y_i - 3\sin\theta)^2 \right]^{1/2} \quad (16)$$

Since P_i could be calculated by (1), we take 500 observations randomly in the tiny observation region and run the regression with d_{1i}, d_{2i}, d_{3i} and P_i (on θ equal to 30, 60, and 90 degree) for 1000 trails. We can get the results of variance, mean value and standard deviation divided by the total numbers of samples ($m = 50$) in Table VII.

TABLE VII
RESULTS OF SIMULATION -6

Area	Value	$\hat{b}_1 - b_1$	$\hat{b}_2 - b_2$	$\hat{b}_3 - b_3$
30	Variance	5.860352864	0.420686763	6.280116097
	Mean	-0.005949672	-0.00186178	-0.006367917
	SD/N	0.004841633	0.001297207	0.005012032
60	Variance	4.734625413	1.576340094	6.310272519
	Mean	-0.022460129	-0.012659197	-0.025872831
	SD/N	0.004351839	0.002511048	0.005024051
90	Variance	3.680904428	3.682455962	7.365952638
	Mean	-0.008842285	-0.008593189	-0.012242425
	SD/N	0.003837137	0.003837945	0.005428058

Simulation based on 1,000 trials on sample size of N=10,201. House price has three distance variables in data producing process. SD/N means Standard Deviation divided by N.

Further, to confirm the relationship of the correlation coefficient R_{12}, R_{13} and R_{23}, we build a regression with the distance d_3 and the other two distance d_1 and d_2. If the coefficient of d_1 and d_2 are β_1 and β_2, β_0 denotes the constant term, then d_3 can be written as the following (7). We set θ=30 degree. We take 500 observations of d_1 and d_2 that following the uniform distribution (tinny region method) randomly. Running the regression, we get the result in Table VIII.

TABLE VIII
RESULTS OF REGRESSION-3

30 degree	Coefficients	Standard Error	t Stat	P-value
Intercept	3.99993314	0.000266201	15025.98452	0
X Variable 1	-0.965945741	0.000134428	-7185.586171	0
X Variable 2	-0.258566108	0.000135428	-1909.247837	0

Regression based on sample size of N=10,201. House price has three distance variables in data producing process.

This result showed that distance variable d_3 does have liner relationship with d_1, d_2 in a tiny study region context. Basically, the influence on d_3 from d_1 and d_2 are different, the effect from d_1 is stronger, with a constant term close to 4.

Adjusting the regression parameters to: $\beta_0 = 4$, $\beta_1 = -0.97$, $\beta_2 = -0.26$. d_3 could be calculated by (7). Running the regression with d_1, d_2 and d_3 for 1000 trails. We can get the variance, mean value and standard deviation divided by N of $\hat{\beta}_0 - \beta_0$, $\hat{\beta}_1 - \beta_1$ and $\hat{\beta}_2 - \beta_2$ in Table IX.

TABLE IX
RESULTS OF SIMULATION -7

Area	Value	$\hat{\beta}_0 - \beta_0$	$\hat{\beta}_1 - \beta_1$	$\hat{\beta}_2 - \beta_2$
30	Variance	0.008670892	0.002147138	0.002127137
	Mean	-0.055189441	-0.001963819	0.257514868
	SD/N	0.000186235	9.26744E-05	9.22418E-05

Simulation based on 1,000 trials on sample size of N=10,201. House price has three distance variables in data producing process. SD/N means Standard Deviation divided by N.

From this result, we could figure out this regression is stable, for the variance and the mean value are small. d_3 has linear relationship with d_1, d_2, and it can be described by (17) in a tiny context. The influence from d_1 and d_2 on d_3 are -0.97 and -0.26, with a constant number 4.

$$d_3 = 4 - 0.97d_1 - 0.26d_2 \quad (17)$$

We did the same analysis when θ was taken on 60 degree. The regression result was laid in Table X

TABLE X
RESULTS OF REGRESSION -4

60 degree	Coefficients	Standard Error	t Stat	P-value
Intercept	3.999529202	0.000265971	15037.45428	0
X Variable 1	-0.865764531	0.00013413	-6454.690115	0
X Variable 2	-0.499955944	0.000141829	-3525.053928	0

Regression based on sample size of N=10,201. House price has three distance variables in data producing process.

TABLE XI
RESULTS OF SIMULATION -8

Area	Value	$\hat{\beta}_0 - \beta_0$	$\hat{\beta}_1 - \beta_1$	$\hat{\beta}_2 - \beta_2$
60	Variance	0.122003719	0.030430334	0.028668118
	Mean	-0.089194122	0.014529395	0.263029106
	SD/N	0.000698581	0.000348886	0.000338633

Simulation based on 1,000 trials on sample size of N=10,201. House price has three distance variables in data producing process. SD/N means Standard Deviation divided by N.

Adjusting the regression parameters to: $\beta_0 = 4$, $\beta_1 = -0.87$, $\beta_2 = -0.5$. d_3 could be calculated by (7). Running the regression with d_1, d_2 and d_3 for 1000 trails. We can get the variance, mean value and standard deviation divided by N of $\hat{\beta}_0 - \beta_0$, $\hat{\beta}_1 - \beta_1$ and $\hat{\beta}_2 - \beta_2$ in Table XI.

From this result, we could consider this simulation is stable, since the variance and the mean value are very small. d_3 has

linear relationship with $d1, d2$, and it can be described by (18) in a tiny context. The influence from $d1$ and $d2$ on $d3$ are -0.87 and -0.5, with a constant number 4.

$$d_3 = 4 - 0.87d_1 - 0.5d_2 \quad (18)$$

Same analysis was performed when θ was taken on 90 degree. The regression result was laid in Table XII.

TABLE XII
RESULTS OF REGRESSION -5

90 degree	Coefficients	Standard Error	t Stat	P-value
Intercept	4.000227861	0.000265451	15069.5336	0
X Variable 1	-0.866079426	0.00013201	-6560.718596	0
X Variable 2	-0.500110073	0.000139732	-3579.063028	0

Regression based on sample size of N=10,201. House price has three distance variables in data producing process.

Adjusting the regression parameters to: $\beta_0 = 4$, $\beta_1 = -0.87$, $\beta_2 = -0.5$. $d3$ could be calculated by (7). Running the regression with $d1, d2$ and $d3$ for 1000 trails. We can get the variance, mean value and standard deviation divided by N of $\hat{\beta}_0 - \beta_0$, $\hat{\beta}_1 - \beta_1$ and $\hat{\beta}_2 - \beta_2$ in Table XIII. This result also showed the simulation is stable, the variance and the mean value are tiny. $d3$ has liner relationship with $d1, d2$, and it can be described by (19) in a tinny context. The influence from $d1$ and $d2$ on $d3$ are -0.87 and -0.5, with a constant number 4.

$$d_3 = 4 - 0.87d_1 - 0.5d_2 \quad (19)$$

The result showed positive variance and very tiny mean value as we expected, which means the fluctuation of $b1$, $b2$ and $b3$ are acceptable, the regression was comparatively stable.

Based on the results we can figure out that when considering the barycenter of $\triangle Q_1Q_2Q_3$ as the center of the tiny observation region, hedonic regression is comparatively stable even the distance variables are increased to three. The variance and mean value showed the fluctuation of $b1$, $b2$ and $b3$ are quite tinny and smooth. This result fits our correlation analysis, and the observations are easy to locate. This is quite important in the research when we are dealing with real database. And the correlation analysis showed that $d3$ has a linear relationship with $d1, d2$. When the value of θ is changing from 30 to 90 degree, the effect from $d1$ on $d3$ is increasing, the effect from $d2$ decayed. And there is a constant term 4 in this model.

V. WIDE REGION SIMULATION

Three Variables Wide Region Simulation

In real database, it will be difficult to gather the samples of the whole observation area. When we are facing the hedonic regression, we have to estimate the house price with some specific region's data. So we have built another wide region regression model to demonstrate the sampling area's location.

In Simulation-10, two circles centered on origin were drawn in Cartesian coordinate as Fig. 11, their radius were taken as r and $3r$. For calculation convenience, r was also taken as equal to 1. Observation region is set as uniform distribution. $Q1(-1,0), Q2(1,0)$ are the first and second urban nodes, r is the half of segment $Q1\ Q2$. $Q3$ is the third node. A is our sampling region's center. And point A was taken just on the radius $r = 1$ circle. If the angle of $\angle SOQ2$ is taken as θ, the coordinate of uniform distributed sample point $A(x, y)$ could be written as $A(\cos\theta, \sin\theta)$. Following the trigonometry $Q3$'s coordinate can be written as $Q3(3\cos\theta, 3\sin\theta)$.

TABLE XIII
RESULTS OF SIMULATION -9

Area	Value	$\hat{\beta}_0 - \beta_0$	$\hat{\beta}_1 - \beta_1$	$\hat{\beta}_2 - \beta_2$
90	Variance	0.237805762	0.062494302	0.05737357
	Mean	-0.080127798	0.113045212	0.260190703
	SD/N	0.000975307	0.000499977	0.000479056

Simulation based on 1,000 trials on sample size of N=10,201. House price has three distance variables in data producing process. SD/N means Standard Deviation divided by N.

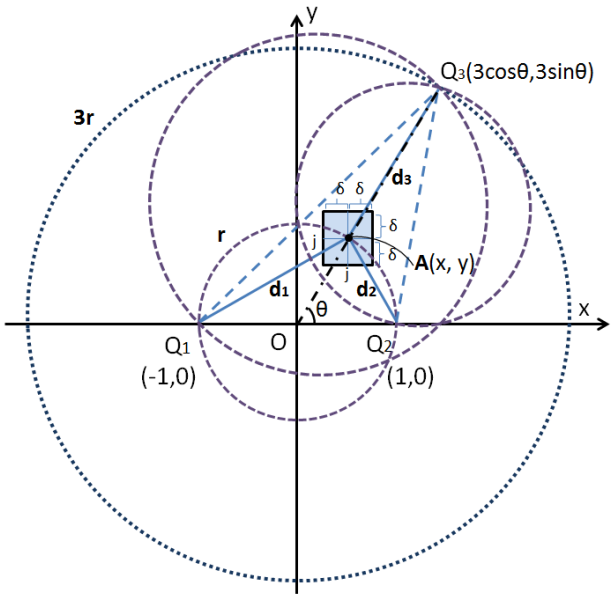

Fig. 11 Illustration of observation region of Simulation-10

Based on the tiny region regression analysis and Simulation-2 and Simulation-3's process, we adjusted the Simulation-6's range δ equal to 0.5, so the region could be considered as a wide region. We marked this regression as Simulation-10. Before performing the regression, we overlapped the correlation coefficient. And we could see the sampling area is on the edge of this overlapping region. The shadow region is marked in Fig. 11.

Make sure other parameter settings as same as Simulation-6. $a = 2500$, $b_1 = -130$, $b_2 = -100$, $b_3 = -80$, $e = 0.001$. The distance d_{1i}, d_{2i}, d_{3i} can be calculated by (10), (11) and (12). House price can be calculated by (1). Run the regression on 30, 60 and 90 degree for 1000 trails. The result is listed in Table XIV.

TABLE XIV
RESULTS OF SIMULATION -10

Area	Value	$\hat{b}_1 - b_1$	$\hat{b}_2 - b_2$	$\hat{b}_3 - b_3$
30	Variance	7.03454E-07	8.41138E-08	7.5162E-07
	Mean	-1.7953E-05	-5.17736E-06	-1.4284E-05
	SD/N	1.67744E-06	5.80048E-07	1.73392E-06
60	Variance	8.2945E-07	2.86064E-07	1.07257E-06
	Mean	-6.55318E-06	-4.93504E-06	-8.86928E-06
	SD/N	1.82148E-06	1.0697E-06	2.0713E-06
90	Variance	5.58429E-07	5.70791E-07	1.1153E-06
	Mean	-3.9663E-06	2.58704E-06	1.21988E-06
	SD/N	1.49456E-06	1.51101E-06	2.11216E-06

Simulation based on 1,000 trials on sample size of N=10,201. House price has three distance variables in data producing process. SD/N means Standard Deviation divided by N.

The regressions showed positive result, the variance, mean value and standard deviation, which is divided by N of the three distance variables, are as small as we expected. They also proved our hypothesis in the correlation analysis. Smaller variance, mean value and standard deviation denote the regressor coefficients' changes are very tiny and acceptable.

Based on the result of correlation analysis, the distance d_3 has a liner relationship with d_1, d_2. And when θ is equal to 30, 60 and 90 degree, the linear equation in wide region will be the same as the tiny region context. We calculated them by the same process in the tiny region analysis, and the coefficients fit (17)-(19). So when the value of θ is changing from 30 to 90 degree, the effect from d_1 on d_3 is increased, and the effect from d_2 decayed. And there is a constant term 4 in this wide region model.

Since the simulations showed positive result when it was realized following the barycenter method. If researchers adopt the data to specific area with three urban nodes, the data on or close to the barycenter distributed by the three nodes are our suggestion.

VI. CONCLUSION AND DISCUSSION

In hedonic pricing method, regression usually employs a variable such as the distance measured from one or some important locations that we considered as a station or an amenity/ hazard in the urban context. When the variables are considered as more than two, terrible multicollinearity will cause fluctuations in regression. Unfortunately these distance variables are not so valid. When the observing location is in some specific area such as the region close to one urban node or sharing the same line with two nodes, multicollinearity problem will appear because of the huge bias caused by these invalid data. They will cause the instability and error in regression such as magnitude variance and mean value.

In the estimation of three specific nodes in space, both wide region and tiny region simulation achieved acceptable variance and mean value when we follow the barycenter sampling method. It is convenient to locate the sampling area, and it fits the demand that we should take samples in the overlapping region with minimum absolute correlation value. Simulation results also showed that the Hedonic price model is comparatively stable under the condition of three distance variables. And considering the correlation coefficients in the regression can lower the bias in analysis, it is possible to reduce the potential collinearity problem. So researchers should avoid gathering the samples as:

1) The samples located at the area are too far away from the three facilities which are considered as specific urban node or amenity/hazard (Q_1, Q_2, Q_3).
2) The samples located at the area are just right on or close to the three urban nodes (Q_1, Q_2, Q_3).
3) And the neighborhood or the location is right on the midpoints of each two nodes.
4) The samples are far from the shadow region that overlapped by the correlation R_{12}, R_{13}, R_{23} circles.

We suggest gather the data as the following methods:

1) Sampling the data which are located on or inside the shadow region which is considered as the region with comparatively small correlation. This area can be located by the correlation analysis.
2) When the sampling area is limited to some specific location, we should take the observation which is located on or close to the barycenter of $\triangle Q_1Q_2Q_3$. Setting θ (the angle of $\angle SOQ_2$) equal to 30, 60 and 90 degree, which can help us locate these sampling region easily.
3) Moreover, when considering add a third node into a two node model, locate the third node on the circle that radius was taken as three times of the circle that passed through the original two nodes on the base. This distribution can also lower the d_2 's effect on d_3. Lower the multicollinearity cause by this three distance variables. And the barycenter area we described are easy to be identified in real world database. They are more effective than other data in regression.

REFERENCES

[1] Lancaster, K. J. *"A new approach to consumer theory", Journal of Political Economy 74,* 1966, pp.132–57.
[2] Sherwin Rosen, *"Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition", Journal of Political Economy, Vol. 82, No. 1,* Jan. - Feb., 1974, pp. 34-55
[3] Silvey, F. *"Multicollinearity and Imprecise Estimation", Journal of the Royal Statistical Society, Series B, 35,* 1985, pp.99-115
[4] Coo, R. D. *"Detection of influential observation in linear regression", Technometrics,19,* 1977, pp.15-18
[5] Herl, Arthur E. and Robert W. Kennard. *"Ridge Regression: Biased Estimation of Nonorthogonal Problems", Technometrics, 12,* 1970a. pp. 55-67.
[6] Hoaglin, D. C., and Welsch, R. E. *"The hat matrix in regression and ANOVA", Amer. Statistician,* 1978, 3217-22.

[7] Neter, J, Kun Neter, J., Kutner, M. H., Nachtshein, C. J.,Wasserman,W., *"Applied Linear Regression Models",3rd ed. Homewood III,* Irwin. 1996.
[8] Danel A. Griffiths, *"Modeling Urban Population Density in a Multi-centered City." Journal of Urban Economics 9,* 1981. pp.298-310.
[9] Eric Heikkila, *"Mullticollinearity in Regression Models with Multiple Distance Measures", Journal of Regional Science, vol.28. No.3,* 1998.
[10] M. Tiefelsdorf, *"Misspecifications in interaction model distance decayrelations: A spatial structure effect" J. Geograph Syst*, 2003, 5: pp25-50
[11] Jaewoo Lee, Keemin Sohn, *"Identifying the Impact on Land Prices of Replacing At-grade or Elevated Railways with Underground Subways in the Seoul Metropolitan Area", Urban Studies Journal Limited,* 2013, DOI:10.1177/0042098013484543
[12] Heather A Sander, Debarchana Ghosh, David van Riper, Steven M Manson, *"How do you measure distance in spatial models? An example using open-space valuation", Environment and Planning B: Planning and Design, volume 37,* 2010, pp.874 – 894.
[13] Justin M. Ross, Michael C. Farmer, Clifford A. Lipscomb, *"Inconsistency in Welfare Inferences from Distance Variables in Hedonic Regressions", Journal of Real Estate Finance and Economics, 43,* 2011, pp.385-400.
[14] Hao Huang, Li Yin, *"Creating sustainable urban built environments: An application of hedonic house price models in Wuhan, China", Springer Science + Business Media Dordrecht,* 2014 J, *House and the Built Environ DOI* 10.1007/s 10901-014-9403-8, *Urban Geography, Vol. 35, No.3,* 2014, pp. 420-439
[15] D. Schmeidler, *"The nucleolus of a characteristic function game", SIAM Journal of Applied Mathematics, 17*, 1969, pp. 1163–1170