http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 13 (13): 2121-2128, 2014
ISSN 1812-5638 / DOI: 10.3923/itj.2014.2121.2128

Study on the Use of Equidistant Binning on Residential Hedonic Price Discretization

[1]Wang Heyong, [1]Hong Ming and [2]Meiling Shyu
[1]Department of E-business, South China University of Technology,
510006, Guangzhou, China
[2]Department of Electrical and Computer Engineering, University of Miami,
FL, United States of America

Abstract: Hedonic price model is the commonly used method in studies of the residential price. Most of the current residential hedonic price models predict the numeric value of residential prices directly, but it is also meaningful to give reasonable intervals to tell the ranges of residential prices. Residential price variables are grouped using equidistant binning and other three discretization techniques for comparison and classification models are built to test the fitness of each discretization techniques in the models. It is the conclusion that equidistant binning is the best.

Key words: Discretization, equidistant binning, residential hedonic price

INTRODUCTION

Valuation of residential prices is an interesting and meaningful issue practically and academically. Housing developers need to have a well-specified criteria to determine sales prices and house buyers Lu (2012) pointed out that currently, it is the most widely used method to determine the estate sales price based on the market price using hedonic price model.

Hedonic price model is commonly applied to residential prices valuation. The model evaluates residential prices from the implicit prices of residential characteristics. It overcomes various defects of traditional methods and can get good results. Various domestic (China) and foreign researches have studied residential price through hedonic price model. Domestically, Wang (2006) analyses urban residential hedonic price model theoretically. Some studies have analysed the factors affecting residential prices in particular regions through hedonic price model and have constructed their particular hedonic price models (Ma and Li, 2003; Wen and Jia, 2006; Guo *et al*., 2006; Hao and Chen, 2007; Zhang and Chen, 2008). Abroad, some studies have discussed the selection of factors affecting residential prices (Butler, 1982; Ozanne and Malpezzi, 1985; Crocker *et al*., 1987).

In practical applications, values of residential prices may be preprocessed so that they are able to satisfy assumptions of a particular model or can be studied in other aspects. The preprocessions can be functional transformation or data discretization, etc. Some of the current studies have been involved in residential price preprocessing. Liu and Zhao (2013) have chosen logarithmic model to show the relationship between residential prices and residential features based on the Enjoy Price Model. Li *et al*. (2011) have classified the residential prices in 12 towns (streets) in Baoshan District in Shanghai. Gu *et al*. (2011) have preprocessed residential prices so that they meet normal distribution. Gu and Xia (2013) have applied a method to forecast real estate prices of some buildings in Nanjing and they have applied hierarchical clustering method to define the state of real estate price in particular time.

Nearly all the current residential price models predict the numeric values directly and these values are approximative, but it is also meaningful to give reasonable intervals to tell the ranges of residential prices. In this study, residential prices are preprocessed using equidistant binning and other three discretization techniques and then residential price classification model are built using decision tree. It proves in experiments that equidistant binning is the best based on empirical data.

SELECTION OF RESIDENTIAL CHARACTERISTIC VARIABLES

The following three types of residential characteristics are commonly considered important (Can and Megbolugbe, 1997; Ma and Li, 2003):

Corresponding Author: Wang Heyong, Department of E-business, South China University of Technology, 510006, Guangzhou, China

Table 1: Selection of residential characteristics

Type	Residential characteristics variables
BSC	Total floors, floor, residential age, residential area, decoration level, orientation, No. of rooms, No. of living rooms, No. of bathrooms, No. of balconies, property type and packing spaces
NEC	Floor area rate, greening rate, public facilities, education facilities, sports and commercial facilities
RC	Area, transport facilities, CBD distance

- Building Structure Characteristics (BSC), such as housing area, orientation
- Regional Characteristics (RC), such as distance from public transportation to the city center
- Neighborhood Environmental Characteristics (NEC), such as residential infrastructures, culture and entertainments

This study has selected 20 residential characteristics including 12 building structure characteristics, 5 regional characteristics and 3 neighborhood environmental characteristics. The variables have been chosen are shown in Table 1.

QUANTIFICATION OF RESIDENTIAL CHARACTERISTICS VARIABLES

Before building residential hedonic price models, non-numerical data need to be quantified. Methods commonly used for quantification include comprehensive index method, numerical method, dummy variables method, Likert scale method and fuzzy mathematics method. The non-numerical residential characteristics variables chosen in this study are quantified using one of the methods mentioned above (Table 2) and the numerical variables retain their original values.

Regional characteristics: In Guangzhou City, economic development level and consumption of residential market in a particular area may differ from the others. Likert scale method is chosen to quantify the following 13 areas in Guangzhou City (Table 3).

Comprehensive index method is chosen to quantify transport facilities according to survey of residential traffic conditions in each area. The original values of this variable are 0 and they increase according to the following rules.

Within 1 km away from the house:

- If there is any subway station, then the values increase by 1
- If there are 10 or more bus lines, then the values increase by 2
- If there are 5 to 9 bus lines, then the values increase by 1

Table 2: Quantification of residential characteristics variables

Type	Variables	Method used
BSC	Floor	Likert scale
	Decoration level	Likert scale
	Orientation	Likert scale
	Property type	Likert scale
	Packing spaces	Dummy variables
RC	Area	Likert scale
	Transport facilities	Comprehensive index
NEC	Public facilities	Likert scale
	Education facilities	Comprehensive index
	Sports and commercial facilities	Comprehensive index

Table 3: Area quantification

Area	Quantified value
Yuexiu	3
Tianhe	3
Haizhu	2
Panyu	2
Liwan	2
Baiyun	2
Huangpu	1
Zengcheng	1
Huadu	1
Conghua	1
Nansha	1
Luogang	1
Around Guangzhou city	1

Table 4: Transport facilities quantification

Within 1 km away from house	Score
No. of (subway station)≥1	+1
No. of (bus lines)≥10	+2
9≥No. of (bus lines)≥5	+1
No. of (bus lines)<5	+0

Table 5: Floor quantification

Conditions of k	Quantified value
k<1/3	0
2/3>k≥1/3	1
k≥2/3	2

- If there are less than 5 bus lines, then values increase by 0

The rules are summarized in Table 4. Function No. of (a) output the number of a.

For example, if the transport facilities of a house meets condition 1 and condition 2 in Table 4, then the final value of the variable is 3 (= 0+1+2). The greatest value of this variable is 3 and the smallest is 0. The higher score, the transport facilities are better.

Building structure features: Likert scale method is chosen to quantify floor according to the principle that the higher floor, the residential price is higher (except the highest floor). Let k = floor/total, the quantification rules are shown in Table 5.

Likert scale method is chosen to quantify decoration level. The higher decoration level, the value is greater (Table 6).

Table 6: Decoration level quantification

Decoration level	Quantified value
Blank	1
Simple	2
Media	3
Top	4
Luxurious	5

Table 7: Orientation quantification

Orientation	Quantified value
East	1
West	1
South	2
North	2
East-West	2
South-East	3
North-East	3
North-West	3
South-West	3
North-South	4

Table 8: Property type quantification

Property type	Quantified value
Ordinary residential	1
Apartment	2
Villa	3

Table 9: Parking spaces quantification

Have parking spaces	Quantified value
Yes	1
No	0

Table 10: Public facilities quantification

Level	Quantified value
Excellent ($n \geq 10$)	3 or 4
Good ($10 > n \geq 2$)	1 or 2
Not bad ($n < 2$)	0

Likert scale method is chosen to quantify orientation according to traditional customer preferences in China. The better orientation, the value is greater (Table 7).

Likert scale method is chosen to quantify property type (Table 8).

Dummy variables method is chosen to quantify parking spaces (Table 9).

Neighborhood environmental characteristics: Likert scale method is chosen to quantify public facilities. Let n represents number of public facilities. The greater n, the value is greater. The result is shown in Table 10.

Comprehensive index method is chosen to quantify education facilities. The original values are 0 and they increase according to the following rules.

Within 1km away from the house:

- If there is any kindergarten, then the values increase by 1
- If there is any primary school, then the values increase by 1
- If there is any secondary school, then the values increase by 1
- If there is any university, then values increase by 1

Table 11: Education facilities quantification

Within 1 km away from house	Score
Kindergartens	+1
Primary schools	+1
Secondary schools	+1
Universities	+1

Table 12: SCF quantification

Within 1 km away from house	Score
Sports facilities	+1
Integrated shopping malls	+1
Banks	+1
Hospitals	+1
Postal services	+1
Leisure facilities	+1

The result is shown in Table 11.

Comprehensive index method is chosen to quantify Sports and Commercial Facilities (SCF). The original values are 0 and they increase according to the following rules.

Within 1 km away from the house:

- If there is any sports facility, then the values increase by 1
- If there is any integrated shopping mall, then the values increase by 1
- If there is any bank, then the values increase by 1
- If there is any hospital, then the values increase by 1
- If there is any postal service, then the values increase by 1
- If there is any leisure facilities, then the values increase by 1

The rule is summarized in Table 12.

DISCRETIZATION TECHNIQUES

Discretization techniques are applied to group a continuous numeric variable. The commonly used discretization techniques include equidistant binning, tiles, mean/standard deviation and MDLP (Minimum Description Length Principle). This four techniques are applied in this study.

Tiles: The main idea of this technique is to divide data in ascending order into several groups using tiles as separate points.

Mean/standard deviation: Separate points of groups are calculated by means (μ) and standard deviations (σ) of the continuous numeric values of variables. For example, values can be divided into 3 groups with the two separate points being $\mu-\sigma$ and $\mu+\sigma$. Values can also be divided into five groups with the four separate points being $\mu-2\sigma$, $\mu-\sigma$, $\mu+\sigma$ and $\mu+2\sigma$.

MDLP: This technique needs an output variable as a guide to group continuous numeric values. The grouped variables must have a better explain for the output variable.

Equidistant binning: Given the width of a group or number of groups, equidistant binning can divide numeric values into several groups and each group has the same width.

Equidistant binning is a very simple discretization technique but it is sensitive to the number of groups a user gives, so it is necessary to pay enough attention to this issue and to give relatively reasonable number of groups. Furthermore, equidistant binning can be affected by outliers. One way to deal with this situation is to remove outliers before discretization but outliers may be of great significance in some time. Class information is not used in equidistant binning, so it is an unsupervised discretization technique.

EXPERIMENTS

Experimental preparation: The experimental data used in this study are collected by two ways. Data of most of characteristics such as building structure features are collected from professional real estate information websites such as gz.soufun.com. Data of some regional characteristics and some neighborhood environmental characteristics can not be collected directly and they are collected from Baidu Map and Guangzhou Electric Map. Totally, 407 complete records are collected.

Analysis software used in experiments is SPSS Clementine.

Experimental steps: The experimental steps are as followed:

- To separate data set into training set and testing set by a particular proportion
- To check the distribution of residential price variable of the whole data set and its subsets (trainings set as well as testing set)
- To group the residential price variable using equidistant binning, tiles, mean/standard deviation and MDLP
- To build decision tree models using C5.0, C and R, CHAID and QUEST according to training set
- To generate the classification accuracy of models built in (3) according to testing set
- To change the proportion of training set of the whole data set and to repeat the above steps

Table 13: Proportions of training sets and testing sets

Proportion of training set (%)	No. of records in training set
50	209
60	229
70	285
80	320

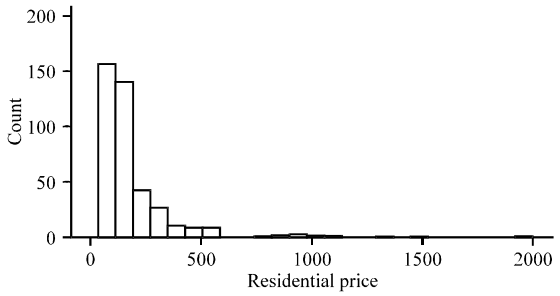

Fig. 1: Distribution of residential price variable of the whole data set

Separating data set into training sets and testing sets: Models are built according to the following four proportions of the whole data set (Table 13).

Residential hedonic price variable discretization: Before grouping residential price variable, it is necessary to observe its distribution. The distribution of residential price variable of the whole data set is shown in Fig. 1.

It indicated in Fig. 1 that the distribution of residential price variable is skewed to the right and the distribution of residential price variables of training sets and testing sets are required to be the same as that in the whole data set, that is, also to be skewed to the right so that different experiments are comparable. Figure 2-5 show the distributions of residential price variables of training sets and testing sets. Figures on the left show the distributions of testing sets and the right show the distributions of training sets.

Form the distributions of training sets and testing sets shown above, it is obvious that most of the values are located in the interval [0, 500] and very few of them are located in the interval [500, 2000] sparsely. One way to group residential price variable is to separate them into three groups, the "high" group, the "median" group and the "low" group. For equidistant binning, three groups are achieved by setting value of its parameter to 3 directly. For tiles, three groups are achieved by setting its parameter to "tertiles". For mean/standard deviation, $\mu-\sigma$ and $\mu+\sigma$ are determined as separate points. However, for MDLP, data can only separate into two groups. Area variable is chosen as the output variable because residential prices in different areas are obviously different (Fig. 6).

The discretization results are shown in Table 14.

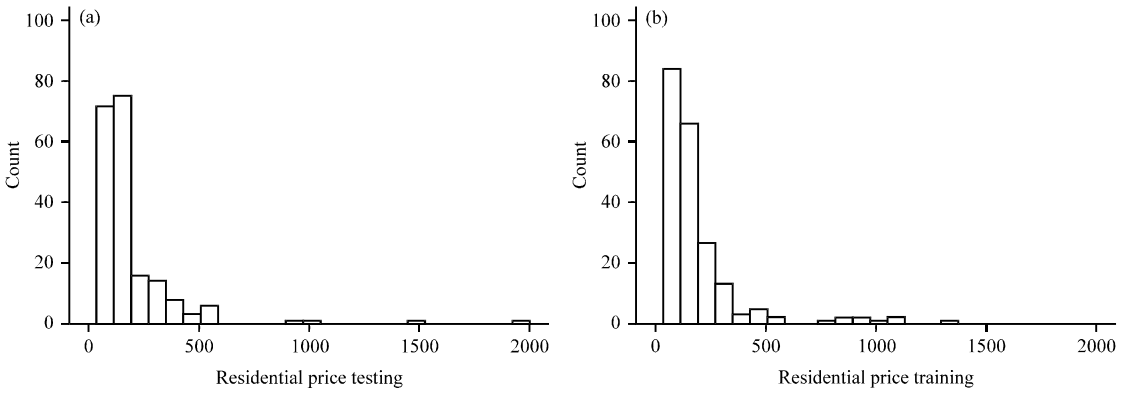

Fig. 2(a-b): Proportion of training set: 50%. Residential price (a) Testing and (b) Training

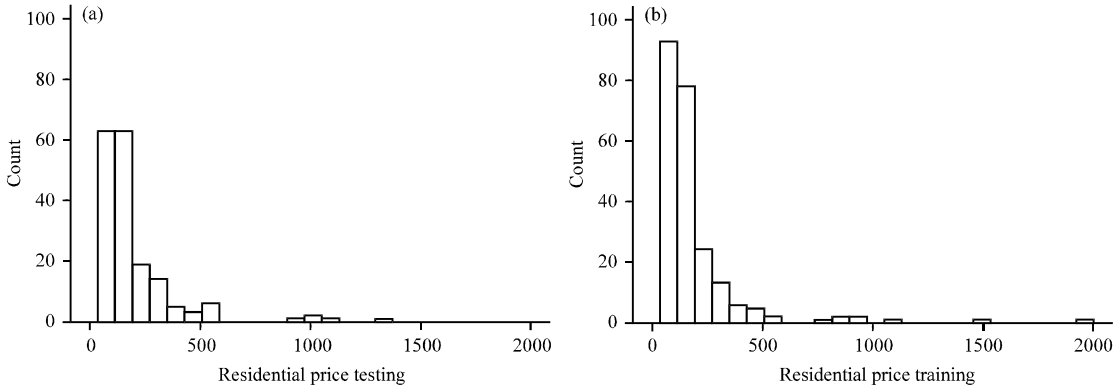

Fig. 3(a-b): Proportion of training set: 60%. Residential price (a) Testing and (b) Training

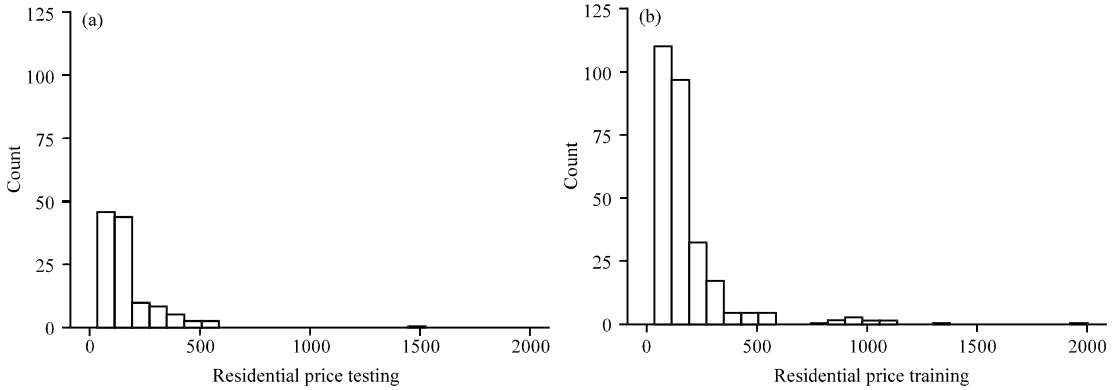

Fig. 4(a-b): Proportion of training set: 70%. Residential price (a) Testing and (b) Training

Table 14: Discretization results

Discretization techniques	Discretization criteria	Result
Equidistant binning	High, median, low	(1344,1999) (689,1344) (34,689)
Tiles		(175,1999) (105,175) (34,105)
Mean±SD		(393.25, +∞) (-10.42,393.25) (-∞,-10.42)
MDLP	Area	(0,102) (102,1999)

Experimental results: Decision tree models are built according to training sets with different proportions using C5.0, C and R, CHAID and QUEST. Corresponding testing sets are used to test the classification accuracy of the models built by training sets with different proportions. The results are summarized in Table 15-18.

Figure 7-10 are based on data in Table 15-18.

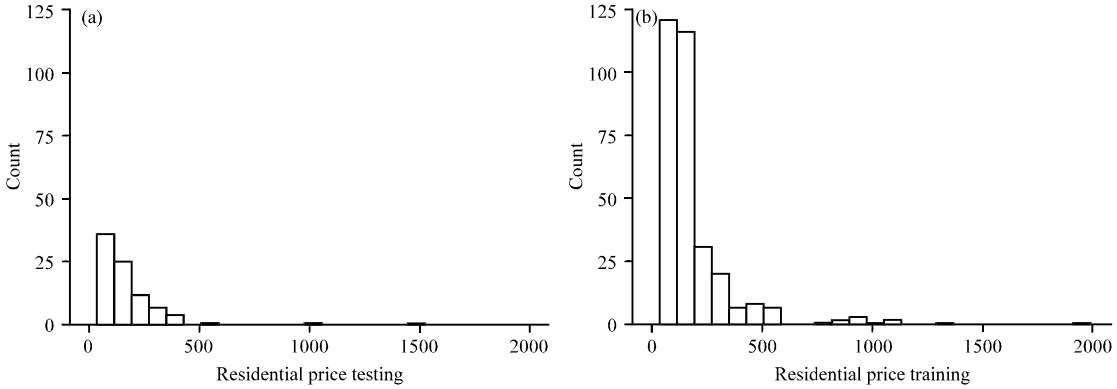

Fig. 5(a-b): Proportion of training set: 80%. Residential price (a) Testing and (b) Training

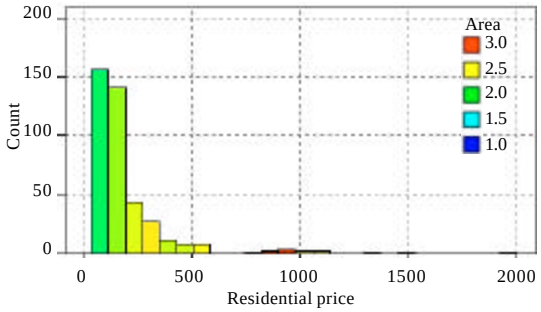

Fig. 6: Residential prices in different areas

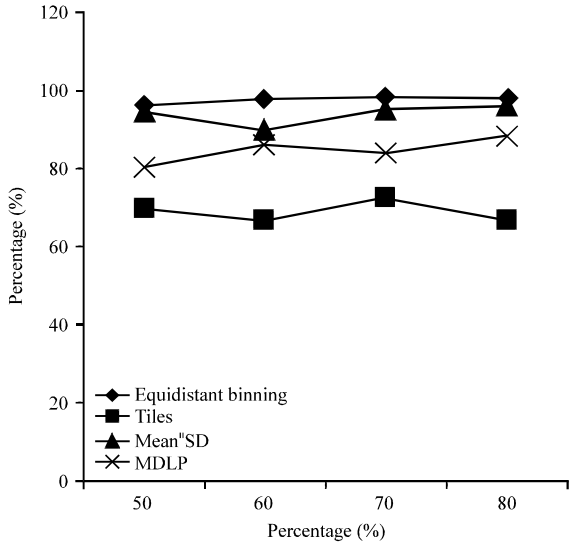

Fig. 7: Classification accuracies of testing sets using C5.0 decision tree models

Table 15: Classification accuracies of testing sets of C5.0 decision tree models

Discretization technique	50%	60%	70%	80%
Equidistant technique	96.46	97.75	98.36	97.70
Tiles	69.70	66.85	72.31	66.67
Mean±SD	94.95	89.89	95.08	96.55
MDLP	80.30	86.52	84.43	88.51

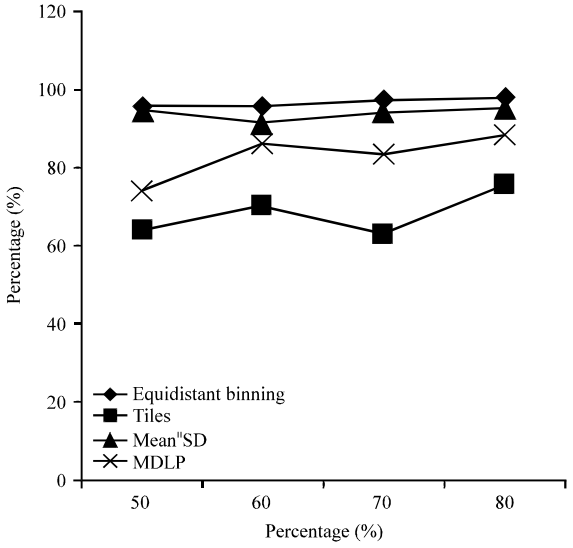

Fig. 8: Classification accuracies of testing sets using C and R decision tree models

Table 16: Classification accuracies of testing sets of C and R decision tree models

Discretization technique	50%	60%	70%	80%
Equidistant technique	95.96	96.07	97.54	97.7
Tiles	64.14	70.22	63.11	75.86
Mean±SD	94.95	91.57	94.26	95.4
MDLP	74.24	86.52	83.61	88.51

Table 17: Classification accuracies of testing sets of CHAID decision tree models

Discretization technique	50%	60%	70%	80%
Equidistant technique	97.98	97.19	95.90	97.70
Tiles	60.10	62.92	64.75	67.82
Mean±SD	95.96	93.26	91.80	95.40
MDLP	80.30	80.34	79.51	88.51

Table 18: Classification accuracies of testing sets of QUEST decision tree models

Discretization technique	50%	60%	70%	80%
Equidistant technique	97.98	97.19	98.36	97.70
Tiles	59.60	60.67	62.30	62.07
Mean±SD	94.95	94.38	95.08	94.25
MDLP	80.81	80.90	82.79	86.21

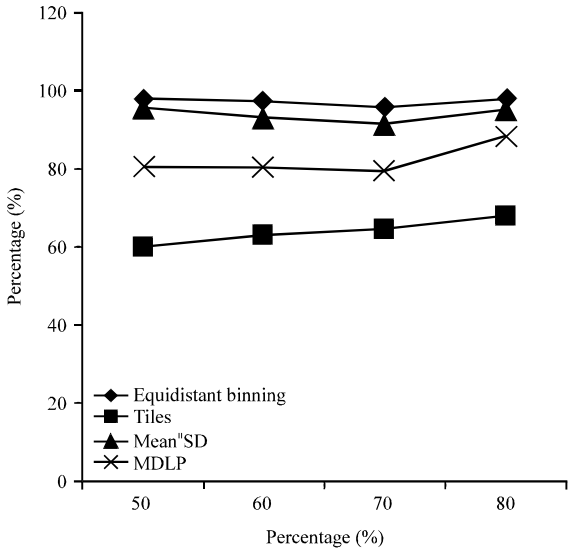

Fig. 9: Classification accuracies of testing sets using CHAID decision tree models

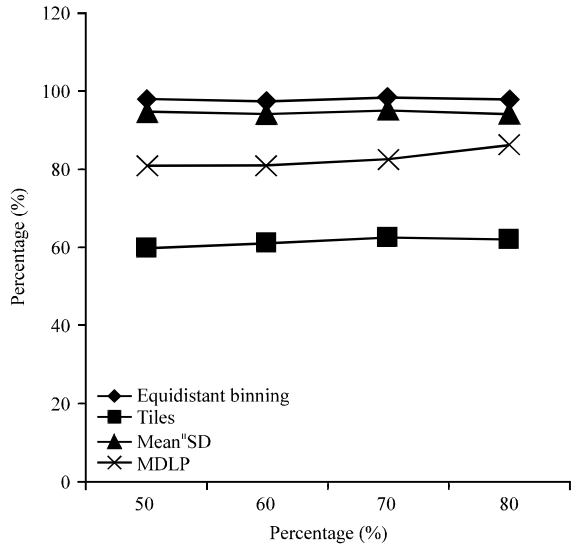

Fig. 10: Classification accuracies of testing sets using QUEST decision tree models

Figure 7-10 show the same result: For testing sets in this study, equidistant binning is the best, the second best is Mean±SD, MDLP is better and the last is tiles.

CONCLUSION

This study studies the application of equidistant binning in residential hedonic price model. First, residential price variable is grouped using equidistant binning and other three discretization techniques for comparison. Then, decision tree classification models are built and test the classification accuracies of testing sets are generated to compare the fitness of different models to prove whether equidistant binning is better than the other three discretization techniques. Finally, experimental results indicated that equidistant binning is the best.

ACKNOWLEDGMENTS

This research was supported by Project of National Social Sciences Foundation, Grant No. 13BTJ005, Social Science Foundation for the Youth Scholars of Ministry of Education of China, Grant No. 10YJC630236, the Fundamental Research Funds for the Central Universities, Grant No. 2013XZD01, supported by the Guangdong Province Science and Technology Fund, Grant No. 2012B091100309 and 2012B040500010.

REFERENCES

Butler, R.V., 1982. The specification of hedonic indexes for urban housing. Land Econ., 58: 96-108.

Can, A.T. and I.F. Megbolugbe, 1997. Spatial dependence and house price index construction. J. Real Estate Fin. Econ., 14: 203-222.

Crocker, J., L.L. Thompson, K.M. McGraw and C. Ingerman, 1987. Downward comparison, prejudice and evaluations of others: Effects of self-esteem and threat. J. Personality Soc. Psychol., 52: 907-916.

Gu, J., M. Zhu and L. Jiang, 2011. Housing price forecasting based on genetic algorithm and support vector machine. Expert Syst. Appl., 38: 3383-3386.

Gu, Y. and L.T. Xia, 2013. The application of weighted Markov chain on forecasting in real estate price. J. Chongqing Univ. Technol. (Nat. Sci.), 27: 125-130.

Guo, W.G., X.M. Cui and H.Z. Wen, 2006. Hedonic price analysis of urban housing: The experiential research on the Hangzhou City. Econ. Geogr., 26: 172-176.

Hao, Q.J. and J. Chen, 2007. Distance to CBD, transport accessibility and geospatial differences of Shanghai residential prices. World Econ. Paper, 1: 22-34.

Li, Z.F., J.W. Li and W. Ji, 2011. Discriminant analysis and prediction of house prices based on support vector machines. J. Hubei Normal Univ. (Nat. Sci.), 31: 60-65.

Liu, Y. and H. Zhao, 2013. Study on city residential prices based on enjoy price model. J. Yuncheng Univ., 4: 17-20.

Lu, Q.H., 2012. Neural networks in commercial residential pricing-case of Hangzhou. J. Zhejiang Univ. Technol. (Soc. Sci.), 11: 337-341.

Ma, S. and A. Li, 2003. House PMCE and its determinations in Beijing based on hedonic model. China Civil Eng. J., 36: 59-64.

Ozanne, L. and S. Malpezzi, 1985. The efficacy of hedonic estimation with the annual housing survey. J. Econ. Soc. Measur., 13: 153-172.

Wang, X.Y., 2006. Theoretical analysis of hedonic model of urban housing. Shanghai Manage. Sci., 4: 68-69.

Wen, H.Z. and S.H. Jia, 2006. Market segment and hedonic price analysis of urban housing. J. Zhejiang Univ. Technol., 36: 155-161.

Zhang, M. and S.M. Chen, 2008. Shanghai real estate prices empirical analysis based on hedonic price theory. Fin. Econ., 6: 72-74.