International Journal of Engineering & Technology, 7 (2.33) (2018) 186-190

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Housing demand forecast based on income section using model tree technique

Hyoung- Seon Lim [1], Sang-Hyun Choi [2 *]

[1] Dept. Bigdata, Chungbuk National University, Chungdae-ro 1, Seowon-Gu, Cheongju, Chungbuk 28644, Korea
[2] Dept. Management Information Systems, Chungbuk National University, Chungdae-ro 1, Seowon-Gu, Cheongju, Chungbuk 28644, Korea
**Corresponding author E-mail: chois@cbnu.ac.kr*

Abstract

Background/Objectives: Mankiw and Weil modified model, which is mainly used in the field of housing demand, has the problem that the added variable has no linear relationship with the age-specific house demand.
Methods/Statistical analysis: In this research, we tried to complement the existing model by proposing aM-W modified model utilizing the Model tree technique. In addition, many poor people need another analysis that understands the characteristics to live in abnormal houses. And, we tried to avoid this problem by reflecting income section. We compare the performance with existing models using the 2005 and 2010 Population and Housing Cencus data.
Findings: First, the error rate of the M - W modified model is greatly affected by the extreme poverty class and the low income class. Second, overall the performance of the model tree dominates, the performance has further improved to produce more of the nodes.In the middle class in which five nodes were created, the error rate decreased by 89%, and the correlation coefficient increased by 0.2566 with 0.0490.Third, it is more accurate to use the "total of income section predicted values" rather than the existing "entire section predicted value". Fourth, in order to express an accurate section error, we propose to judge "not the total of income section errors" but "total absolute value of income section error".
Improvements/Applications: In this research, there is a limitation that generalization of results is inappropriate. For further research, it is considered appropriate to apply the Random forest method to generalize the results.

Keywords: *Housing Demand Forecast; Mankiw and Weil; Model Tree; Income Section; Population and Housing Cencus.*

1. Introduction

One-member household of Korea 's household composition by number of family members based on 2015 is 27.2%, two-member household is 26.1%, one and two-member household now occupies more than half of total households [186], The government needs to supply housing smoothly. In this situation where the composition of households changes with time, the field of housing demand prediction has been studied steadily by necessity. Many researchers studying housing demand forecasts have used Mankiw and Weil (1989, M-W) model [2] using cross-sectional data of population structure. The M-W basic model is a model that estimates the housing price as an independent variable by age-specific population data, and predicts the dependent variable using the linear regression technique. The function pfor predicting the residential area of either household by the M-W model is as follows.

$$p = \sum_{i=0}^{k} \alpha_i d_i$$

When K is 0, it means an age less than 1 year old, in case of k, it is an age of k or more.
α is the derived coefficient value of age i.
d_i is a population dummy variable at the age I at that age in the area, expressed as a value of 0 or 1. However, the goal of the M - W model is not what type of household is required for how much residential area, but at that point it is how many houses in total for that area are needed. Therefore, the model pthat estimated the total residential area of the area at that time is as follows.

$$P = \sum_{i=0}^{k} \alpha_i D_i$$

And $D_i$is the sum of the population dummy variables for the age i of 1 year old at that time, that area. Mankiw and Weil's house price forecasting model has the advantage of being able to predict house prices by making use of stable population materials and a lot of interest in its performance but, limitations on research that relies heavily on population data and does not use economic indicators at all are raised [3]. Therefore, the basic model of M - W required extended research, and as a result, many researchers proposed a more developed model. Domestic researchers have proposed M-W modification model including income and housing cost, which is the core determinant to more prospects for long-term housing demand systematically [3].
However, the M-W based model can reflect changes in housing demand due to income and housing cost by adding economic indicator variables, but there is a problem of lower predictive power than existing models. The reason is that income and expenses, which are economic indicator variables, are not grasped because they do not have a linear relationship with age-specific house demand [5]. In order to overcome these drawbacks, M-W models using other analytical methods such as nonlinear regression techniques [5] and quantile regression techniques [6] were proposed.

In this research, we tried to complement the explanatory power and predictive power of the M-W modified model using decision trees. The most important part of the decision tree is whether to start discrimination with several variables. Decision trees use the concept of purity to proceed with classification tasks. Purity is the degree to which the same type of data gathered in the group, and the decision tree can be a task of dividing the group so that the purity of the data becomes the fastest. Methods for calculating purity are diverse and differ from the proposed algorithm. The C4.5 algorithm that works well for various problems and recognized versatility uses entropy to measure purity [7].

The Model Tree technique used in this study is an evolution model of the regression tree method for classifying numerical dependent variables by deformation tree [8]. Unlike a regression tree that averages the values of nodes in the tree, the model tree present a linear regression model. It is generally said to exhibit higher performance than regression trees. Existing schemes are not appropriate using one linear regression model in situations where economic indicator variables and age-specific housing demand do not have a linear relationship. However, in order to complement such problems, we consider that a model tree method that creates multiple linear regression models is suitable, and I would like to propose a modified model of M-W applying the model tree method.

Therefore, in this study, the model tree method was used instead of the existing M-W model analysis method. The low-level prediction and explanation of the M-W correction model are supplemented by using the model tree method which is expected to achieve high performance.

In addition, the problem of using a single linear regression technique together with all income sections is not limited to this. Looking at the survey of residential floor households conducted since 2005 [1], the total of abnormal housing furniture in Seoul City standard basement 355, 427, attic room 34,098, shack, tenement house, plastic house, cave, dugout, etc. in 2005 is 399,530 households. Considering that the number of household in Seoul city is 3,309,890 in 2005, the proportion of abnormal households is about 12% of Seoul city as a whole. In general, if you judge the income group of sub 19% as a low income group and the lower 9% income group as extremely poor, most of them can be expected to be housing in abnormal houses. Also, looking at Housing actual condition and support plan of housing poor [8] announced at the 2004 Korea Research Institute for Human Settlements, we can learn the background of the formation of abnormal housing and the actual condition of housing. The main residents of abnormal houses are mainly low-income households, camouflaged move-in, simple daily workers, homeless people, young children and urban low income groups. Among them, housing poor people who are relatively close to the common people live in a good underground dwelling. In this way, most abnormal housing layers can be judged as financially vulnerable extremely poor and low income people. In order to predict the housing demand of such an abnormal housing, it is judged that it is necessary to understand the characteristics of each type of residence and to conduct additional analysis. Analysis of all income sections together without such additional analysis makes it difficult to expect accurate results, so in this research, we try to analyze all sections of income by section.

2. Materials and methods

2.1. Data

In the survey that can be utilized for building the M - W modification model, there are Korea Housing Survey [10], Household trend Survey [1] and Population and Housing Cencus[1]. Korea Housing Survey and Household trend Survey, which includes household income and housing cost, is not only suitable for constructing a modification model, but also has the advantage that surveys are carried out annually. However, since it was not investigated geographically, it has the disadvantage that analysis by region is impossible. On the other hand, the total investigation of population housing does not include income and housing cost, construction of M-W modification model is not easy there is a disadvantage that the investigation cycle is 5 years, but since the area is created with division the area another analysis is possible. In terms of the characteristics of the M - W modification model in which income is added as a variable, it was judged that the regional division was important, and Population and Housing Cencus data was selected as a model construction material of this research, and the analysis target was concentrated in Seoul City. As mentioned, since population and Housing Cencus can`t find income variables, its prediction is necessary and built an income model. We utilized the data of Household Travel Diary Survey [10] 2010) with similar internal variables to Population and Housing Cencus. Effectiveness is important because housing demand forecast is a field that actually needs to predict the future. Therefore, after building a model using the 2005 Population and Housing Cencus data, I would like to confirm the predictive power through 2010 Population and Housing Cencus data. As in existing families, we did not consider variations in model along time flow.

2.2. Analysis method

First of all, we built an income model and preceded research to estimate income variables (Phase 0). After that, the house demand function was constructed by utilizing the estimated income variable for the M - W modification model (Phase 1). At this phase, we will build a model using the existing linear regression method and the newly proposed model tree method, and compare its performance with other researchers. After that, assign the independent variable of the fiscal year to be predicted (in the case of this research, 2010) to the function, calculate the area of the house, and then predict the total area. (Phase 2). Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) were further confirmed to stably verify the predictive power of the model. First, we analyzed all income sections, and further analyzed income sections. Income section analysis was also done in the same way.

2.2.1. Phase 0

Generally, when building an income model, recognize the model as a censored regression model and use the Tobitmodel[11]. However, in the case of Household Travel Diary Survey, households with no income are categorized and not tabulated in order to investigate income in the section. Therefore, the construction of the income model was recognized and accessed as a general regression model. The model constructed is as follows and explanatory power is 77.3%.

IN-
COME=192.672N2ZONE11010+182.443N2ZONE11020N2ZONE+202.827N2ZONE11030......+ (-62.051House_status3) + (-46.080House_status6) +93.987CAR1

The variables used are as follows.

N2ZONE is an area dummy variable.

Family is the number of household members.

H_job dummy is a dummy variable for classification of occupation such as management / clerical / management, professional / technical worker, sales worker, skilled worker in agriculture and fishery, production / transportation / workers, other occupations.

House_Type dummy is a classification dummy variable of residential type such as single-family house, townhouse, multi-family / multi-family house, other housing.

House_Status dummy is charter, rent, and other housing occupancy classification dummy variable.

CAR 1 dummy is a dummy variable as to whether or not to own a passenger car.

2.2.2. Phase 1

As mentioned earlier, "Phase 1: Construction phase of housing demand forecasting model" used the data of Seoul City for 2005 by Population and Housing Cencus for all of the linear regression technique and model tree technique. A dummy variable of 1 year old unit, a single household dummy variable, and an estimated income variable were set as independent variables, and the area of the house was designated as a dependent variable and analyzed. First, the analysis method of the linear regression M - W modification model is as follows.

1) Linear Regression M - W modification model(LRMW)

Remove the missing value displayed in 999 or NULL value. After applying the linear regression model, search outlier. Exclude outlier from data set and improve performance.

2) Model Tree M-W modification model(MTMW)

Remove the missing value displayed in 999 or NULL value. In order to compare with conditions like a linear regression model, model tree analysis was carried out excluding outlier observed in 1).

After that, the area of the house was calculated by substituting the independent variable of the year to be predicted (in this study, 2010) into the constructed function. We further confirmed the correlation, MAE, RMSE and added stability to the prediction of performance.

2.2.3. Phase 2

In "Phase 2: Housing area prediction stage", Population and Housing Cencus 2010 year was used. Enter the independent variable of 2010 Population and Housing Cencus in the model constructed in Step 1 to calculate the forecast value of housing demand, and we judged the performance compared with the actual demand of housing in 2010.

In order to conduct additional analysis of income section, income distribution confirmation was preceded [Figure 1].

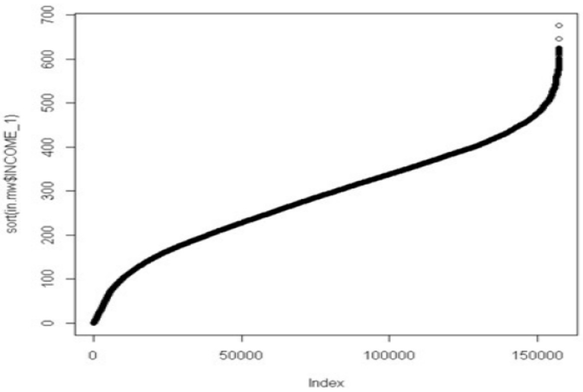

Fig. 1: The Linear Graph Sorted by Income.

Depending on the slope of the graph, the extreme poverty class with monthly incomes of less than 750,000won are low income class with monthly income of 750,000wom or more and less than 1,300,000won, middle class with a monthly income of 1,300,000won or more and less than 4,750,000won, monthly income of 4,750,000won We divided these into high income class and analyzed them.

Similarly, Phase 1 and Phase 2 were performed for the analysis method.

The tool used for the analysis was R.3.3.2 version, and linear regression analysis was progressed using the lm () function, and a model tree analysis was carried out using the m5p () function of the R Weka package.

3. Results and discussion

The function L constructed with LRMW is as follows and the R-Squared value is 0.6369.

L= 8.5025D00 + 11.0102D01 + 10.2745D02 + ... + 10.0687D85 + 119.4532ALONE + 0.0341INCOME

i of Di means the age of the population, which is a household member dummy variable of "when i = 00, under 1 year old", "when i = 85 is over 85 years old". ALONE is a dummy variable with or without one-member household, and INCOME stands for an estimated income variable.

An independent variable for Population and Housing Cencus 2010 is substituted for the derived function, and actual values and predicted values are compared [Table 1].

Table 1: Comparative Analysis of Income Integration Model

	LRMW	MTMW
Actual valuep(m2)	7,477,579	
Predicted valueq(m2)	6,958,488	8,127,567
Erorr;p-q	519,091	-649,988
Erorr rate (%)	6.94	-8.69
Correlation coefficient	0.0801	0.1842
MAE	75.9907	66.1575
RMSE	113.1693	83.9495

In 2010, the actual total residential area was 7,477,579m^2, and the total residential area predicted based on the function derived to LRMW was 6,958,488m^2. The error was 519,091m^2; the correlation was 0.0801, which was a very low level. The model constructed with MTMW is the same as in Fig. 1, and like the LRMW, the actual value is compared with the predicted value. The estimated value was 8,127,567m^2, and it gave an error of 649,988 m^2 compared with the actual value of 7,477,579m^2. The model of MTMW yielded an error of 130,897m^2 from the error area of LRMW difference from 519,091m^2. However, it is difficult to judge that the value of the total area is a value obtained by adding all the area values of individual household and that the predicted value was a more accurate analysis approximating the actual value. The correlation, MAE, RMSE that can mathematically confirm the degree of error MTMW is more dominant. To further analyze this, we analyzed the income section in consideration of the characteristics of the poor according to the preliminary household. In order to advance income segment analysis, we first confirmed the income section [Figure 2].

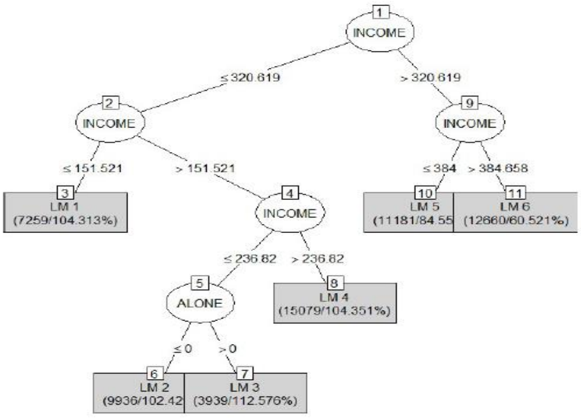

Fig. 2: Income Integrated Model Tree M - W Modified Model (MTMW).

Depending on the slope of the graph, the extreme poverty class with monthly incomes of less than 750,000won are low income class with monthly income of 750,000wom or more and less than 1,300,000won, middle class with a monthly income of 1,300,000won or more and less than 4,750,000won, monthly income of 4,750,000won We divided these into high income class and analyzed them.

The results of the analysis are tabulated for comparison [Table 2], [Table 3].

Table 2: LRMW Analysis of Income Section

LRMW	Extreme poverty class	Row income class	Middle class	High income class	Entire section
Actual valuep (m^2)	688,592	719,596	5,830,950	238,441	7,477,579
Predicted valueq (m^2)	919,355	891,280	544,007	245,418	6,958,488
Erorr;p-q	-230,763	-171,684	384,943	-6,977	519,091
Erorr rate (%)	-33.51	-23.85	6.60	-2.92	6.94
Correlation coefficient	0.0634	0.0968	0.0490	0.2683	0.0801
MAE	109.6514	100.6349	75.3812	28.5163	75.9907
RMSE	140.5954	130.165	113.2408	41.5864	113.1693

Table 3: MTMW Analysis of Income Section

MTMW	Extreme poverty class	Row income class	Middle class	High income class	Entire section
Actual valuep (m^2)	688,592	719,596	5,830,950	238,441	7,477,579
Predicted value q(m^2)	866,282	881,684	5,874,012	243,967	8,127,567
Erorr;p-q	-177,690	-162,088	-43,062	-5,526	-649,988
Erorr rate (%)	-25.80	-22.52	0.73	-2.31	-8.69
Correlation coefficient	-0.005	0.0948	0.2566	0.2430	0.1842
MAE	65.8649	80.7017	66.3144	33.5428	66.1575
RMSE	78.9656	93.9837	84.2326	45.4617	83.9495

As a result of comparative analysis of the income section model, in the case of extreme poverty and low income class, both error rates of both models were 2 digits of both, which had higher error rates than the middle class and the high income class. Therefore, it can be thought that poor people have to think about other analysis methods that understand the person's characteristics more than the general M - W model.

It is impossible to correlate directly with extremely poor and low income class, but the main finding is that the error rate of the existing MW model is greatly affected by the poor.

In addition, when analyzed by income section, MTMW performance was found to be high overall. In the case of the extremely poor class when using MTMW, the error rate of about 23% LRMW ratio could be reduced. Also, in the case of the middle class, the error rate was reduced by 89%, and the error rate was the income section layer where the error rate was greatly improved. Even in the low income and high income class, the error rate slightly decreased.

Therefore, the results are different between 'the predicted value of the entire section' and 'the sum of the predicted values of the income section', and 'the sum of the predicted values of the income section' was more accurate. In the case of LRMW, 'the predicted value of the entire section' was 6,958,488m^2, 'the sum of the predicted values of the income section' was 7,502,060m^2, and the error from the actual value was 519,091m^2, -24, 481m^2. In the case of MTMW, 'the predicted value of the entire section' was 8,127,567m^2, 'the sum of the predicted values of the income section' was 7,865,945m^2, and the error from the actual value was -649,988 m^2, -388,366m^2.

In other words, the reason why the error of 'the predicted value of the entire section' is higher than the LMW of MTMW can be grasped from the income section analytically. In the case of LRMW It is judged that the error of -422,447 between the extremely poor and the low income is offset with the error of the middle class + 384, 943, and the relative error rate was relatively small, In the case of MTMW, since the error of the middle class has been greatly improved, it is analyzed that -339,778 of the low income group is penetrated as it is, and the total error is relatively high. In order to express accurate interval errors, we propose to judge not by the sum of income section errors but by the sum of absolute values. The error A of the whole section can be expressed as follows.

A is the error of the extreme poverty class, B is the error of the low income class, C is the error of the middle class, and D is the error of the high income group. When calculating the whole section error of this study by the conventional method, MTMW error is higher in LMMW case 519,091m^2 and MTMW case 649,988m^2. However, as a result of calculation by the proposed method, MTMW is judged to be a more accurate model at LMMW 794,637m^2 and MTMW 388,366m^2. In this way, when analyzing only the entire section as in the conventional model, there are many unrecognized elements, so the result can be incorrectly interpreted. Analysis based on special understanding should be added especially for extreme poverty and low income brackets.

In other words, it is thought that it is desirable to analyze by dividing into income sections rather than the conventional method of analyzing the whole section at a time. In addition, in this study, we observed that MTMW is superior to existing LRMW model in terms of analytical model over all sections. However, in fact, it can`t be asserted that the model tree is suitable for the M-W modified model rather than linear regression, and more research is needed.

4. Conclusion

In this research, in order to complement the limit of the M-W modified model mainly used in the demand forecasting field of housing, we analyzed income section analysis and applying a model tree. Utilizing the data of "Population and Housing Census 2005", we predicted the sum total of residential areas in 2010, and did not consider variations according to time. A summary of the results is as follows. First, the error rate of the M - W modified model is greatly affected by the extreme poverty class and the low income class. In the case of extreme poverty and low income classes, both models had a very high error rate compared with the middle class and the high income class with a two-digit error rate. This can be interpreted in the sense that the M-W model is not suitable for the poor, we must think of other analytical methods that understand their characteristics. Second, overall the performance of the model tree dominates, the performance has further improved to produce more of the nodes. In the middle class in which five nodes were created, the error rate decreased by 89%, and the correlation coefficient increased by 0.2566 with 0.0490. Third, it is more accurate to use the "total of income section predicted values" rather than the existing "entire section predicted value". In other words, "M-W model derived with four income section models" was superior to "M-W model derived with one income section model". Fourth, in order to express an accurate section error, we propose to judge "not the total of income section errors" but "total absolute value of income section error".

However, in this research, there is a limitation that it is irrational to generalize the result with a decision tree based model. In other words, it can`t be generalized that the model tree is always suitable for M - W modified model rather than linear regression, and more additional research is needed. For additional research, it is considered that a random forest method that can complement the shortcomings of this research and generalize the results is appropriate. Also, research based on the characteristics of the poor must be followed.

In this research, M - W modified model was analyzed mainly in the income section, and the model tree method was applied to improve the performance. It is expected to be considered in various ways for researches on housing demand forecast based on M - W modified model.

5. Acknowledgment

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2017-2013-0-00881) supervised by the IITP (Institute for Information & communications Technology Promotion).

The analyzed data were provided by Smart City Research Center in LHI (Land & Housing Institute).

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2017-2013-0-00881) supervised by the IITP (Institute for Information & communications Technology Promotion).

References

[1] http://kostat.go.kr/portal/korea/index.action.

[2] Mankiw, N. G., & Weil, D. N., The baby boom, the baby bust, and the housing market. Regional science and urban economics, 1989, 19(2), pp. 235-258.

[3] Swan, C, Demography and the demand for housing A reinterpretation of the Mankiw-Weil demand variable. Regional Science and Urban Economics, 1995. 25(1), pp. 41-58.

[4] Chung Eui-Chul, Cho Sung-Jin, Demographic Changes and Long-term Housing Demand in Korea,Journal of Korea Planning Association, 2005, 40(3), pp. 37-46.

[5] Choi Seong-ho, Lee Chang-moo, Non-Linear Mankiw-Weil Model on Housing Demand -The case of Seoul Metropolitan Area -, Journal of the Korea Real Estate Analysis Association, 2009,15(2), pp. 117-130.

[6] Kim Mikyoung, Lee Changmoo, Forecasting Distribution of Dwelling Size Using Quantile Regression Model, Journal of the Korea Real Estate Analysis Association, 2015, 21 (3), pp. 45-62.

[7] Quinlan, J. R., C4. 5: Programming for machine learning, Morgan Kauffmann, 1993, 38.

[8] Wang, Y., & Witten, I. H., Induction of model trees for predicting continuous classes, 1996.

[9] Hong In-ok, Housing actual condition and support plan of housing poor, Korea Research Institute For Human Settlements, 2004,PLANNING AND POLICY, pp. 32-40 http://www.dbpia.co.kr/Article/NODE01168209.

[10] http://www.molit.go.kr/portal.do.

[11] McDonald, J. F., Moffitt, R. A, The uses of Tobit analysis. The review of economics and statistics, 1980, pp. 318-321.