ESTIMATION OF REAL ESTATE ASSET PRICING MODELS

Felipe Alonso Arias Arbeláez[†]
Francisco Iván Zuluaga Díaz[‡]

EAFIT University

2016

Abstract. In this project we aim to develop 4 different methods in order to estimate the real market price of 380 properties owned by Midtown Realty Group in Miami, Florida. We used the ordinary least squares, generalized method of moments, artificial neural networks and fuzzy inference systems. The comparison between the 4 models was made using the root mean squared errors (RMSE) with an interesting result showing that the best method to estimate housing price given our data set is the artificial neural network using the correct network architecture. Some further work is proposed in order to make more comparison between the models and define the best model for housing price.

Keywords: Neural Networks, Fuzzy Inference System, Hedonic Price, Housing Price, Ordinary Least Squares, Generalized Method of Moments

[†]Department of Economics, EAFIT University. email: fariasa@eafit.edu.co
[‡]Department of Mathematical Sciences, EAFIT University. email: fzuluag2@eafit.edu.co

Contents

Introduction

When it comes to estimating assets it is important to emphasize the great contribution econometric models have had to explain this problem. As for the real estate sector is no exception, many econometric models have explained the basis of valuation of real estate assets of all kinds: commercial, residential, etc. The problem arises when these valuation models and estimations are not compared with other models. No one could ever say that a model is good if it is not compared to other options. In literature are some solutions to the valuation of real estate assets with multiple regression models and atificial intelligence models (Borst, 1991).

It is very important for a real estate company to know the value of their assets based on relevant known information, there is where the problem of valuing these assets in the best way arises; some researchers like Rossini (1997), Lai (2006) and Can (1992) propose methods of simple linear regressions and other authors like Kusan, Aytekin, and Ozdemir (2010) propose boldest estimates based on artificial intelligence methods such as neural networks and fuzzy inference systems.

I am currently working on a portfolio management company in real estate based in South Florida and we have almost 500 properties of which have organized database for exercise. Mainly it seeks to estimate the price of the properties, but the academic background is to compare estimations using econometric models and using artificial intelligence models to define what is best in terms of minimizing the error and curve fitting. It's important to realize that our study is based on cross-sectional data so there are not a lot of models to do like in times series data, that's the reason we only propose 2 models for the econometric estimation that can be sucesfully compared and 2 models based on artificial neural networks and fuzzy inference systems.

The general objective we are seeking in this paper is to assess real estate assets of Midtown Realty Group LLC using econometric methods and artificial intelligence methods; in order to get that result we need to perform some specific tasks such as estimate the price of real estate assets using the least squares model log-lin using ordinary least squares, generalized method of moments, neural networks and fuzzy logic systems; and to identify the difference and effectiveness of models in estimating real estate assets. All of these seeking a global result of contribute to the development of estimation models for the real estate sector in order to avoid a crisis like the one presented in 2008 in part by the mispricing of real estate assets.

Since the 2008 crisis that was caused, mainly, by problems with real estate assets, it has been implemented robust estimates such as the ones by (Evans, James, & Colins, 1991) and (Tay & Ho, 1991) regarding the valuation of real estate assets. Recall that in the 2008 crisis housing bubble mainly caused by excessive price increase in housing were reported.

This leads us to justify that it is important for the market to have more scientific basis for assessing assets of this nature, can not just leave to others the assessment because it could happen similar crisis. In this way it seeks to Midtown Realty Group LLC, as market maker may have for buying and selling prices chords to the market and not according to bubbles that can be generated. This would lead to greater investor confidence to see that prices for buying and selling properties within the portfolio is not being subject of subjective opinion but objective studies and methods of robust estimation.

Finally the theoretical justification is that so far no one has developed its own neural network, and already well said (Kosko, 1992) in his book, the estimation methods for artificial intelligence should be adjusted for each problem, no networks or fuzzy inference systems are general for all problems.

Literature Review

Rossini (1997) in their study compares the dynamics of neural networks with linear regression. They use Australian data taken from Upmarket that are of public domain and estimate the price based on three models. The variables used to estimate their models are the date of sale of the property, the sale price, the suburb where the property is located, arrangements made to the property, total habitable area, the area (based on what the government defines for Australia), the number of rooms, the total area of the building where the property is located, the condition of the unit, the type of walls, roof type, the building style and the year in which the building was constructed. Which is interesting about Rossini (1997) is that they define many variables to try to explain how the price of a real estate asset is constructed. The results obtained by Rossini (1997) are very interesting considering the amount of variables (for the case of OLS) or entries (for the case of the neural network). In total 10 variables were used to explain the price of the assets in the linear regression; he obtained a joint significance of 98%. The neural network was defined by 10 input neurons, 10 hidden layers and one output. They conclude that it is better to use models of linear regression than neural networks; also they conclude that the key is the training of the neural network and the performance that it can achieve, they conclude that the ordinary least squares models are better because they reduce the error more than neural networks does.

In China, Lai (2006) used linear regression models and also compared them with neural networks methods in a city called Kaohsiung as study, as regressors in the case of the econometric model and inputs to the neural network he used the location, type of street that had in front of the property, property type, structure, the date on which was built the property, total area, the (commercial or residential) area and the area where the building was located.

Lai (2006) got his regressions with 5% significance in all regressor variables, and R^2 of 69.4 %, indicating that the model had a great explanation based on the regressors data. An interesting topic was the analysis made on the basis of economic variables, for example the age of the building should have an inverse relationship with the price of the property,

according to economic theory; a direct relation was obtained, implying that not necessarily an old building will be cheaper than a new one; there are more important variables that can determine the price, and age of the building is not one of them.

As for the neural network,Lai (2006) did not build their own neural network but used the *Alyuda* software. The results were better than those of the linear regression against those obtained by Rossini (1997).

In other studies, McManus and Mumey (2002) created a patent where estimated using a model of *Splines* the price of real estate assets. Although this method will not be used in our study, it is important to highlight the work of the authors, which use a method of estimating nonparametric as is the method of splines, use market data and create an application that estimates the price of real estate assets.

Hedonic Models started to arise in the 90's with Can (1992) studies, he stated that there are a lot some useful methodologies for cross-section data. The most important theory of hedonic models is based on the next equation (Equation taken from Can (1992)):

$$P(H) = f(h_1, h_2, ..., h_k)$$

Such a simple equation that says a lot about the estimation. As Can (1992) said: "(this model) establishes a functional relationship between the observed household expenditures on housing, P(H), and the level of characteristics contained in vector H".

The final conclusion of Can (1992) is based on the neighborhood variable, he used some spatial econometrics to sucesfully explain the spatial problem in the regression; however he realized some other linear regressions with simple estimators excluding the number of bedrooms in the unit, that's the most important difference between his model and the ones of Rossini (1997) and Lai (2006).

One of the most interest thing in econometric evaluations are the assumptions, for example in the variance we need to have homoskedasticity, which is a very strong assumption. In all the studies mentioned here, they had problems with heterskedasticity, so Goodman and Thibodeau (1995) studied the causes of the violation of this assumptions and they found the age of the property will cause heteroskedasticity because the depreciation of real estate assets is nonlinear in the reality, so this variable need to be treated different, that's the reason to involve log-lin models in the especification of hedonic models. This problem was also studied by (Fletcher, Gallimore, & Mangan, 2000) as an extention of Goodman and Thibodeau (1995) study; they said that not only the model present heteroscedasticity because of the age but also the external area of the property can generate problems with the variance. As a final conclusion, Fletcher et al. (2000) claimed that EGLS models can correct this problem on OLS models, making the estimators consistent and asymptotically efficient.

Some interesting topics of analysis in hedonic price house is based in taking into account some variables based on the spatial distribution like distance between a house and the nearest metro station, nearest to a Walmart, etc. Liao and Wang (2012) realized that those variables were important and created a fictional data set to try the spatial regressions. They assumed some distances to be important in the model and made some comparison between OLS, GMM and spatial quantile regression. They concluded also that regressions made with GMM, instrumental variable or least squares in two stages are asymptotically equivalent.

For the FIS models we found Kusan et al. (2010) worked with them in order to achieve the appropiate housing price. They use 4 inputs for their model such as city plans, location based on the nearness to some medical, cultural and educational buildings in ther home town; then they compute the error of the FIS model using the root mean squared error shown in Equation 1.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(T_i - P_i)^2} \tag{1}$$

Using the difference between the estimated data and the real data. This Equation is going to be used in our discussion to compare the errors.

Methodology

The following are variables that are taken from the database of properties in Midtown Realty Group, a company based in USA that have 500 properties but for this exercise we chose 380 because some of the 500 are in another city and have different sspecifications; some data like the number of bedrooms and bathrooms was taken from Miami Dade web page were by law are the public description of each property. In parenthesis is we define the name of each variable to be used in the model.

- **Property Name (ID):** because Midtown Realty Group want some confidenciality with the data we changed the name of the property with a number between 1 and 380.
- **Number of bathrooms (Bath):** variable taken from Miami Dade public record and is between 1 and 5.
- **Number of bedrooms (Bed):** variable taken from Miami Dade public record and is between 1 and 5.
- **Size of property in square foot (SqFt):** variable taken from Miami Dade public record, this is the total area of the unit measured in square feet.
- **Property location (City):** dummy variable that is 1 when the property is based in the center area (Brickell, Midtown, Downtown, etc), and 0 when it's outside the center of the city.
- **Property age (Age):** this is just the difference between today's date and the year the property was built; this is going to be expressed in years.

- **Square of property age (Age2):** Just the square of the property age.
- **Type of property (Type):** Midtown Realty Group have a lot a different properties but for this analysis we are interested in just analyze the residential segment, so we dropped the commercial properties. Type of property is a dummy variable that is 1 when the property is an apartment and 0 when is a house.
- **Market price of the property (Price):** This is the assessed price in 2016; information taken from Miami Dade public records.
- **Actual rent of the property (Rent):** All the properties Midtown Realty Group have are rented, so we take into account the actual amount paid for the monthly rent.

For ease we are going to divide by 1000 the variables rent and Market Price.

Econometric Methods

The econometric estimations are based on the next equation:

$$ln(Price) = \Sigma * \Psi + \varepsilon \tag{2}$$

where,

$$\Sigma = [\beta_0 \quad \beta_1 \quad \cdots \quad \beta_8], \qquad \Psi = \begin{bmatrix} 1 \\ Bath \\ Bed \\ SqFt \\ City \\ Age \\ Age^2 \\ Type \\ Rent \end{bmatrix}$$

Because we have cross sectional data the model is quite simple and there is not a lot of models we can define; since our aim is to obtain the minimun error in an estimation we define two econometric models that are equivalent but differs in how they calculate errors. In Equation 2 we define the price of each property as a linear relation. This price is going to be dependant of the number of bathrooms, number of bedrooms, total area measured in square feet, location, age, type (house or apartment) and rent.

It is important to note that we are going to work with a log-lin model which have the dependant variable expressed in natural logaritm. This is because it's easier to analyze the relation between the price and the independant variables; with the natural logaritm we can do the analysis with percentage instead of real values (Chumncy & Simpson, 2005).

Ordinary Least Squares (OLS)

There is no better summary of the OLS method that the one given by (Chumncy & Simpson, 2005): "Ordinary Least Squares, referred as OLS, is one of the most common techniques used in multivariate analysis. Regrettably, it is also probably the technique most misused".

It is important to note that Equation 2 defined in the Econometric Methods section is a linear equation with an intercept and also an ε; the importance of this method is based on the capability of the model to decrease the sum of the square of those ε called estimation errors. In Equation 3 we defined the basic optimization in order to compute OLS.

$$\min \sum_{i=1}^{n} \varepsilon_i^2 \tag{3}$$

As this method is a linear estimation of the relation between the dependant variable and the independant variables, to minimize the square of the errors we are reducing the distance between the estimated points and the real points. There are some steps we need to do in order to implement the method (Wooldridge, 2009):

1. Compute the distance between the estimated points and the real points in the data set.

2. Take the square of the residual (estimation errors) in order to make the distances all positive.

3. Estimate the line that minimizes in all way the sum of the squared errors.

Generalized Method of Moments (GMM)

Since the OLS is a very specific case of the GMM method, we introduce the GMM method to estimate Equation 2 in order to avoid the endogenous problem that may be presented in the regressors, at this point we can minimize even more the error having the same estimation for the regressors. So the methodology of this estimation is quite simple, we minimize the sum squared errors as well but taking into account some instruments to avoid the endogenous problem, so basically we supose that the error and some regressors do not have an orthogonal relationship, giving us some endogenous problems (Gujarati, 2010).

Since the calculation steps are the same as OLS we follow the same rules but adding the endogenous problem between the regressors and the residuals to calculate the estimators with reduced variance.

Artificial Intelligence Methods

The aim of the Artificial Intelligence methods is based on the cognitive learning the humans have and the power of the machines that can achieve that kind of learning. Basically what people want is to enable a computer to do some cognitive processes than only can be made in a human brain. So AI is a representation of what a brain can do on how we compute information in our heads. There are a lot of methods that try to do this congnitive aproach, but in this paper we are going to focus on two well-known methods: Neural Networs and Fuzzy Inference Systems.

Artificial Neural Networks (ANN)

The inspiration of neural networks is the study of the central nervous system, (Haykin, 1999) said that neural networks "has the capability to organize its structural constituents, known as *neurons*, so as to perform certain computations many times faster than the fastest digital computer in existence today."

A neural network is defined by layers where there are neurons, as it would function in a human brain. Basically you have 3 layers, one input, one hidden and one output. The relationships between neurons in each layer are called weights. It is noteworthy that the number of neurons and layers do not follow a logical behavior rules, it simply depends entirely on the researcher (Haykin, 1999).

Neural networks in its basic operation have 3 types of parameters:

1. The pattern of connections between different layers of neurons.
2. The learning process that updates the weights of each connection.
3. The activation function that converts input neurons to outputs.

The general architecture of a multilayer perceptron-network can be seen in Figure 1.

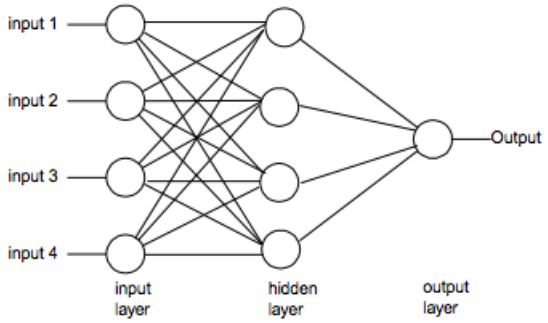

Figure 1: Architecture of a network. Taken from Rossini (1997)

Borst (1991) explains that "there is an input layer, a hidden layer, and an output layer. In mass appraisal, the input neurodes represent the input data in much the same fashion as the X, (independent variables) in the linear model. The output layer represents the output sought by the model of the process of interest. In mass appraisal, one output neuron would be used to represent estimated selling price or perhaps estimated rent. The hidden layer allows for the combination of input data in a near infinite number of ways".

Network training process is by which the weights of the input variables of the neural network for minimizing variation in the curve fit to the actual data, or which is similar fit, minimize error prediction to compare methods.

Now the question is how to model the neural network and how do training for the network to "learn" and can estimate with minimal error. The algorithm *back propagation* is a supervised learning algorithm that solves the problem by using the method of descent gradient with momentum Haykin (1999). Basically what the propagation algorithm do in order to make the network learn is to take a database in and find patterns that have each minimization of the cost function is efficient, making the network each who can start, at least, where to start looking for the weights of the variables. At the end of the day a network would be able to take specific property data and based on what it already learned, estimate selling prices with minimal error (Kosko, 1992).

In this simulations is neccesary to define the variables that are going to be used:

- **Inputs:** independant variables.
- **Outputs:** dependant variable.
- **Hidden layers:** We need to define tue number of hidden layers and the kind of activation function; the most used is the hyperbolic tangent.
- **Output layers:** also it is important to define the number of hidden layers and the activation function; the most used is the sigmoid.
- **Max ephocs:** the number of simulations that we are going to do.
- **Bias:** in this methods it is important to define a coefficient associated to the bias of the network.
- **Momentum:** This variable allows the back propagation algorithm to avoid local minima.
- **Learning rate:** in the first steps of learning we need to define a learning rate for the neural network, this rate compares each estimation in each epoch to define if the neural network is improving or getting worst.

Fuzzy Inference Systems (FIS)

The latest model is using fuzzy logic described by Zadeh (1965), professor Lotfi exposes us generally the problem of classical logic. It has argued that it can not be white or black, we have in the logic gray parts that do not correspond to an exact answer; in other words, professor Lotfi assures us that logic can not be bivariate (with answers 1 or 0), but can give a range of possible responses between 0 and 1. Based on this, Zadeh (1965) states that the values or weights each variable can take is described by rules that follow the study variables; in this same vein, it is possible to estimate using fuzzy logic rules for each of the variables and letting it fluctuate in a world of possibilities. The generation of several fuzzy sets for the variables and their relationships with each other create the answers to the study variables.

Perhaps the greatest contribution of Zadeh (1965) are the membership functions, which say each variable has a function that links a possibility in the set of 0 and 1, thus leaving to be false and true, but having several options in an infinite set of numbers.

Fuzzy inference systems (FIS) use the principle of fuzzy logic raised by Professor Lotfi. In this project we will use a system of fuzzy inference Mamdani-Sugeno type, which has a *fuzzyfication* stage, which converts an input variable in a fuzzy set using membership functions; a set of rules if-then, which will direct the membership functions to link inputs with outputs; and the last component is the *defuzzyfication*, which transforms the output fuzzy set of numbers.

Data

Before we make the estimation of the models it's important to analyze the data and define what we expect in the simulations, as we said we have a data set of 380 properties all in Florida in Miami-Dade and Broward County. In Table 1 we can see the descriptive analysis for the main variables involved in the estimation of each model.

Variable	**Mean**	**Std. Dev.**	**Min**	**Max**
Bedroom	1.94	0.91	1	4
Bathroom	1.77	0.77	1	4
SqFt	1100.245	495.42	425	4433
Age	24.49	12.13	0	60
Type	0.55	0.50	0	1
City	0.63	0.48	0	1
Rent	1.47	0.75	0	9
Market Price	161.28	163.91	55	1364.26

Table 1: Descriptive statistics of the variables.

For the bedroom variable we can see that the mean of the total observations is near two, so most of the properties have 2 bedrooms in an range from 1 to 4 possible bedrooms; for the bathrooms we have almost the same statistics as the bedrooms, almost 2 in the mean and the same range from 1 to 4. The SqFt variable which involves the total area of the property we have that the mean of the sample is 1100 square feet and we have properties from 425 to 4433 square feet.

The age is such an important variable and since we are working with the assumption of perfect markets we do not care much about the depreciation of the assets in books, so we only take into account a simple relation between age and price: if a property is older then the price is high, this is not always true but works logically 90% of the times (Goodman & Thibodeau, 1995). Table 1 shows that the mean is 25 years old for the properties, having new properties (0 age) and the oldest that are from 1956.

The type of the property is well distributed between houses and apartments because the mean is nearly 0.5, so we expect to have this variable as significant in the econometric models; this can explain a lot the relation between the price and variables such as the type of the property. Similarly we have the variable city that shows a huge number of properties are in the center of the city, so this variable is expected to be not significant in the model.

The rent and the market prices are expressed in thousands of dollars, so we have that the mean price for rent is $1.470 with rents from $0 (new properties) and $9.000 in a monthly basis. With the market price the average property is $161.280 having properties from $55.000 up to $1'364.260.

The previous analysis can be shown more easily in Figure 2

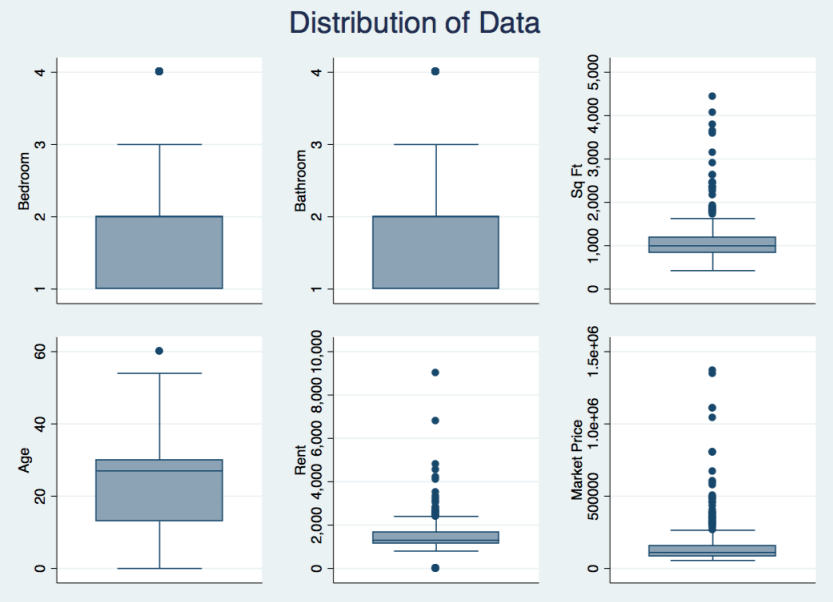

Figure 2: Distribution of data with box plots using Stata®.

Box plots shown in Figure 2 show that the age of the properties, the number of bedrooms and bathrooms have a normal distribution that oscilates between two values consistently, but for the total area, rent paid and market price we can see that we do not have a consistent range, so we have a lot of outliers that can affect the estimations but at the same time we can not elimiate them because they have relevant information. These outliers need to be treated in a special way to avoid erroneous estimates.

A priori we can say that the price of an unit is directly related to the number of bathrooms and bedrooms the most, so in Table 2 we can see the number of properties with each amount of bedrooms and bathrooms.

Baths \ **Beds**	1	2	3	4	Total
1	128	15	0	0	143
1.5	5	5	1	0	11
2	6	124	47	6	183
2.5	0	8	4	1	13
3	0	3	3	1	7
4	0	0	0	23	23
Total	139	155	55	31	380

Table 2: Relation between bathrooms and bedrooms in all the units

As shown in Table 2 we can see that the common distribution is to have 1 bathroom/1 bedroom and 2 bathroom/2 bedroom; the logic is that an unit with one bedroom is not going to have 4 bathrooms, as the number of bedrooms increase it is expected to have more bathrooms. So in order we have more 1/1, then 2/2 followed by 3/2, being the first number bedrooms and the second number bathrooms.

Finally it is important to analyze the cross correlation between the variables involved in the estimations, in Table 3 we can see the cross correlations.

	Price	Beds	Baths	SqFt	Age	Rent	Type	City
Price	1.0000							
Beds	0.4598	1.0000						
Baths	0.5093	0.8383	1.0000					
SqFt	0.6345	0.6883	0.6683	1.0000				
Age	-0.1665	-0.2059	-0.1804	-0.2576	1.0000			
Rent	0.4434	0.3888	0.3670	0.6491	-0.2609	1.0000		
Type	-0.0979	-0.1906	-0.1166	-0.2705	0.2396	-0.1721	1.0000	
City	-0.0202	-0.0544	-0.0902	-0.0851	-0.1133	0.0663	0.0000	1.0000

Table 3: Cross correlations between variables

Based on Table 3 we can see that we do not have autocorrelation problems in the variables involved in the regressions and estimations. We expected to have a high correlation between bathrooms and bedrooms because those are possibly the most important variables involved in the price of each property; also the area measured in square feet have high correlation with the price. Since Type and City are negative and so small compared to the other variables we expect these variables to be not as important as the other ones, maybe giving us coefficients that are not statistically significant.

Estimations

Fist we simulated the econometric models and then the artificial intelligence ones.

Estimations by OLS

In Table 4 we can see the estimation for Equation 2, the significance of each regressor and the standard error. We developed 3 variations of the main model exposed in Equation 2, the first one is estimating the model with all the expressed variables, the second one in estimating the model with only the number of bathrooms, the number of bedrooms, the total area variable (SqFt) and the rent; and the last model using only number of bathrooms, number of bedrooms and the total rent paid by the property. This models were defined by us based on the studies made by Borst (1991) and because of the common believe of being those variables the most important in order to define the market price of each property.

Variable	**(1)**	**(2)**	**(3)**
β_0	4.4495*	3.8499*	3.8769*
	(0.1602)	(0.0638)	(0.0602)
$Bath$	0.1545*	0.1222†	0.1828*
	(0.0540)	(0.0562)	(0.0702)
Bed	0.0590‡	0.0492	0.1140*
	(0.0319)	(0.0342)	(0.0315)
$SqFt$	0.0002†	0.0004*	-
	(0.0001)	(0.0001)	
Age	-0.0513*	-	-
	(0.0087)		
Age^2	0.0008*	-	-
	(0.0001)		
$Rent$	0.2100*	0.1863†	0.2970*
	(0.0785)	(0.0905)	(0.0705)
$Type$	0.0952‡	-	-
	(0.0509)		
$City$	-0.0006	-	-
	(0.0457)		
R^2	0.55	0.48	0.44
RMSE	0.3896	0.4173	0.4313

Significance Levels: * 1%, † 5%, ‡ 10%.
Estimation using robust regression for heteroskedasticity.

Table 4: Estimations using OLS method

As we can see in Table 4, from 8 variables we got 4 significant at 1% level, 1 at 5% and 2 at 10% for the first model. Only the city (location of the property) was not significant.

Taking into account the results shown in Table 4, we can put Equation 2 as is shown in Equation 4. The R^2 nearly the 50% in the three models which means the data explains the model in about 50%, this is a good number taking into account that we are working with cross-sectional data. Also we have the MSE which is the variance of the residual; this is going to be our error comparative in the next section to define the best estimation method for this problem.

$$lnPrice_i = \Sigma_i * \Psi_i \tag{4}$$

where

$i = 1, 2, 3;$ each i is the defined variation of the general model

$$\Sigma_1' = \begin{bmatrix} 4.4495 \\ 0.1545 \\ 0.0590 \\ 0.0002 \\ -0.0513 \\ 0.0008 \\ 0.2100 \\ 0.0952 \end{bmatrix}, \qquad \Psi_1 = \begin{bmatrix} 1 \\ Bath \\ Bed \\ SqFt \\ Age \\ Age^2 \\ Rent \\ Type \end{bmatrix}$$

$$\Sigma_2' = \begin{bmatrix} 3.8499 \\ 0.1222 \\ 0.0492 \\ 0.0004 \\ 0.1863 \end{bmatrix}, \qquad \Psi_2 = \begin{bmatrix} 1 \\ Bath \\ Bed \\ SqFt \\ Rent \end{bmatrix}$$

$$\Sigma_3' = \begin{bmatrix} 3.8769 \\ 0.1828 \\ 0.1140 \\ 0.2970 \end{bmatrix}, \qquad \Psi_3 = \begin{bmatrix} 1 \\ Bath \\ Bed \\ Rent \end{bmatrix}$$

We defined the final equations in terms of matrix because of the amount of variables involved in the estimation, we show also the matrix Σ_i transposed for ease. We eliminate City or Location because it was not statistically significant in each model.

It is important to analyze the results, because we are working with log-lin models, each coefficient tell us the increment or decrease in the dependant variable in terms of percentage. In Table 5 we can see the percentage each independant variable affect the final price of each property in each model.

Variable	**(1)**	**(2)**	**(2)**
Bath	15.45%	12.22%	18.28%
Bed	5.9%	4.92%	11.40%
SqFt	0.02%	0.02%	-
Age	-5.13%	-	-
Rent	21%	18.63%	29.70%
Type	9.51%	-	-

Table 5: Percentage each independant variable affect the price

The most important variables in our data set that define the price of each property are number of bathrooms, rent paid and number of bedrooms. The sense of this is based on the preference for each person to require a big place to live so supply and demand make the market price go up. Also the rent show us that if we increase the rent in 100 dollars, the price of the property is going to increase in 21% for the first model, 18.63% for the second one and the almost 30% in the last one; telling us that the market price is defined the most for data of the rent paid in each property.

In Miami-Dade and Broward county the appraisal of the properties are based on the listings in MLS[1], so the offer made by the person interested in the property is based on the number of bedrooms, bathrooms and rent that is being paid in that moment. Also the type of the property was significant in the model, it tell us for the first model that if the property is an apartment, the price of the property is going to increase in about 9.51%; we exclude this variable in the next estimations because this is highly corralated with bedrooms and bathrooms as shown in Table 3, so it can gave us problems of autocorrelations because of the few variables chose for the estimations (Lai, 2006).

All tests are shown in Appendix A, we found an initial problem with heteroskedasticity which was corrected with a robust regression. All the assumptions made for the OLS estimator are fully obtained.

[1]Software that allows realtors to show properties for rent or to buy them.

Estimations by GMM

In Table 6 we can see the estimation for Equation 2, the significance of each regressor and the standard error, also using variations of the main model as in the OLS estimations.

Variable	**(1)**	**(2)**	**(3)**
β_0	4.4488*	3.8499*	3.8768*
	(0.1398)	(0.0634)	(0.0599)
$Bath$	0.1545*	0.1222†	0.1828*
	(0.0529)	(0.05584)	(0.0698)
Bed	0.05898‡	0.0492	0.1140*
	(0.0311)	(0.03396)	(0.0313)
$SqFt$	0.0002†	0.0004*	-
	(0.0001)	(0.0001)	
Age	-0.0512*	-	-
	(0.0082)		
Age^2	0.0009*	-	-
	(0.0001)		
$Rent$	0.2099*	0.1863†	0.2970*
	(0.0765)	(0.0899)	(0.0701)
$Type$	0.0951‡	-	-
	(0.0499)		
RMSE	0.3653	0.3287	0.4150

Significance Levels: * 1%, † 5%, ‡ 10%.

Estimation using robust standard errors for heteroskedasticity.

Table 6: Estimations using GMM method

With this method we expected to have the same results as in the OLS method because OLS is a special case of GMM; the aim of estimate with this model is to use the robust errors estimation in order to avoid heteroskedasticity and to minimize even more the error, in next section we will compare the two models in terms of error, in this case all the GMM calculated errors are smaller than OLS ones.

It is important to note also that the coefficientes estimated are similar to the ones estimated by the OLS method but not equal; this is because of the moments used in each method; since in OLS they are predeterminated, in GMM we define all the regressors as instruments for the number of moments.

In Table 7 we can see the results in terms of percentage, same as the ones shown in the OLS estimations.

Variable	**(1)**	**(2)**	**(2)**
Bath	15.45%	12.22%	18.28%
Bed	5.89%	4.92%	11.40%
SqFt	0.02%	0.04%	-
Age	-5.12%	-	-
Rent	20.99%	18.63%	29.70%
Type	9.51%	-	-

Table 7: Percentage each independant variable affect the price

Estimations by ANN

The most important variables involved in the artificial neural networks are the number of hidden layers and the neurons per layer; in the literature we found that more layers and neurons do not give better results (Tay & Ho, 1991); the problem is to know the exact amount where the neural network is failing. Table 8 show the parameters used to estimate the market price of the properties using artificial neural networks. This is also known as the *architecture* of the neural network.

Inputs	8
Outputs	1
Hidden Layers	Hyperbolic Tangent
Output Layer	Sigmoid
Max. Ephocs	2500
Bias	0.1
Momentum	0.2
Hidden Layers	1
Hidden Neurons per Layer	4
Learning Rate	0.5

Table 8: Architecture and training parameters

After some experiments changing the most important values, we have that the minimum error is given by the configuration shown in Table 8. We can see in Table 9 the results of 9 networks with different number of neurons per hidden layer. The best one (measured in minimum error) is the network 13.

Network	Layers	Neurons	Learning Rate	Momentum	RMSE
11	1	2	0.5	0.2	0.2225
12	1	3	0.5	0.2	0.2242
13	1	4	0.5	0.2	0.2130
14	1	5	0.5	0.2	7.3331
15	1	6	0.5	0.2	6.5983
16	1	7	0.5	0.2	28.9812
17	1	8	0.5	0.2	15.2134
18	1	9	0.5	0.2	45.2325
19	1	10	0.5	0.2	50.9012

Table 9: Estimation of some networks changing number of neurons per hidden layer.

As shown in Table 9, when the number of neurons increase, the error increase. With 4 neurons in 1 hidden layer we obtain good results. In Figure 3 we can see the real output with the estimated output with the neural network using the *architecture* shown in Table 8. Also we include the minimization of the error among the 2500 ephocs in Figure 4

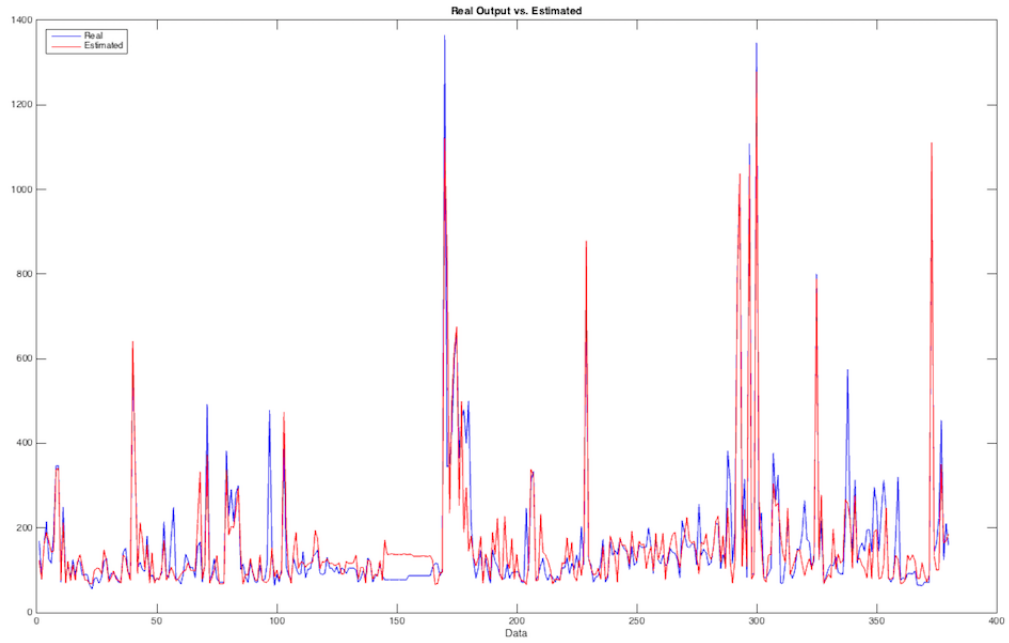

Figure 3: Real output and estimated output for properties market price using ANN in Matlab®.

It is interesting to note that if we define more ephocs, the error is going to be minimum, but we can not define more than 5000 ephocs because the algorithm is not going to be as fast as we need it to be; it is not efficient when you have a method running for long time to estimate a variable.

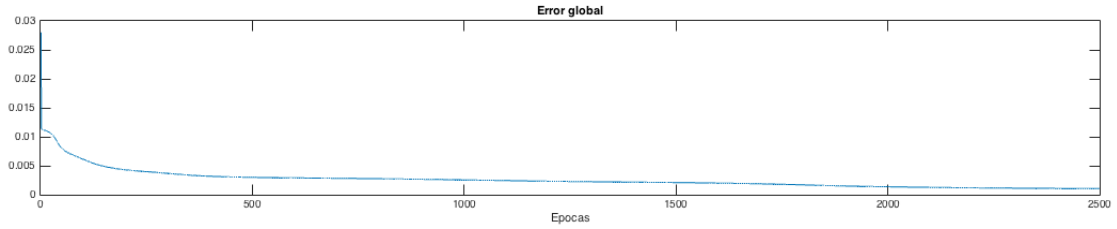

Figure 4: Decreasing of the error in 2500 ephocs using Matlab®.

Estimations by FIS

For the FIS simulation we need to reduce the inputs to 3 in order to follow Zadeh (1965) recommendations for optimal simulations, the main problem with a lot of inputs is that we need to define rules for each combination of imputs, so when we have a lot of imputs we are going to have a lot of possible combinations that are not going to be analytical. In Table 10 we can see the variables chosen to perform the simulations and each membership function.

Variable	Category	Function	Parameters
Bathrooms	Few	Gaussian	[1.68 1]
Bathrooms	Average	Gaussian	[0.8033 2.5]
Bathrooms	Lot	Gaussian	[1.792 4]
Bedrooms	Few	Gaussian	[1.331 1]
Bedrooms	Average	Gaussian	[1 2.5]
Bedrooms	Lot	Gaussian	[1.511 4]
TotalArea	Small	Gaussian	[964.3 425]
TotalArea	Average	Gaussian	[1006 1890]
TotalArea	Big	Gaussian	[887.9 4433]
Price	Low	Gaussian	[61.492 55]
Price	Average	Gaussian	[61.374 375.882]
Price	High	Gaussian	[565.123 1364.26]

Table 10: Membership Functions

Based on the work made by Kusan et al. (2010), we need to define the range of values were each variable have a possible membership function. The variables used for the estimation are going to be the number of bathrooms, the number of bedrooms and the total area (SqFt) in order to explain the market price of each property. We showed in estimations by OLS and GMM that are the most important ones in statistical estimations, that is the reason we use them.

The categories defined are only 3: lot, average and few. This are going to be the membership functions, each variable is going to be involved in one of them based on the value it may have, all the functions are going to be Gaussian based on Kusan et al. (2010) analysis.

In Table 10 the column parameters told us the range of each variable in each membership function, this was defined using the help of Matlab®, but always having the maximun and minimum values defined in Table 1.

Now that we have the functions defined, it is important to create the rules each variable is going to follow in order to obtain the price estimated. In Table 11 we can see the rules defined using the experience of the people involved in the company Midtown Realty Group.

#	**Bathrooms**	**Operator**	**Bedrooms**	**Operator**	**TotalArea**	**Operator**	**Price**
1	Few	AND	Few	AND	Small	THEN	Low
2	Few	AND	Few	AND	Average	THEN	Average
3	Few	AND	Few	AND	Big	THEN	High
4	Average	AND	Few	AND	Small	THEN	Low
5	Average	AND	Few	AND	Average	THEN	Average
6	Average	AND	Few	AND	Big	THEN	High
7	Lot	AND	Few	AND	Small	THEN	Low
8	Lot	AND	Few	AND	Average	THEN	Average
9	Few	AND	Average	AND	Small	THEN	Low
10	Few	AND	Average	AND	Average	THEN	Average
11	Few	AND	Average	AND	Big	THEN	High
12	Few	AND	Lot	AND	Small	THEN	Average
13	Few	AND	Lot	AND	Average	THEN	Average
14	Few	AND	Lot	AND	Big	THEN	High
15	Average	AND	Average	AND	Small	THEN	Average
16	Average	AND	Average	AND	Average	THEN	Average
17	Average	AND	Average	AND	Big	THEN	High
18	Average	AND	Lot	AND	Small	THEN	Average
19	Average	AND	Lot	AND	Average	THEN	Average
20	Average	AND	Lot	AND	Big	THEN	High
21	Lot	AND	Average	AND	Small	THEN	Average
22	Lot	AND	Average	AND	Average	THEN	Average
23	Lot	AND	Lot	AND	Small	THEN	Average
24	Lot	AND	Lot	AND	Average	THEN	High
25	Lot	AND	Lot	AND	Big	THEN	High

Table 11: Knowledge Rules

In Table 11 we can see some easy to understand rules, for example rule number 4 tell us that if bathroom is defined in the membership function as AVERAGE, the number of bedrooms are FEW and the area measured in square feet is SMALL, then the price is going to be LOW. That is how each rule was performed, for this experiment we have 25 rules.

Using the rules and the membership functions we can see the relation between the variables and the market price. In Figure 5 we can see those relations.

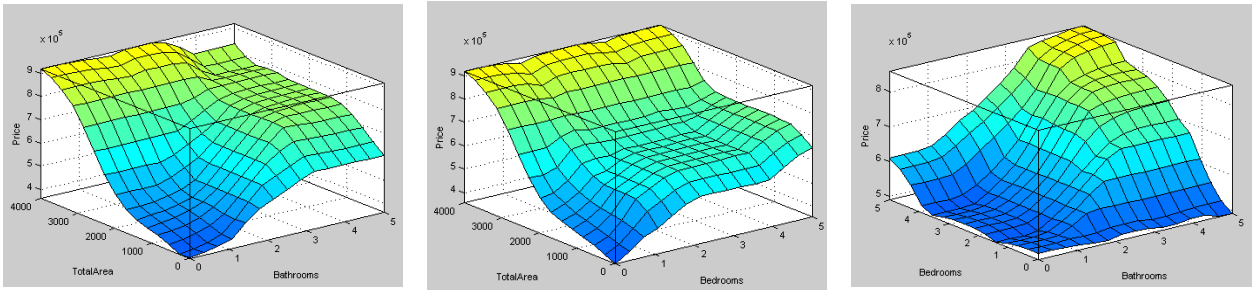

Figure 5: Relation between variables in order to achieve market price

In order to obtain the error, we use the Equation 1 of RMSE shown in the literature review suggested by Kusan et al. (2010).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(T_i - P_i)^2} = 0.8423 \tag{5}$$

In Figure 6 we show the estimated FIS using the rules compared with the real output, this is the defuzzification process mentioned in the methodology. As you can see, the estimation is not very good in terms of curve fitting and minimizing the error.

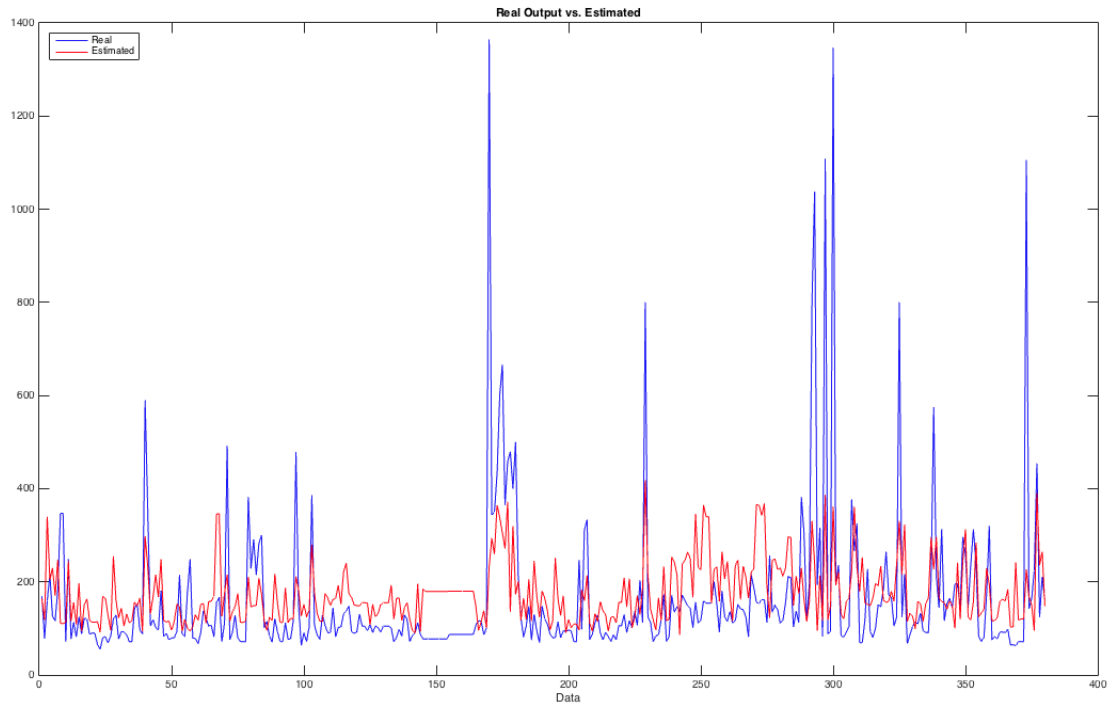

Figure 6: Real output and estimated output for properties market price using FIS in Matlab®

Comparison

First we will compare both econometric methods because we estimated the coefficients involved in the market price; then we will compare the 4 methods using the RMSE. In Table 12 we combined Tables 4 and 6.

Variable	**(OLS_1)**	**(OLS_2)**	**(OLS_3)**	**(GMM_1)**	**(GMM_2)**	**(GMM_3)**
β_0	4.4495* (0.1602)	3.8499* (0.0638)	3.8769* (0.0602)	4.4488* (0.1398)	3.8499* (0.0634)	3.8768* (0.0599)
$Bath$	0.1545* (0.0540)	0.1222† (0.0562)	0.1828* (0.0702)	0.1545* (0.0529)	0.1222† (0.05584)	0.1828* (0.0698)
Bed	0.0590‡ (0.0319)	0.0492 (0.0342)	0.1140* (0.0315)	0.05898‡ (0.0311)	0.0492 (0.03396)	0.1140* (0.0313)
$SqFt$	0.0002† (0.0001)	0.0004* (0.0001)	-	0.0002† (0.0001)	0.0004* (0.0001)	-
Age	-0.0513* (0.0087)	-	-	-0.0512* (0.0082)	-	-
Age^2	0.0008* (0.0001)	-	-	0.0009* (0.0001)	-	-
$Rent$	0.2100* (0.0785)	0.1863† (0.0905)	0.2970* (0.0705)	0.2099* (0.0765)	0.1863† (0.0899)	0.2970* (0.0701)
$Type$	0.0952‡ (0.0509)	-	-	0.0951‡ (0.0499)	-	-
$City$	-0.0006 (0.0457)	-	-	-	-	-
R^2	0.55	0.48	0.44	-	-	-
RMSE	0.3896	0.4173	0.4313	0.3653	0.3287	0.4150

Significance Levels: * 1%, † 5%, ‡ 10%.

Table 12: Estimations using OLS and GMM methods

As we expected, coefficients in both methods are almost the same, having some little variations. The important change is located in the standard deviation of each coefficient, and then the RMSE is minimum in the GMM just because this method is using robust errors and aim to minimize even more the error. The best model (chose by the minimum error) is the number two estimated by GMM, in Equation 6 we show the final relation.

$$lnPrice = 3.8499 + 0.1222 * Bath + 0.0492 * Bed + 0.0004 * SqFt + 0.1863 * Rent \quad (6)$$

GMM in the estimations made is better in all way than OLS, that is because GMM is the generalized method and OLS is just a specific case of GMM, but more over is because the procedure to minimize the error.

On the other hand we have the artificial intelligence methods which can be compared using the RMSE and the curve fitting. In Figure 7 we can see the estimations made with the two methods.

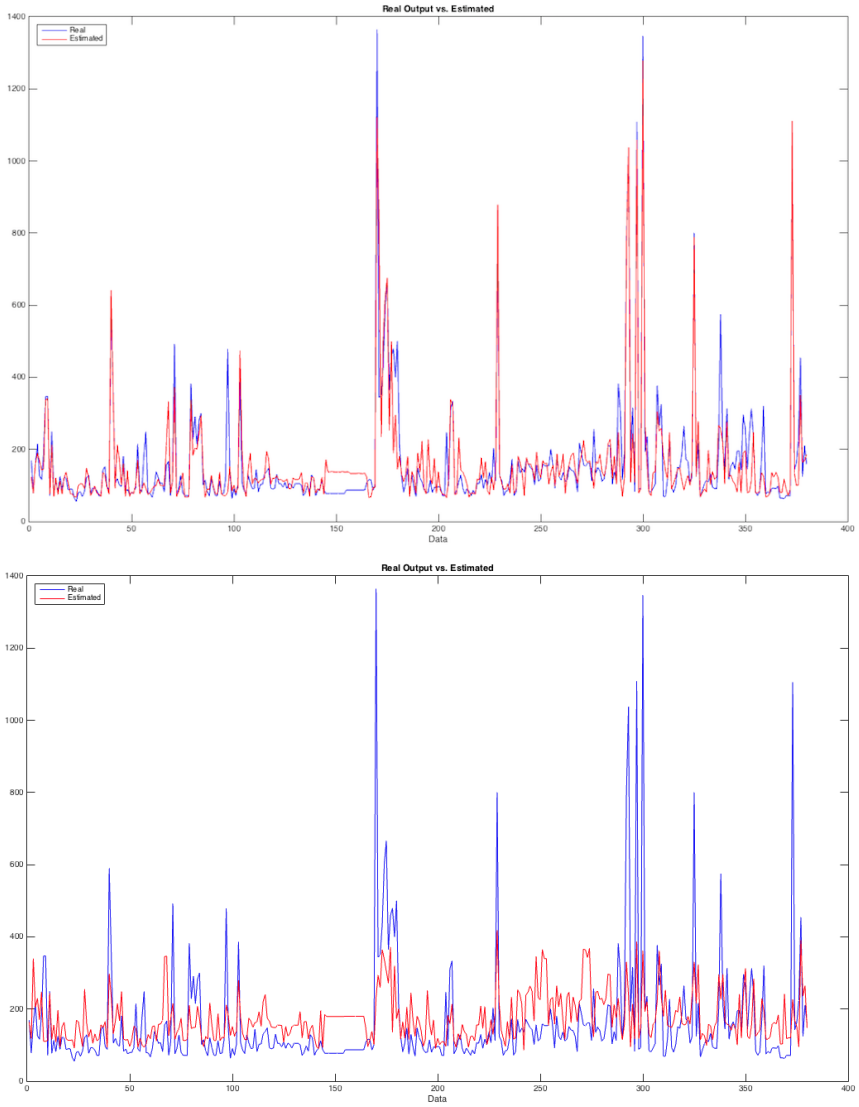

Figure 7: Real output and estimated output for properties market price using ANN (top) and FIS (bottom) in Matlab®

Graphically we can see that with the artificial neural network we have a very good estimation, even the outliers in the data set are explained with good fit by the network; on the other hand the estimations made by the FIS were not that good. This can be explained as Zadeh (1965) confirmed in his studies and Kusan et al. (2010) proved: the FIS is going to work good if the modeler have experience in the area he is trying to explain, because it is neccesary to have rules that fit exactly what the logic of the problem says. One of the problem of this model is that we need to define every single rule for the variables combinations, so we can say that maybe some rules that was important we did not included.

Finally we compare the 4 methods using the RMSE of each estimation, in Table 13 we can see each error.

Method	**RMSE**
OLS	0.3896
GMM	0.3653
ANN	0.2130
FIS	0.8423

Table 13: RMSE of each estimation method

We can see that the best estimation method to obtain the market price for each property is the artificial neural networks with 0.2130. It is important to note that this is not going to happen in all data sets, a lot of variables are involved in this results, as Kusan et al. (2010) said, "because of these limited number of data and factors in certain narrow range, the model cannot be extended for general applications".

Conclusions

As we discussed in the previous section, the best estimation method for assess the housing price of the 380 properties was artificial neural networks. Lai (2006) got similar results as we got using OLS and ANN for his data set, explaining that the nature of the data have nonlinear relations, and therefore neural networks are the most optimal estimation method. On the other hand we have the FIS models that also explain satisfactorily nonlinear relationships of the variables with the dependent variable; however, the problem we have in these models is that they must be well defined number of inputs and the relationships they have with each other, that is, the rules that explain the fuzzy inference model. It is possible that in our study have been omitted relevant rules, that is the reason for a poor estimate of the price of the properties. The inclusion of more rules for the fuzzy model should be given the entire set of possibilities in the region (Kusan et al., 2010).

It is recommended as future work to take time series data and no cross-section data. As we saw, rent paid tends to be a significant variable in the model, so that the historical rent paid could be influential in terms of property prices. Another important variable that should be taken into account as time series is the depreciation of the properties, as these should be treated differently as explained Goodman and Thibodeau (1995) that makes a

transformation of variables age of the property to treat depreciation nonlinearly.

Model comparison was made through minimization of error as suggested by Lai (2006) and Crone and Voith (1992), could have different results if the models are compared based on more curve fitting and the similarity measure distance to the observed points and estimated points.

In terms of explanation we would chose the econometric models because they give us more information about the problem we are modelling, artificial intelligence are good for models of black box were we do not have any information available; because in this problem we know everything about the model, it is better to fit an econometric model.

References

Borst, R. (1991). Artificial neural networks: The next modelling/calibration technlogy for the assessment community? *Property Tax Journal*, *10*(1), 69-94.

Can, A. (1992). Specification and estimation of hedonic housing price models. *Regional Science and Urban Economics*, *22*, 453-474.

Chumncy, E., & Simpson, K. (2005). *Methods and designs for outcomes research.* American Society of Health-System Pharmacists.

Crone, T. M., & Voith, R. (1992). Estimating house price appreciation: A comparison of methods. *Journal of Housing Economics*, *2*, 324-338.

Evans, A., James, H., & Colins, A. (1991). Artificial neural networks: An application to residential valuation in the uk. *Journal of Property Valuation and Investment*, *11*(2), 195-204.

Fletcher, M., Gallimore, P., & Mangan, J. (2000, August). Heteroscedasticity in hedonic house price models. *Journal of Property Research*, *17*(2), 93-108.

Goodman, A., & Thibodeau, T. (1995). Age-related heteroskedasticity in hedonic house price equations. *Journal of Housing Research*, *6*(1), 25-42.

Gujarati, D. (2010). *Econometria* (5th ed.). McGraw-Hill.

Haykin, S. (1999). *Neural networks a comprehensive foundation.* Pearson Prentice Hall.

Kosko, B. (1992). *Neural networks and fuzzy systems.* Prentice Hall Inc.

Kusan, H., Aytekin, O., & Ozdemir, I. (2010). The use of fuzzy logic in predicting house selling price. *Expert Systems with Applications*, *37*, 1808-1813.

Lai, P.-Y. (2006). Analysis of the mass appraisal model using artificial neural network in kaohsiung city. In *23rd pan pacific congress of appraisers, valuers and counselors.*

Liao, W.-C., & Wang, X. (2012). Hedonic house pprice and spatial quantile regression. *Journal of Housing Economics*, *21*, 16-27.

McManus, D., & Mumey, S. (2002, June 4). *System and method for providing house price forecasts based on repeat sales model.* Google Patents. Retrieved from `https://www.google.com/patents/US6401070` (US Patent 6,401,070)

Rossini, P. (1997). Application of artificial neural networks to the valuation of residential property. In *Third annual pacific-rim real estate society conference.*

Tay, D. P., & Ho, D. K. (1991). Artificial intelligence and the mass appraisal of residential apartments. *Journal of Property Valuation and Investment*, *10*(2), 525-540.
Wooldridge, J. M. (2009). *Introduccion a la econometria: Un enfoque moderno* (4th ed.). Cengage Learning.
Zadeh, L. A. (1965). Fuzzy sets. *Information And Control*, *8*, 338-353.

Appendix A

We made the tests for the econometric models in order to avoid problems with the estimations; basically the OLS and GMM methods have the assumptions that there is no multicollinearity between the variables, that the model hace correct specification, the residuals have a normal distribution and same variance in the estimators (Homoskedasticity). We did estimations with robust regression for OLS and robust standard errors for GMM to avoid heteroskedasticity so there is no need to make a test for this assumption (Gujarati, 2010).

Multicollinearity

In Table 14 we did VIF test in order to determine the multicollinearity in the model. It is well-known that if each variable is below 10 then we are not having multicollinearity in the model (Wooldridge, 2009). See that we do have for Age and Age2 but that is because they are the same variable, Age2 is just Age squared. The mean VIF is less than 10 so in the global model we are free of multicollinearity.

Variable	VIF	1/VIF
Age	16.35	0.061157
Age2	15.73	0.063586
Bedroom	3.78	0.264270
Bathroom	3.67	0.272654
SqFt	3.35	0.298573
Rent	1.82	0.549795
Type	1.17	0.852522
Mean VIF	6.55	

Table 14: Variance Inflation Factor.

Ommited Variables

Finally, it is quite important to define whether the model is specified correctly, an error in the model specification can occur when one or more relevant variables are omitted from the model or one or more irrelevant variables are included in the model. Ramsey-RESET test is used, which basically create variables and include them in the model looking to see if there is statistically signicant ones. Next we show the Ramsey-RESET test to the estimated model, and you can see that it accepts the null hypothesis that the model has no relevant variables omitted, so it is concluded that the model have correct specification.

Ramsey-RESET test using powers of fitted values
Ho: model has no omitted values

$$F(3, 91) = 0.69$$
$$Prob > F = 0.5617$$

Normality in Residuals

As defined in Gujarati (2010), it is important to prove that the residuals of the model have a normal distribution; Figure 8 show that a normal distribution fit almost completely in the residuals based on the histogram.

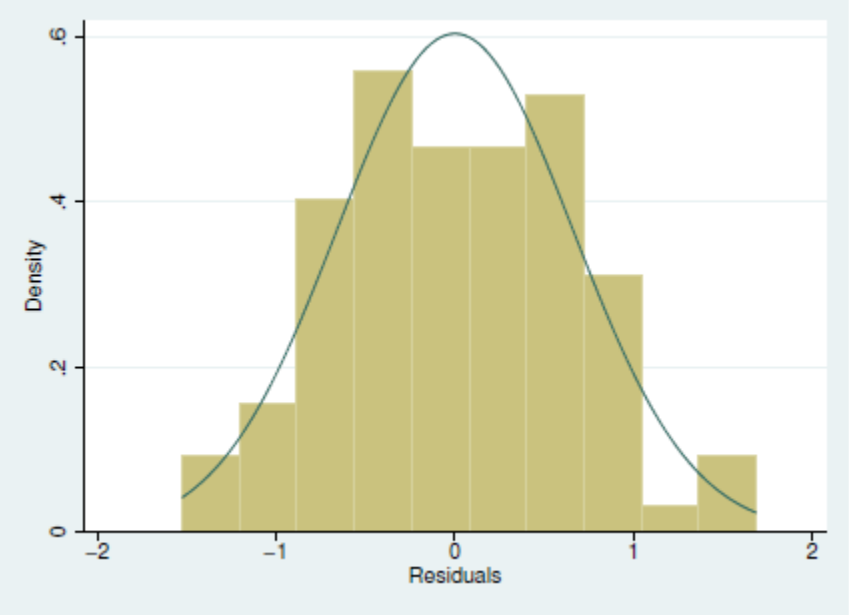

Figure 8: Kernel for residuals compared with a normal distribution.