The 16th Pacific Rim Real Estate Society Conference

A Hybrid Model for Long-Tailed Distribution Property

Chung-Hsien Yang[1] Vickey C. C. Lin[2] Ying–Hui Chiang[3] Ming-Fang Yu[4]

Abstract

Since the valuation of typical properties is popular, the long-tailed distribution property is a special issue for the valuation model. Due to the poor accuracy and low hit rate of the estimation in a traditional hedonic model, we developed a hybrid model of OLS and quantile regression for long-tailed property. We also developed a process to build a hybrid model by classifying and identifying properties.

The model has improved the accuracy of Automated Valuation Models (AVMs), especially with long-tailed properties. By using the data of Taipei, the hit rates within 10% and 20% error in OLS are 0.5224 and 0.806, less than 0.5373 and 0.8209 in quantile regression. The MAPE in OLS is 0.1212, more than 0.1197 in quantile regression.

Keywords: Long-Tailed, Hedonic Price, Quantile Regression

[1] Assistant Professor, Department of Real Estate Management, National Pingtung Institute of Commerce, 51 Min-Sheng E. Road, Pingtung, Taiwan 900, E-mail:turtlekk@npic.edu.tw
[2] Professor, Department of Land Economics ,National Chengchi University, Taipei, Taiwan. E-mail:cclinv@nccu.edu.tw
[3] Assistant Professor, Department of Land Economics ,National Chengchi University, Taipei, Taiwan. E-mail:yinghui@nccu.edu.tw
4.Researcher, Department of Land Economics ,National Chengchi University, Taipei, Taiwan. E-mail:mingfang@nccu.edu.tw

Introduction

Real Estate Appraisals have been a professional and individual issue. Because of some reasons, like more efficient, faster, cheaper, even automated, there are mass valuation methods that be developed. In 1970's, Rosen(1974) developed the Hedonic theory and started more studies of the mass valuation. Now, The Hedonic price model is a popular method to valuate properties.

The Basel 2 accord has been adopted since 2008, and the property risk management has been a very important topic. The Basel 2 requires the lenders to monitor the collateral values. In the section 509.3 of the Basel 2 accord, "the bank is expected to monitor the value of the collateral on a frequent basis and at a minimum once every year. More frequent monitoring is suggested where the market is subject to significant changes in conditions. Statistical methods of evaluation (e.g. reference to house price indices, sampling) may be used to update estimates or to identify collateral that may have declined in value and that may need re-appraisal. A qualified professional must evaluate the property when information indicates that the value of the collateral may have declined materially relative to general market prices or when a credit event, such as default, occurs." Thus, we need a stables, accurate, and fast tools in appraisal, and the mass appraisals with Econometric model could be good solutions.

Many researches of Mass Appraisal have focused on Hedonic Price method for past 30 years, but the parametric and OLS regression always has unstable errors, so that there were the low stability and accurate in appraisal. Because the OLS regression is used to obtain estimates for the conditional mean of some variables, the parametrics is the only one number used to summarize the relationship between dependent variable and each of the independent variables. In particular, this method assumed that the conditional distribution is homogenous (Reck, 2004). Therefore, we often see a simple statistics and appraisal tricks, e.g., adjust size $50/sq.ft., or adjust building **depreciation** $10/sq.ft.,(Dell, 2004). This is not fit the actual state. There are different conditional means in different outcomes of some attributes. Chang and Chang(2006) examined the conditional means of different housing prices in the building area and

which were significantly different from zero in Taipei. The Liao and Chang(2006) also examined the effect of real estate brokerage services which were significantly heterogeneous across the conditional price distribution.

For the stability and accuracy, a hybrid method of mass appraisal may be necessary. What is the solution of hybrid method? Allen(2002) thought the hedonic price still was the popular core method[1]. There were some weaknesses of method by single method. The parametric regression always exist a fix conditional mean effect and selection bias of sample distribution (Koenker and Hallock, 2001).

Quantile regression, first introduced in Koenker and Bassett (1978), on the other hand, allows different estimates to be calculated at different points of the conditional distribution. In other words, no assumptions about the homogeneity of the conditional distribution are needed in quantile regression (Reck, 2003). The advantage is that quantile regression fits the robust hypothesis using the empirical quantile. In real estate appraisal, we may need to estimate at every point of conditional distribution, high housing prices and low housing prices both have difference conditional mean. In particular, the tails of distribution seem significantly different from mean or median. We must try to construct a process and a model to take everyone point of conditional distribution, into account including the tails. The quantile regression seems to be a good tool.

We try to construct a hybrid method (OLS and quantile regression) and a process of mass appraisal in this paper. In order to be a core of Automated Valuation Model system (AVMs), it needs a robust, accurate, stable model. We use the transaction data of housing of Taipei and compare the OLS regression with quantile regression.

Methodology

The base of quantile regression had been introduced by Koenker and

[1] The CSW company(www.cswcasa.com) and Solimar company(www.solimar.net) use the hedonic price method as the core of real estate automated valuation model system(AVMs).

Bassett(1978), we have refered Koenker and Bassett(1978) and Kuan(2004) and Chang and Chang(2006) to explain the quantile regression.

Given the data $(y_t, x'_t)'$ for t=1…T, where x_t is k x 1, consider the following linear specification:

$$y_t = x'_t \beta + e_t$$

This specification can approximate a particular conditional quantile of y_t provided that β is estimated properly.

The θ^{th} quantile regression estimator of β can be obtained by minimizing its sample counterpart, i.e., the average of asymmetrically weighted absolute errors with weight θ on positive errors and weight $(\theta - 1)$ on negative errors:

$$V_T(\beta;\theta) = \frac{1}{T}\left[\theta \sum_{t:y_t \geq x'_t\beta} |y_t - x'_t \beta| + (1-\theta) \sum_{t:y_t < x'_t\beta} |y_t - x'_t \beta|\right] \quad (1)$$

For θ =0.5, 2 times (1) is exactly the objective function for LAD estimation:

$$V_T^m(\beta) = 2V_T(\beta;0.5) = \frac{1}{T}\sum_{t=1}^{T} |y_t - x'_t \beta| \quad (2)$$

Hence, a regression estimated via the method of LAD is in effect a special case of conditional quantile regression and is usually referred to as a "median regression."

Let ρ_θ denote the so-called "check" function such that $\rho_\theta(a) = \theta a$ if $a \geq 0$ and $\rho_\theta(a) = (\theta - 1)a$ if $a \leq 0$. We can then write (1) in a compact form:

$$V_T(\beta;\theta) = \frac{1}{T}\sum_{t=1}^{T} \rho_\theta(y_t - x'_t \beta) = \frac{1}{T}\sum_{t=1}^{T} \left(\theta - 1_{[y_t - x'_t\beta < 0]}\right)(y_t - x'_t \beta) \quad (3)$$

Where 1_A is the indicator function of the event A. The first order condition of minimizing (3) is

$$\frac{1}{T}\sum_{t=1}^{T} x_t\left(\theta - 1_{[y_t - x'_t\beta < 0]}\right) = 0 \quad (4)$$

Koenker and Bassett(1978) showed that, when x_t are nonstochastic, together with other regularity conditions, the quantile regression estimator $\hat{\beta}_\theta$ is consistent for β_θ and asymptotically normally distributed when it is suitably normalized. (5) is simplifies to note that the asymptotic variance-covariance matrices in these cases still involve the unconditional density function f_e and hence are not easy to estimate. In practice, f_e is typically estimated by a nonparametric kernel estimator or by bootstrapping.

$$\sqrt{T}\left|\hat{\beta}_\theta - \beta_\theta\right| \overset{A}{\sim} N\left(0, \frac{\theta(1-\theta)}{\left|f_{e(\theta)}(0)^2\right|} IE\left(x_t x'_t\right)^{-1}\right) \tag{5}$$

Because of the features of quantile regression, we can estimate the conditional distribution of real estate price, and we also can test the distributions of tails to get the accuracy estimator.

Data and Model

We use the transaction records of housing of Taipei city in 2007. It's from "The Quarterly Report of Taiwan Real Estate transaction". The resource of the data is the best choice of transaction data in Taiwan. The quarterly report of data is accounted for 10%-12% around whole Taiwan real estate transaction.

In order to test the accuracy, we randomly select 10% of samples[2] to test the model and use the other 90% of samples to construct our model. As to simplify the data, we limit the building type to apartment building. To reduce the influence of limited value, we delete the observations while are greater than 99th quantile or less than 1th quantile of housing price(per unit area). We also delete the building area around 17 pings to 60 pings. After deleting the limited value, there are 3,178 observations in Taipei, 2,860 observations for constructing model and 318 observations for testing data.

[2] The 10% random sample is according 12 county in Taipei, everyone county be selected 10% random.

In the model form, we set a linear-linear regression function form:

$$P = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \varepsilon \quad (6)$$

P: housing price; x_i : housing attributes; β_0 :intercept; ε : error term

The variables setting list of dependent and independent is in following table1:

Table1 The variables list

Variables	Unit	Specification
TOTLPRIC	NT dollars	Housing price, dependent
BUILAREA	Pings	1 ping Equal to 36 ft.sq
BUILARE2	NA	Square of level ground
DHCLS	Years	Building age
DHCLS2	NA	Square of years
TOTFLOOR	Floor	Amount of building floor of overground
FLOOR	Floor	The floor of observation
FLOOR2	NA	Square of Floor
TYPE	Dummy	If the building type is apartment without elevator, TYPE=1, other TYPE=0
ROAD2[3]	Dummy	If the median price of road between 25th to 50th all of median price of the roads, ROAD2=1, other ROAD2=0.
ROAD3	Dummy	If the median price of road between 50th to 75th all of median price of the roads, ROAD3=1, other ROAD3=0.
ROAD4	Dummy	If the median price of road more then 75th all of median price of the roads, ROAD4=1, other ROAD4=0.
CAR	Dummy	If the observation is with parking lot, CAR=1, other CAR=0
D300[4]	Dummy	If the observation away from MRT station is less than 300 meters, D300=1, other D300=0
LANDX	Pings	If the land area are more then 10 pings, we set x=1, other x=0; the LANDX=land area * x
SALEQ1 SALEQ2	Dummy	Seasonal variables, if the transaction day between Jan. to Mar., SALEQ1=1, other SALEQ1=0; if the day between

[3] We hope set a new location variable in the county. We calculate the median of housing unit price in every road and lane, then sort it by road and lane. And we calculate the 25TH , 50TH , 75TH quantile percentage of all "median road and lane price". If someone road or lane is greater than 75th price, we set the road or lane that is the class 4 price section, less 75th price but more then 50th , we set the class 3 section, otherwise, we set the class 2 and class 1 section.

[4] We try to set a space attribute, use the (x,y) coordinate of GIS database from Taipei city government, then got everyone building and MRT station coordinate in Taipei city . We can calculate the distance of everyone building to MRT station, then we try it and found the 300 meters that is best fit setting.

SALEQ3		Apr. to June, SALEQ2=1, other SALEQ2=0; if the day between July to Sep., SALEQ3=1, other SALEQ3=0
Location	Dummy	We set 11 county dummy variables in 12 county of Taipei, the 11 dummy variables are following: L100, L103, L104, L105, L106, L110, L111, L112, L114, L115, and L116; the "wanhua"(L108) county is the base.

Empirical Process

We first ran OLS regression, and then detected[5] the outliers. In order to test the performance of model, we use two different ways to check it, Mean Absolute Percentage Error(MAPE) and Hit Rate. The MAPE checks errors and deviations. And the Hit Rate provides a level of prediction for the model. We set the error range of hit rate to be 10% and 20%. The formulas of MAPE and hit rate are following:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{P_i - \hat{P}_i}{P_i}\right| \times 100\% \quad (8)$$

$$HitRate = \frac{count\left[(P-(1-a\%)P) \le \hat{P} \le (P+(1+a\%)P)\right]}{n} \times 100\% \quad (9)$$

P: the real housing price; $\hat{P}$:the prediction of housing price;

a%: the percentage of error range; n: the numbers of observations

The results, hit rates with 10% and 20% error, and the MAPE are the following table2:

Table2 The parameter estimate of OLS, the result of the hit rates and MAPE

Variable	Parameter
Intercept	5.42571**
BUILAREA	0.04225**
BUILARE2	-0.00017**
DHCLS	-0.0072**
DHCLS2	0.000146**
TOTFLOOR	-0.00415**

[5] In OLS regression, we use the DFFITS method to detect the outlier(Lin, 1997)

FLOOR	-0.01949**
FLOOR2	0.00151**
TYPE	-0.07126**
CAR	-0.02469**
ROAD2	0.14419**
ROAD3	0.22969**
ROAD4	0.31209**
D300	-0.01056
LANDX	-0.00089
SALEQ1	-0.02544**
SALEQ2	-0.00571
SALEQ3	-0.01045
L100	0.50977**
L103	0.04968**
L104	0.35734**
L105	0.5125**
L106	0.67356**
L110	0.43067**
L111	0.32483**
L112	0.18096**
L114	0.2211**
L115	0.21434**
L116	0.09929**
Adj R^2	0.9236**
D-W	2.006
Collinearity Index	12.92
Observations	2,697
Hit Rate with 10% and 20% Error	0.5224 / 0.8060
MAPE	0.1212

*: 5% significance level; **: 1% significance level

There was a problem with quantile regression, we can not get the actually transaction price of property which will be valued. But we need the price to identify the quantile. So we try to build a process for mass appraisal of quantile regression. First, we compare the distribution of some quantile prices by different attributes. Second, we check the attributes patterns of higher and lower quantile. Third, we separate the insample by pattern. If the insample matches the pattern of higher

quantile, we set the outsample to run with higher quantile regression. If the insample matches the pattern of lower quantile, we set the outsample to run with lower quantile regression, the others were set to run with 50^{th} quantile regression.

We set 10^{th}, 30^{th}, 50^{th}, 70^{th}, 90^{th} quantile and compare distributions of the quantile prices by building area, building age and road price class[6]. The distributions with quantile are following in figure1, figure2, and figure3:

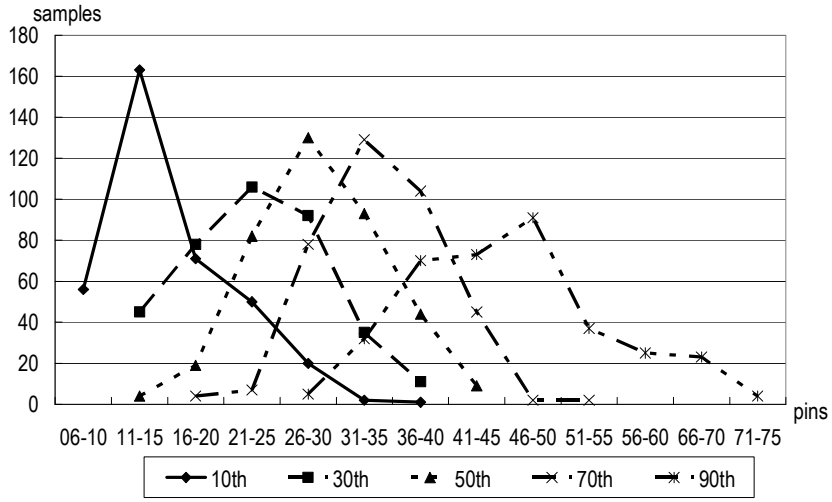

Figure 1 The distribution of building area with quantiles

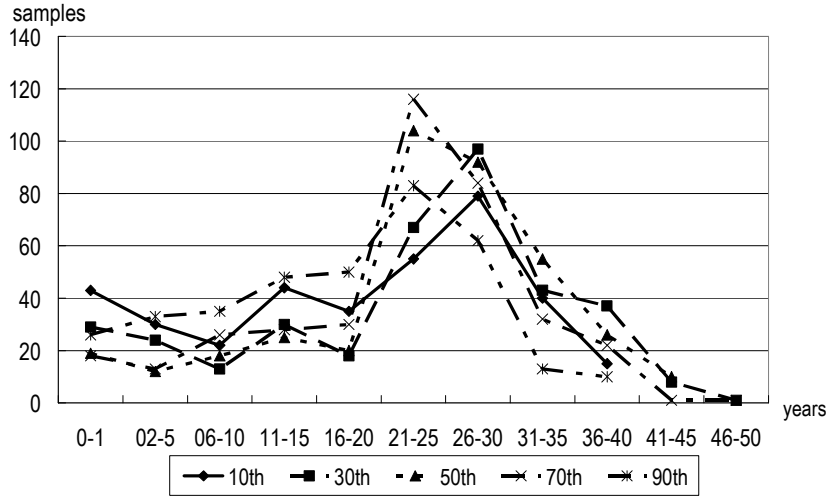

Figure 2 The distribution of building age with quantiles

[6] The standardize estimate of the three attributes are more than 0.1, so we think these attributes are important.

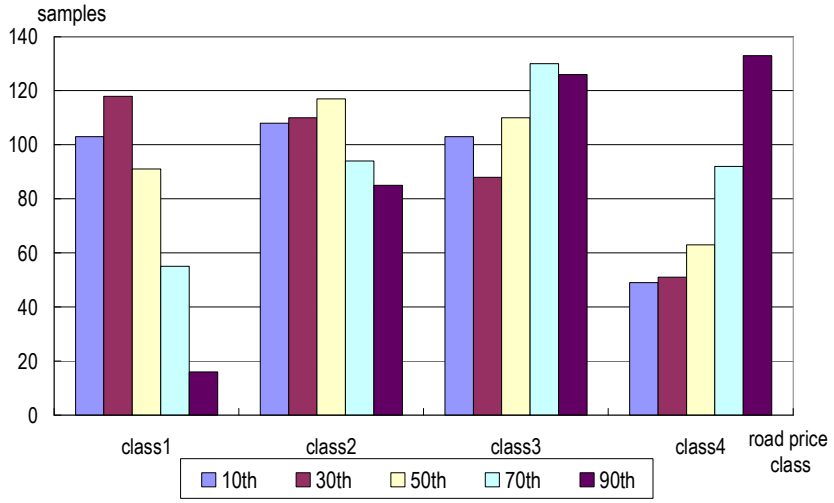

Figure 3 The distribution of road price class with quantiles

Then, we separated the attributes into 5 sections. If the building area(BA), building age(BG), and road price class(RC) match the rules of following table 3, then the outsample would be set to the quantile.

Table 3 The rules which quantile of outsample

Rules	The quantile of outsample
$\mu^{10th}-\sigma^{10th}$<BA<$\mu^{10th}+\sigma^{10th}$, and BG>25years, and RC=class1	10^{th}
$\mu^{30th}-\sigma^{30th}$<BA<$\mu^{30th}+\sigma^{30th}$, and BG>25years, and RC=class2	30^{th}
$\mu^{70th}-\sigma^{70th}$<BA<$\mu^{70th}+\sigma^{70th}$, and BG<=25years, and RC=class3	70^{th}
$\mu^{90th}-\sigma^{90th}$<BA<$\mu^{90th}+\sigma^{90th}$, and BG<=25years, and RC=class4	90^{th}
which others or both setting	50^{th}

We get the results of 5 sections quantile model by insample is following the table4:

Table 4 The parameter estimate of quantile regression with 10^{th}, 30^{th}, 50^{th}, 70^{th}, 90^{th}

Variable	10^{th}	30^{th}	50^{th}	70^{th}	90^{th}
Intercept	5.2462**	5.319**	5.378**	5.4744**	5.7596**
BUILAREA	0.0438**	0.0446**	0.0434**	0.0426**	0.0404**
BUILARE2	-0.0002**	-0.0002**	-0.0002**	-0.0002**	-0.0001**
DHCLS	-0.0081**	-0.0064**	-0.0052**	-0.0072**	-0.0119**
DHCLS2	0.0002**	0.0001**	0.0001**	0.0001**	0.0002**
TOTFLOOR	-0.0005	-0.0014	-0.0033**	-0.0049**	-0.0072**
FLOOR	-0.0276**	-0.0207**	-0.0166**	-0.0124**	-0.0228**

FLOOR2	0.0019**	0.0015**	0.0013**	0.001**	0.0018**
TYPE	-0.0814**	-0.0742**	-0.0582**	-0.0662**	-0.0675**
CAR	-0.0099	-0.0363**	-0.0254**	-0.0247*	-0.0246
ROAD2	0.1455**	0.1477**	0.1403**	0.1205**	0.1087**
ROAD3	0.213**	0.2285**	0.2275**	0.2072**	0.1833**
ROAD4	0.3023**	0.3173**	0.3258**	0.3041**	0.2498**
D300	-0.0121	-0.0047	-0.0038	-0.0062	0.0014
LANDX	-0.0004	-0.0004	-0.0007	-0.0006	0.0003
SALEQ1	-0.0178	-0.0099	-0.0176*	-0.0391**	-0.0728**
SALEQ2	-0.0117	-0.0063	-0.0061	-0.0053	-0.0164
SALEQ3	-0.0087	-0.0044	-0.0075	-0.0093	-0.0193
L100	0.4961**	0.4693**	0.4936**	0.5073**	0.5404**
L103	0.0393	0.0445	0.0475*	0.0613**	0.032
L104	0.3425**	0.3278**	0.3477**	0.3647**	0.3777**
L105	0.4943**	0.4804**	0.5077**	0.527**	0.5175**
L106	0.6486**	0.6459**	0.6684**	0.6833**	0.6791**
L110	0.4461**	0.4104**	0.425**	0.4396**	0.4417**
L111	0.2808**	0.2932**	0.3108**	0.3287**	0.3526**
L112	0.1312**	0.1454**	0.1707**	0.2052**	0.2196**
L114	0.2193**	0.2094**	0.2116**	0.222**	0.2018**
L115	0.2064**	0.2037**	0.2106**	0.2**	0.1924**
L116	0.0947**	0.0824**	0.0843**	0.1095**	0.0774
Observations	2855	2855	2855	2855	2855

*: 5% significance level; **: 1% significance level

Then, we set the outsample to different quantile by table3, and estimate the prices of outsample by quantile parameters of table4. As a result, the table5 shows the MAPE and Hit Rate of OLS regression and quantile regression. We find the MAPE of OLS regression (12.12%) is greater than MAPE of quantile regression (11.97%), and the hit rates (both 10% error is 52.24% and 20% error is 80.6%) are also less than quantile regression (53.73% and 82.09%). From the table5, we could find the quantile regression by attributes is identified to be greater than OLS regression.

Table 5 The MAPE and Hit Rate of OLS and quantile regression

	MAPE	Hit Rate of 10% error	Hit Rate of 20% error
OLS	0.1212	0.5224	0.8060
Quantile	0.1197	0.5373	0.8209

Conclusion

This paper has developed a method to improve the effect of valuation with OLS regression and quantile regression. In our process of quantile regression, we think that the quantile regression is better than OLS regression; it is more robust and stable. It's very useful to build a AVMs of stability and accuracy.

This paper still has some problems to be answered. What is the process of mass appraisal in hybrid method? How to build more rules for more attributes? How to improve the hit rate of OLS regression in small size sample? Is there any other method which can joint to the hybrid method? ex: bootstrapping regression, Grid Adjustment, etc.

Reference

Allen, L. J.

2002 "Automated Underwriting: Collateral Assessment Alternatives", FFIEC Risk Management Planning Seminar 2002.

Calhoun, C. A.

2001 "Property Valuation Methods and Data in the United States," Housing Finance International, 16(2): 12-23.

Chang, E. W. and Chang, C.O.

2006 "The Extension of Mass Housing Appraisal with Hedonic Price Using Quantile Regression," 16th Chinese Society of Housing Studies Conference, Taipei, Taiwan, 16 Dec. 2006. (in Chinese)

Chen, Colin

2005 "An Introduction to Quantile Regression and the QUANTREG Procedure," SAS Institute Inc.

Clapp, J. M.

2003 "A Semiparametric Method for Valuing Residential Locations: Application to Automated Valuation," *Journal of Real Estate Finance and Economics*, 27(3): 303-320.

Dell, G.

2004 "AVMs: The Myth and the Reality; the Problem and the Solution," *Valuation Insights and Perspectives*, 9(3): 12-52.

Fisher, J. D.
2002 "Real Time Valuation," *Journal of Property Investment & Finance*, 20(3): 213-222.

Koenker, R. and Bassett, G..
1978 "Regression Quantiles", Econometrica, 46(1): 33-50.

Kuan, C. M.
2004 "An Introduction to Quantile Regression," Institute of Economics Academia Sinica.

Kummerow, M. and Hanga Galfalvy
2002 "Error Tradeoffs in Regression Appraisal Methods," *In Real Estate Valuation Theory*, edited by Ko Wang and Marvin Wolverton. Norwell, Mass.: *American Real Estate Society*, Kluwer Academic Publishers, 2002.

Kummerow, M
2002 "A Statistical Difinition of Value," *The Appraisal Journal*, 70(4): 407-416.

Lai, Pi-Ying
2006 "Analysis of the Mass Appraisal Model Using Artificial Neural Network in Kaohsiung City," *2006 Pan Pacific Congress(PPC)*, San Francisco, US, 16~19 Sep. 2006.

Liao, C. J. and Chang, C. O.
2006 "Asymmetric Price Effects of Residential Real Estate Brokerage Service Using Quantile Regressions," Journal of city and Planning, 33(1): 1-16, Taiwan. (in Chinese)

Lin, Vickey
1997 "Fluctuation of Housing Prices: 1971-1995---Theory and Application," Book published by Chai –Yan, Taiwan.(in Chinese)

Meese, R. & Nancy Wallace
1991 "Nonparametric Estimation of Dynamic Hedonic Price Models and the Construction of Residential Housing Price Indices," *AREUEA Journal*, 19(3): 308-332.

Pace, R.K.
1993 "nonparametric Methods with Applications to Hedonic Models," *Journal of Real Estate Finance and Economics*, 7: 185-204.

1995 "Parametric, semiparametric, and nonparametric estimation of characteristics values within mass assessment and Hedonic pricing models," *Journal of Real Estate Finance and Economics*, 11: 195-217.

Peng, C. W. and Yang, C. H.
2007 "Potential Impacts of AVMs on Real Estate Appraisers," Journal of Housing Study, 16(1), not yet published. (in Chinese)

Reck, C. G.
2003 "Heterogeneity and Black-White Labor Market Differences: Quantile Regression

with Censored Data 1979-2001," University of Illinois at Urbana-Champaign, 2003.

2004 "Three Essays Exploring Heterogeneity Using Quantile Regression," Ph.D., University of Illinois at Urbana-Champaign, 2004.

Rosen, Sherwin

1974 "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *The Journal of Political Economy*, 82(1): 33-55.

Shiller, R. J. & A. N. Weiss

1999 "Evaluating Real Estate Valuation Systems," *Journal of Real Estate Finance and Economics*, 18(2): 147-161.

The Appraisal Foundation

2006 "2006 Uniform Standards of Professional Appraisal Practice."

Waller, B. D.

1999 "The Impact of AVMs on the Appraisal Industry," *Appraisal Journal*, 67(3): 287-293.

Koenker, R. and Hallock, K. F.

2001 "Quantile Regression," Journal of Economic Perspectives, 15(4): 143-156.