博士論文（要約）

Method for Introducing Multiple Distance Variables in Hedonic Analysis and Its Application to Real Estate Price Analysis

（ヘドニック分析における複数の距離変数の導入方法と不動産価格分析への適用）

WANG YAN

王　焔

論文題目　**Method for Introducing Multiple Distance Variables in Hedonic Analysis and Its Application to Real Estate Price Analysis**

（ヘドニック分析における複数の距離変数の導入方法と不動産価格分析への適用）

氏　　名　**WANG YAN**

王　焰

In urban context, urban nodes that are considered as amenity or hazard facilities will affect house price and rents. One of the classic evaluations on this topic is the hedonic pricing method. In classic hedonic analysis, distance variables that are measured from each urban landmark are employed for the analysis. However, the effects of distances to these urban nodes on house prices sometimes do not represent the true prices of the households. Distance variables that have been measured on the same urban area suffer from problems such as spatial multicollinearity, which is usually presented in regression results as magnitude variance and mean value of the parameters. Also, this unstable element will lead to ill analysis result, so in the past years, hedonic pricing method is becoming unpopular for the blemish caused by multiple variables. In this research, I provided an estimation system to identify and choose the data with less bias, and a specific sampling method for locating the sample area to avoid the spatial multicollinerity problems for the two and three distance variable's case of the analysis of urban real estate. Comparing to other researchers, I improved the hedonic pricing model by using a simpler method rather than using statistical methods and mathematical solutions.

This dissertation is divided into eight chapters. After the introduction of the contents, the background of hedonic pricing model is elaborated in Chapter 1. Then, literature related to the sampling method and function transformations of hedonic pricing model is reviewed. Two research studies that inspired me a lot are introduced in detail in chapter 2. A sampling-based method for achieving the research goal of solving the multicollinearity problem and stabilizing the hedonic regression is proposed, and the introduction of the research method based on previous works is laid out in Chapter 3. From Chapter 4 to Chapter 5, the circumscribed circle sampling method and correlation analysis based on theoretical data base are introduced. Following the order of two nodes case and three nodes case, the simulations are conducted both under the context of wide region and tiny region. The sample size raged from 500 to 1000, with each simulation run for 1000 trails. The

coefficients of distance variables are estimated to assess the stability of the hedonic regression. In Chapter 6 the data selection principles and the major sampling rules of the two nodes case and three nodes case are proposed. To verify the theoretical hypothesis with the real data base, two data sets of Tokyo area are used for different simulations. The results of the two nodes case and three nodes case both showed positive reaction, and supported the recommendation that have been made in the ideal model stage. Unlike the ideal data analysis, the actual data set will be much more complicated for the situation of the urban context. Therefore not only the distance variables but also other variables should be involved in the regression. Still, most of the multicollinearity problem that begot from the distance variables is solved by the circumscribed ring method. Based on the analysis results, conclusion, limitations and discussion are provided in the last chapter.

Considering the correlation coefficients of the distance variables can identify the area with less bias, and using a sampling method, which takes the samples right on or in the neighborhood of the circumcircle of the specific urban nodes, reduce the bias caused by the observation itself in regression. Furthermore, to reduce the potential collinearity problem, researchers should avoid using data located in the distant area and the location right on or close to the nodes and the midpoint of these two types of nodes. In wide region sampling method, taking samples both inside and outside the circle that passed through the nodes will reduce the bias, the high-valued correlation area should be avoided too. When the database is built in a limited area, fixing the value of θ to specific constants, will simplify the calculation. In tiny region method, the area near the circumscribed circle of the two nodes should be sampled, and the tiny area located inside this circle with high density and the tiny area located outside this circle can be sampled simultaneously under the precondition with a certain number in each area. These deliberately chosen samples can cancel out the multicollinearity with each other mathematically and stabilize the regression.

Furthermore, when triple urban nodes are considered in the estimation of urban space, the multicollinearity problem from the regressor will exacerbate the influence in hedonic regression. Simulation showed positive results and the third node should be identified carefully. As simulations outcome has proven, samples with negative and positive correlation coefficient signs can cancel out to resolve the multicollinearity problem. However the sampling area should be selected carefully for this cancellation depending on the complexity of the three distance variables correlation. Following the circumcircle sampling method with three urban nodes in both in wide region and tiny region context, researchers could take samples right on the circumscribed circle of the nodes or sample inside and outside this circle simultaneously. The samples located in the distant area from the three urban nodes and the samples located on or in the neighborhood of the three sides of the triangle formed by these three nodes should be avoided. Also, in actual database analysis, taking a range in this circle’s neighborhood will form a ring area, which will simplify the process of identifying the sampling area. In tiny region sampling method, the correlation should be calculated first, and the area that is located on the circumscribed circle of the three nodes is recommended.

In addition, when researchers are targeting real database in hedonic analysis, the recommendation is to follow the wide region context except in the situation when the density of the data fits the sampling minimum limitation. Value of variance inflation factors (VIF) can clarify the correlation of the distance variables, so monitoring this value and making sure that it is under the acceptable range when adjusting the range of the sampling ring area will validate the efficiency of the simulation. Since I have proved that the distance variables are have a linear relationship, the area including the three sides of the triangle that are formed by these three nodes will lead to terrible bias in hedonic regression. Also, inside the triangle area, the correlation among the three nodes will be more complicated. Samples should be taken inside and outside of the circumcircle with positive and negative correlation at the same time. But the sampling location should be chosen wisely and carefully. The data analysis of Tokyo area also showed that when we are dealing with the actual data-base, collecting and employing effective samples will help researchers achieve more steady regression and save a lot of effort in the hedonic analysis. This dissertation also offered a method for minimization of the correlation between the distance variables in hedonic regression model. The theoretical model and real data analysis are different. For the theoretical model, the density and distribution of the sample points are quite important for the optimization of the sampling region selection phase. And the data balance method developed from the ideal model realized the area analysis which, correlates high distance variables. This method will be quite useful to researchers for optimizing the balance of the simple location and sample size according to the correlation coefficient signs. Moreover, comparing to other methods, the circumcircle sampling method and the data balance method that have been proposed in this research will help researchers elevate the hedonic analysis to a level of higher validity and accuracy.

A good sampling scheme could ensure that analysis model is more applicable and simplified. Comparing to other methods that optimize and reduce the residual of the data base by statistics or mathematical means, this research provided a workable sampling method to choose more valid data in the first step. This research will be a meaningful reference for research on multiple distance variables in hedonic regression and urban real estate analysis. But limitations of the uniform distribution and potential error of constant number and random error will also influence the hedonic regression result. Besides, there are many random data, which increase the chances of potential error. When we are conducting the hedonic analysis, usually we are using the two dimensional base for the distance measurement, but in the geographical algorithm, there will be loss in the transformation from the latitude and longitude coordinates. In this thesis, the classic measurement of the distance is used, but there are measures such as network distance and Manhattan distance which are quite common in geography analysis, the transformation of the functions will affect the analysis and also the optimization of the correlation coefficient in various ways. In actual database there are some

other ways to stabilize the regression and minimize the error when sampling in the downtown area. This research only offered a direction and basic method of choosing the optimal sampling area to reduce the bias introduced by the multicollinearity among the distance variables.

Keywords: hedonic regression, urban node, distance variables, multicollinerity, collinearity, Tokyo area