Pattern Recognition Techniques applied to Evaluation Engineering Problem

Maria Teresinha Arns Steiner; Anselmo Chaves Neto**; Sílvia Neide Bráulio; Valdir Alves*

*UFPR – *Engineering Production Department; **Statistical Department*
CP: 19081; CEP: 81531-990, Curitiba, Paraná, Brazil

Summary
The purpose of this paper is to present a Pattern Recognition methodology composed by Multivariate Statistical Analysis techniques, in order to build a Multiple Linear Regression statistical model to evaluate real estates according to their characteristics (variables, attributes). First, a Clustering Analysis was applied to the data of each urban estate class (apartments, houses or plots) to obtain homogeneous clusters within each class. Next, the Principal Components Analysis (P.C.A.) was applied to solve the multicollinearity problem that may exist among the variables in the model. The scores of the principal components are then the new independent variables and with them, the Multiple Linear Regression model was adjusted for each cluster of similar estates, within each class. This methodology was applied to estates in the city of Campo Mourão, Paraná, Brazil. The model for each similar cluster within each class of evaluated estates presented an adequate adjustment to the data and a satisfactory predictive capacity.

Keywords:Evaluation Engineering, Clustering; Principal Components Analysis, Multiple Linear Regression.

1. Introduction

The real estate market is one of the most dynamic areas of the tertiary economic sector and its main difficulties to evaluate goods come from the estates' characteristics (attributes, variables), which are quite heterogeneous and can keep a relation between them. Estate evaluation, whether for tax collection, for sale, security for financing or others, in general is subjectively made, based upon the personal experience of estate managers, and of other professionals, who compare the data of the estate that is being negotiated with those other estate transactions. In most cases, no scientific procedure is systematically used for this purpose.

The purpose with this paper is to propose a Pattern Recognition methodology based on statistical techniques, able to forecast an estate's value by considering the historical records of similar estates. These value records are those defined in deals that were closed in the past. For such, we considered as a case study the estate market in the city of Campo Mourão, Paraná, Brazil, and in the apartments, houses and plots classes. This way, once a statistical model is obtained for better representing the analyzed market, during a certain period, one will be able to forecast the market value (price) of any estate with the maximum possible precision.

This paper is organized as follows: in Section 2, the problem in the Evaluation Engineering area is delimited by presenting the main norms and concepts related to the theme and some related papers are discussed; in Section 3, we describe the data for the practical problem under consideration. In Section 4, we succinctly describe the statistical techniques used in this work and also present the proposed Pattern Recognition methodology; in Section 5, is described the results of applying the proposed methodology to the problem's data. Finally, in Section 6, the conclusions for the work are presented.

2. Evaluation Engineering

According to [6], Evaluation Engineering is as a part of engineering that includes knowledge from this area, from architecture and others (social, exact and of nature) with the purpose of technically determining the value of a certain good, its rights, fruits and reproduction costs, thus subsidizing decisions with respect to values and involving goods of all natures. Its practitioners may be: engineers, architects, agronomists, each one within their professional qualifications and according to the laws of the Federal Engineering and Architecture Council (or *Conselho Federal de Engenharia e Arquitetura - CONFEA*).

The first works in the Evaluation Engineering area published in Brazil, of which there are records, are dated of the beginning of the 20th century. Methods to evaluate plots were introduced in 1923 and from 1929 on they started to have a systematized use [7].

The Brazilian Association of Technical Norms (or *Associação Brasileira de Normas Técnicas - ABNT*) is the National Forum for Norms. The first norms for estate evaluation appeared in the mid 1950s and were organized by public entities and institutes. The first pre-project of ABNT norms in Evaluation Engineering is dated 1957 and the first Brazilian Norm for Evaluation of Urban Estates is dated 1977, NB-502/77 [6]. This norm was revised in 1989 and originated NBR 5676 (or NB-502/89), registered at INMETRO.

According to NB-502/89, real estates may be classified according to: use (residential, commercial, industrial, institutional or mixed); class (plot, apartment, house, office, store, shed, garage vacancy, mixed, hotel, hospital, theater, club or recreation areas); and clustering (allotted

Manuscript received November 5, 2009
Manuscript revised November 20, 2009

area, house condominium, apartment building, housing development, store group, office building, group of office buildings, group of store units, shopping center or industrial complex). However, we must point out that this work only used data that correspond to estate from the apartment, house and plot classes.

It is interesting to notice that a part of an estate's value can be considered random because there are countless influences in defining its value, this is, one may think of the estate's final value based on a most probable value, increased or decreased of and unpredictable part and according to certain punctual influences. This way, an estate's value follows this statistical model: $Y = \mu + \varepsilon$, where Y is the negotiated value (price); μ is the probable value and ε is the stochastic disturbance term; thus, the expectation for Y is $E(Y) = \mu$.. For further details consult [3], among others.

According to [8], the real estate market has a behavior that is different from other goods markets due to the special characteristics estates show, especially the countless sources of divergence and dissimilarity they present, thus making impossible to compare them directly. Among the factors that distinguish estates from one another, one can mention: long life, fixed spatial position, singularity, high maturing term and high cost of units.

ABNT (NBR5676/90) [1] divides the evaluation methods into two great groups: direct and indirect methods. A method is considered as being direct when the value resulted from the evaluation does not depend on others [6]. Direct methods are divided into market data comparative method (defines values by comparing similar market data) and improvements reproduction costs comparative method (appropriates the improvements' value). According to [6], the use of direct methods has been preferred and when there is enough market data for their use, they are the choice.

A method is considered indirect when it needs the results from some direct method. Indirect methods are divided into income method (defines the value in function of an existing revenue or forecasted by the good in the market, this is, by the good's economic value); unevolutional method (value is estimated by technical-economical feasibility studies for its use) and the residual method (it calculates the difference between the estate's total value and the improvements' value, considering the marketability factor).

Regarding precision levels, evaluation tasks may be classified as follows: expeditious strictness level (the value is obtained without using any mathematical instrument), normal (uses statistic methods and there are requirements with respect to data collection and treatment), strict (the value, which is a result of the method employed, shall have a maximum confidence level of 80%, with null hypothesis tested to the maximum significance level of 5%) and the strict special level, which is characterized by finding a statistical model, the most comprising as possible, this is, one that incorporates the greatest number of characteristics that may contribute to form the value.

The function estimated to form the value must be efficient but not biased. The null hypothesis over the regression model must be rejected only to the maximum significance level of 1% (ANOVA). Null hypothesis for the regression model parameters should be tested to the significance level of 10% for the unilateral test (test "*t*") or 5% on each branch of the bilateral test. The following basic conditions should be analyzed with respect to residues of the model adjusted to the data: have a Gaussian distribution, variance homogeneity and independence. Thus, residues must be Gaussian, independent and identically distributed, this is, $\varepsilon_i \sim N(0,\sigma^2)$.

There are some papers, in the literature, that deal with Evaluation Engineering. One can mention, for instance, the work of [11], which compares the predictive performance of Artificial Neural Networks with the Multiple Regression Analysis, for selling residential houses. Several comparisons were made between the two models varying: data sample size, functional specification and time forecast. In the work [2], the authors examine the effect a view to a lake (Lake Erie, E.U.A.) has on the value of a house. In this study were considered those prices based on the transaction of houses (market price). Results show that besides the variable "view", which is significantly more important than the others, built area and plot size are also important.

In [5], the authors compared the Linear Regression and the Artificial Neural Networks techniques to carry out an estimate of costs for selling or renting estates in the city of Porto Alegre, RS, Brazil. Two databases were evaluated: 1) 1,600 estates offered for sale, 20 attributes each and 2) 500 estates offered for rental, 85 attributes each. From the total number of attributes, only six were selected to train the models. In [9], it is presented two tools for evaluation engineering: generalized linear models and Neural Networks applied to 50 urban plots from three districts in the city of Recife, PE, Brazil.

In [12], it is also made a comparative study between the use of Neural Networks and Multiple Regression Analysis to estimate the sales value of real estates, regarding the offer of 172 middle and low income apartments in the real estate market in the city of Belo Horizonte, MG, Brazil. In [4], the author presented a work that uses Neural Networks to determine the influence variable "accessibility" has upon the value of urban plots, comparing them with the Multiple Regression model, in two cities in São Paulo's countryside (São Carlos and Araçariguama), Brazil, The mentioned variable presented a weight over the final estate's price greater than 34%.

3. Problem Description

The use of the Pattern Recognition methodology proposed here was applied to urban estates from the classes, apartments, houses and plot, in the city of Campo Mourão, Paraná, Brazil. The sample was built with 119 estates (classes), being 44 from the apartment class, 51 from the house class and 24 from the plot class. They are all located in the city's urban area and 80 of these are located inn the city's central area.

Attributes are of the qualitative and quantitative types; the apartments are listed in Attachment 1 and, as can be noticed, they total 21, already divided into clusters (further sections), 17 in cluster A, 19 in cluster B and 19 in cluster C.

4. Pattern Recognition Methodology

The Pattern Recognition methodology to reach the goal consists of the following statistical techniques from the Multivariate Analysis area:

1. Clustering Analysis: through this technique we try to determine the clusters of homogeneous items for each class of estate. In this analysis we used the Euclidian Distance and Ward's method was used as connection method.
2. After forming the homogeneous clusters, discriminants were built with two purposes: evaluate the consistency of the clusters that were obtained and also allocate future items in each one of the clusters that form each class.
3. Following, the Principal Components Analysis was applied to each one of the clusters, from each class, to substitute the values of the original variables by the principal components' scores and circumvent the eventual multicollinearity problem.
1. Finally, a Multiple Linear Regression model was adjusted for each one of the clusters of each estate class. The cash price, called value, was considered the answer variable to the model.

4.1 Description of the Statistical Techniques

Multivariate Analysis is a set of techniques used, among others, to solve problems related to: 1) Covariance structure of random vector $\underline{X}$ (summarized in the covariance or correlation matrix) through Principal Components Analysis; Factor Analysis and Canonic Correlation Analysis; 2) Items Clustering (Cluster Analysis); 3) Pattern Recognition and Classification [10]. In this section we will succinctly describe the multivariate statistical techniques that were used.

a) **Clustering Analysis**

Clustering Analysis consists in a technique that has the purpose of forming homogeneous clusters of objects (estates). Clusters are formed based on their distances (Euclidian, Mahalanobis, among others) or similarities and on a connection method between the partial clusters. The distance that is usually used is the Euclidian Distance: $d(x,y)=\sqrt{\sum_{i=1}^{p}(x_i-y_i)^2}$. The mostly used connection method is Ward's, which minimizes the "loss of information" when "joining" two clusters, by using the criterion of minimizing the sum quadratic error, $\mathrm{SQE}=\sum_{j=1}^{n}(\underline{x}_j-\overline{\underline{x}})'(\underline{x}_j-\overline{\underline{x}}).$]

b) **Quadratic Discrimination Score for Recognition and Classification**

In this study we used the recognition and classification rule based on the minimum total probability of error defined by the quadratic score for population (cluster) *i*, given by:

$$d_i^Q(\underline{x})=-\frac{1}{2}\ln|\Sigma_i|-\frac{1}{2}\left(\underline{x}-\underline{\mu}_i\right)'\Sigma_i^{-1}\left(\underline{x}-\underline{\mu}_i\right)+\ln(p_i)$$

$i = 1, 2, \ldots, g,$

where p_i is the probability that this item belongs to population Π_i; $\underline{\mu}_i$ and Σ_i are respectively the average vector and the covariance matrix of population *i*. These parameters are generally unknown and, therefore, we work with their estimates $\underline{\overline{x}}_i$ and S_i. With respect to the p_i, one can take them as the proportions of the clusters groups' sizes; $\underline{x}_0$ is recognized as belonging to Π_k if: $d_k^Q(\underline{x}_0) > d_i^Q(\underline{x}_0)$ $\forall\ i = 1, 2, \ldots, g$, with $k \neq i$.

c) Principal Components Analysis

Be the random vector $\underline{x}$ with *p* correlated components. This relationship's structure can be summarized in covariance matrix Σ or in correlation matrix ρ. It is known from the Spectral Decomposition Theorem that $\Sigma = P\Lambda P'$ or $\rho = P\Lambda P'$, where *P* is the orthogonal eigenvectors matrix and Λ is the eigenvalues diagonal matrix. Thus, there are *p* non-correlated Principal Components represented by linear combinations $Y_i = \underline{e}_i'\underline{X}$, which recompose this covariance structure, where e_i and λ_i, $i = 1, 2, \ldots, p$ are, respectively, the eigenvectors and eigenvalues of Σ or ρ.

Besides, it is well known that $V(Y_i) = \lambda_i$ expresses the importance of each principal component. A number $m < p$ of Principal Components can represent a significant part of the total variation and it is possible to use them instead of the *p* original variables. A criterion to determine the number of Principal Components to be considered was suggested by Kaiser, in 1960 [10]. It consists in taking a number m of Principal Components that is equal to the number of eigenvalues $\lambda_i \geq 1$. Moreover, it is interesting to consider the part of the variation explained by the m Principal Components above (around) 90%, this is, by extending the mentioned criterion, eigenvalues smaller than "1" may be considered, provided they are close to "1". When applying the Principal Components Analysis, the scores of its m principal components are obtained. This way, matrix $\boldsymbol{X}$ of the model of order n x p is transformed into matrix $\boldsymbol{E}$ of order n x m, $m < p$, corresponding to the scores of the m Principal Components.

d) Multiple Linear Regression for Forecasting

In order to obtain the value of a variable Y in function of other variables X_i, independent from one another, we use a Multiple Linear Regression model, given by: $\underline{Y} = X\,\underline{\beta} + \underline{\varepsilon}$; where $\underline{Y}$ is the observed answers vector of the n observations (estates), $\boldsymbol{X}$ is the model's matrix of order n x p; $\underline{\varepsilon}$ is the errors vector of dimension n and $\underline{\beta}$ (to be estimated) is the parameters vector of dimension p.

Once defined the item's (estate) cluster k of class ℓ, based on the Clustering and on the Recognition and Classification, the adjusted Multiple Linear Regression model is used to estimate the estate's j value by: $\hat{y}_j = \underline{esc}'_j\,\underline{\hat{\beta}}$, where $\underline{esc}_j$ is the components' scores vector $\underline{\hat{\beta}}$ is the parameters estimated vector.

5. Results

The cluster of the apartment class were formed by the Clusters Analysis described in item a of Section 4.1, above. The result indicated that three clusters make up the apartments class, as shown in Figure 1, below. Cluster 1 contains 38.64% of the analyzed apartments, cluster 2 has 22.73% and cluster 3 has 38.64%,

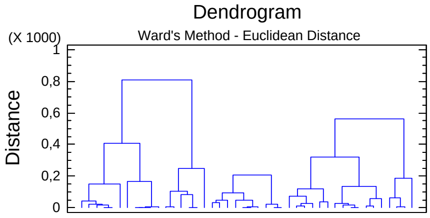

Figure 1. Dendrogram of the three classes formed with the 44 apartments

After the Cluster Analysis, a Discriminant Analysis was made, using the Quadratic Scores, as described in item b of Section 4.1, showing that the classification of the 44 apartments into three classes (clusters) were corrected. We have that from the 17 apartments that belong to class 1, from the 10 that belong to class 2 and from the 17 that belong to class 3, all were also classified correctly. This way, we have a precision of 100%. Thus, results were consistent: from the 44 observations that adjust to the model, 100% were correctly classified. The interpretation of the clusters that were obtained was made according to the attributes in each class, being 17 from cluster A, 19 from cluster B and 19 from cluster C, as we have already mentioned.

Next, we present the most determining aspects in each one of the three clusters:

Cluster 1: all apartments are located downtown; they are located in buildings with at least seven floors; with at least a garage vacancy; buildings have glazed covering; have a minimum area of 160 m^2; at least one elevator; more than two bedrooms; all have a suite; all of them have complete maid lodgings; and prices are higher than R$115,000.00.

Cluster 2: all apartments are located downtown; they are located in buildings with at least 13 floors; buildings have glazed facing, or of marble or granite; more than one garage vacancy; have a minimum area of 220 m^2; buildings newer than 15 years; more than two bedrooms; all have a suite; all of them have more than one elevator; all have complete maid lodgings; and prices are higher than R$175,000.00.

Cluster 3: all apartments have only one garage vacancy; exclusive area is smaller than 132 m^2; low buildings and prices range from R$30,000.00 to R$110,000.00.

Next, the Principal Components Analysis was applied to the data of the original explicative data and obtaining m = 6 components and their scores, as shows Table 1, below.

Comp. Number	Eigenvalue	Percent of Variance	Cum. Percent
1	4.38168	25.775	26.775
2	3.68641	21.685	47.459
3	2.53730	14.925	62.385
4	1.93819	11.401	73.786
5	1.39890	8.229	82.015
6	0.95637	5.626	87.640

Table 1. Principal Components Analysis of the apartments that belong to cluster 1 (m = 6)

Through the results in Table 2, at the end, we can notice that the first component has higher weights in the original variables (in boldface): distance from schools (dschool); distance from supermarkets (dsmarket); distance from hospitals (dhospital); preservation conditions (conservation) and number of bathrooms (bath). The second component has higher weights in variables: number of elevators (elevator); number of bedrooms (nbedr) and the estate's apparent age (ageapparent). The third component has higher weights in variables: number of elevators (elevator); number of rooms (nrooms); how the building is covered (pbuilding); number of bathrooms (bath); number of sitting rooms (nsitrooms); finishing quality (finquality). The fourth component has higher weights in variables: how the building is covered (pbuilding); number of floors the building has (nfloor); area built (area); actual age (agereal); vacancy in garage (vacancy); number of sitting rooms (nsitrooms); finishing quality (finquality). The fifth component has the highest weights: number of vacancies in the garage (vacancy); how the building is covered (pbuilding); number of floors the building has (nfloor) and number of elevators (elevator). Finally, the sixth component has higher weights in variables: how the building is covered (pbuilding); preservation level (conservation) and number of vacancies in the garage (vacancy).

The scores the six components supplied for the 17 apartments are in Table 3, at the end. These are the explicative variables' values that were considered to adjust the linear regression model.

While adjusting the Multiple Linear Regression model $\underline{Y} = X\,\underline{\beta} + \underline{\varepsilon}$**,** it was noticed that the fifth and sixth components were not significantly important, because their *p*-values were greater than 0.05 and, therefore, they were discarded and only the first four were considered, as shown in Table 4, at the end.

The R^2 statistics that measures the adjustment's quality is given by:

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \text{com } 0 < R^2 < 1$$

Supplied the value of R^2 = 0.956557, this is, the adjusted model explains around 96% of the market's price variability. Therefore, the Multiple Linear Regression equation for the apartments belonging to class 1, which describes the relation between price and the four independent components is given by the following equation:

(1) Price = 17412.0 + 18607.3 Y_1 + 4386.74 Y_2 + 7100.19 Y_3 – 23492.7 Y_4

The Analysis of Variance, contained in Table 5, at the end, shows that the hypothesis of no regression is rejected, this is, the model above is truly significant.

The necessary premises to use the linear model and the applied tests were all checked and satisfied by the residues, this is, $\varepsilon_i \sim N(0, \sigma^2)$. The values forecasted by equation (1), adjusted, and the observed values and the error percentages in the forecast a presented in Table 6, at the end.

In the same way, the analysis carried out for the 10 apartments that belong to class 2 resulted in six Principal Components that explain 92.44% of the original data's variability and to the scores that compose the model's matrix of order (10 x 6) the Multiple Linear Regression model was adjusted. The adjustment's determination coefficient was of R^2 = 0.998942, this is, the adjusted model explains almost 100% of the market price's variability.

As for the 17 apartments that belong to cluster 3, the Principal Components Analysis showed that the first seven components explain 88.949% of the original data's variability. Adjusting the model to the matrix of scores of order (17 x 7) supplied a determination coefficient of R^2 = 0.893306, this is, the adjusted model explains close to 90% of the market price's variability.

6. Conclusions

In this paper we propose a Pattern Recognition methodology based on multivariate statistical techniques to forecast prices of urban real estate. This methodology is composed by the following techniques: Clustering Analysis, in which "similar" estates are clustered in terms of their attributes; Quadratic Determinant Scores, in which the consistency of those clusters is checked and one has a criterion for allocating a new item. Next, the Principal Components Analysis is applied in order to obtain $m < p$ components, as well as their independent scores to substitute the original p variables, thus circumventing the multicollinearity problem. Finally, the Multiple Linear Regression model of values vector $\underline{Y}$ is adjusted against the explicative variables summarized in matrix $\boldsymbol{E}$ with order (n x m), this is, $\underline{Y} = \boldsymbol{E}\underline{\beta} + \underline{\varepsilon}$, which supplies an estimate of estate's $\underline{x_0}$ value through equation $\hat{y}_0 = \hat{\underline{\beta}}' \underline{e_0}$, where $\underline{e_0}$ is the vector of correspondent scores.

This methodology was applied to the other two estate classes (51 houses and 24 plots) with a result that was considered quite satisfactory. The quality of the adjustment to the variables, now truly independent, generated the determination coefficients shown in Table 7.

Class	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Apts.	0.95655	0.99894	0.89330	-
Houses	0.91937	0.99226	0.95555	0.96918
Plots	0.97755	0.99745	-	-

Table 7. Values of R^2 for the clusters of the three classes

Therefore, given a new estate in the city of Campo Mourão (apartment, house or plot), of which one wishes to have a value estimate, initially the cluster to which this estate belongs, must be checked and the quadratic scores must be applied. Once the cluster is identified, one can use the Multiple Linear Regression model that corresponds to such cluster. For the apartment class, cluster 1, the defined model is presented in equation (1), section 5. This same way, we have the models for the other situations. This methodology is generic and can be used for any city by obtaining the models definitions for the several different situations in each city.

The multivariate Pattern Recognition methodology that was presented for forecasting real estates prices is reliable, highly appropriate and reaches results with quite satisfactory precision levels. This way, it may serve as support for estate managers when defining estates prices, as well as form people and companies who want to realistically evaluate their assets. One must be aware that the defined Multiple Linear Regression models must be periodically readjusted due to the country’s highly dynamic economy and growth.

References

[1] ABNT (Associação Brasileira de Normas Técnicas), Avaliação de Imóveis Urbanos (NBR 5676 e NBR 502), ABNT, Rio de Janeiro, 2004.

[2] BOND, M. T.; SEILER, V. L.; SEILER, M. J. Residential Real Estate Prices: a Room with a View, The Journal of Real Estate Research, v. 23, n. 1, p. 129-137, 2002.

[3] BRAÚLIO, S. N. Proposta de uma Metodologia para a Avaliação de Imóveis Urbanos baseada em Métodos Estatísticos Multivariados, Dissertação de Mestrado em Métodos Numéricos em Engenharia (Programação Matemática), UFPR, Curitiba, PR, 2005.

[4] BRONDINO, N. C. M. Estudo da Influência da Acessibilidade no Valor de Lotes Urbanos atrabés do uso de Redes Neurais, Tese de Doutorado em Engenharia Civil (Transportes), USP-São Carlos, SP, 1999.

[5] CECHIN, A. L.; SOUTO, A. & GONZÁLEZ, M. A. Análise de Imóveis através de Redes Neurais Artificiais na Cidade de Porto Alegre, Scientia, v.10, n. 2, p. 5-32, 1999.

[6] DANTAS, R.A. Engenharia de Avaliações: uma Introdução à Metodologia Científica, São Paulo: Pini, 2003.

[7] FIKER, J. Avaliação de Imóveis Urbanos, São Paulo: Pini, 1997.

[8] GONZÁLEZ. M. A. S. & FORMOSO, C. T. Análise conceitual das dificuldades na determinação de modelos de formação de preços através da análise de regressão, Engenharia Civil – UM, 8, p. 65-75, 2000.

[9] GUEDES, J. C. Duas Ferramentas Poderosas a Disposição do Engenheiro de Avaliações: Modelos Lineares Generalizados e Redes Neurais, Anais do XI COBREAP, Guarapari, ES, 2001.

[10] JOHNSON, R. A. & WICHERN, D. W. Applied multivariate statistical analysis, New Jersey: Prentice Hall, 1998.

[11] NGUYEN, N. & CRIPPS, A. Predicting Housing Value: A Comparison of Multiple Regression Analysis and Artificial Neural Networks, The Journal of Real Estate Research, vol. 22, no. 3, p. 313-336, 2001.

[12] PELLI NETO, A. & ZÁRATE, L. E. Avaliação de Imóveis Urbanos com a utilização de Redes Neurais Artificiais, Anais do IBAPE – XII COBREAP, Belo Horizonte, MG, 2003.

Maria Teresinha Arns Steiner got her Master's and Ph.D.'s degrees in Production Engineering, on Operations Research area, at Federal University of Santa Catarina, Brazil, and her Pos-Doc, at the Technological Institute of Aeronautics, São José dos Campos, SP, Brazil. She is an Associate Professor at Federal University of Paraná, Curitiba, Paraná, Brazil. She lectures on Engineering Undergraduate Programs and on Numerical Methods in Engineering Graduate Program.
e-mail: tere@mat.ufpr.br

Anselmo Chaves Neto got his Master's degree in Statistical at UNICAMP, Campinas, SP, Brazil and his Ph.D.'s degree in Electrical Engineering at Catholic Pontifícia University of Rio de Janeiro, RJ, Brazil. He is an Associate Professor at Federal University of Paraná, Curitiba, Paraná, Brazil. He lectures on Engineering and Statistical Undergraduate Programs and on Numerical Methods in Engineering Graduate Program. e-mail: anselmo@ufpr.br

Sílvia Neide Bráulio got her Master’s degree in Numerical Methods in Engineering at Federal University of Paraná, Curitiba, Paraná, Brazil.

Valdir Alves got his Master’s degree in Numerical Methods in Engineering at Federal University of Paraná, Curitiba, Paraná, Brazil.

Table 2. Weights of the original variables in each one of the six Principal Components of the apartments that belong to class 1

variable	component	component	component	component	component	component
	1	**2**	**3**	**4**	**5**	**6**
pbuilding	0.089390	0.246161	**-0.526886**	**- 0.979668**	**0.416273**	**-0.570934**
elevator	-0.526883	**-0.413666**	**0.726574**	-0.152862	**0.350890**	-0.068226
vacancy	0.029296	-0.115928	-0.142546	**-0.344625**	**0.513443**	**0.376004**
area	0.289871	-0.133004	0.088613	**-0.431775**	-0.000990	-0.085048
nfloor	0.133738	-0.035762	0.243017	**0.470840**	**0.391287**	0.066217
level	0.208736	0.285383	0.022108	0.102197	0.246664	-0.278490
nrooms	0.225747	0.063101	**-0.531922**	-0.021247	0.019333	0.067219
nsitrooms	0.223168	-0.107456	**-0.355162**	**0.311851**	-0.024862	0.274322
nbedr	0.127221	**0.400858**	-0.255580	-0.002345	-0.100400	0.113880
bath	**0.308159**	-0.027083	**-0.397666**	0.147821	0.005187	-0.255486
dschool	**-0.343604**	0.278393	-0.136115	-0.045763	0.233369	0.169827
dhospital	**-0.363872**	0.279642	-0.118547	-0.052729	-0.026404	0.158290
dsmarket	**-0.381969**	0.262674	-0.096780	-0.010771	0.212156	-0.057252
finquality	0.096366	0.245521	**0.347794**	**0.306414**	-0.070997	0.024395
conservation	**0.327714**	0.141950	0.174912	0.072385	0.268025	**0.455252**
agereal	0.263027	0.240884	0.114627	**-0.410822**	-0.194047	0.064233
ageapparent	0.220789	**0.338124**	0.255992	-0.193006	-0.014837	0.108000

Table 3. Scores of the six principal components of the apartments that belong to cluster 1

	component	component	component	component	component	component
Real States	1	2	3	4	5	6
1.	-2.151200	-0.741581	1.277210	0.867893	-0.379761	-2.779916
2.	-0.215068	0.281233	3.151900	1.286650	2.814090	-0.021131
3.	0.554301	0.073913	-0.793701	2.646601	-1.499770	0.619036
4.	-0.896709	0.062992	-0.310830	0.478471	1.861330	0.619036
5.	2.972630	-1.509750	2.21074	0.225224	-0.129701	0.859525
6.	2.931970	-1.362010	-3.424170	-1.012550	1.872970	0.012452
7.	-0.142195	1.896950	-0.454084	-1.250770	0.675696	0.381225
8.	4.010450	-1.836110	-0.735231	-0.644179	-0.672452	-1.737787
9.	0.663582	2.928590	-1.21238	1.96973	0.009390	-0.404564
10.	0.28557	2.255900	-1.17425	1.96289	-0.827946	0.66776
11.	0.115707	3.100540	0.897485	-2.17611	-0.685172	-
12.	0.115707	3.100540	0.897485	-2.17611	-0.685172	-
13.	2.30919	-1.28958	-0.548849	-0.223296	-0.267799	0.384237
14.	-2.46238	-1.34315	-0.643239	-0.47653	-0.0581804	0.0206226
15.	-2.61556	-1.39672	-0.737628	-0.729764	0.151438	-0.342991
16.	-2.84039	-2.01583	-0.605112	-0.48337	-0.895518	1.09295
17.	1.98277	-2.20593	2.20466	-0.264193	-1.28344	1.02096

Table 4. Adjustment of the Multiple Linear Regression Model for the apartments that belong to class 1 and t Test.

Parameter	Estimate	Standard Error	t Statistic	p-value
CONSTANT	174412.0	3150.0	55.3689	0.0000
PCOMP_1	18607.3	1551.15	11.9958	0.0000
PCOMP_2	4386.74	1691.11	2.594	0.0235
PCOMP_3	7100.19	2038.4	3.48323	0.0045
PCOMP_4	-23492.70	2332.26	-10.0729	0.0000

Table 5. Analysis of Variance of the Regression Model’s Adjustment for apartments belonging to class 1

Source	Sum of Squares	Df	Mean Square	F-Ratio	p-value
Model	4.45699E10	4	1.11425E10	66.06	0.0000
Residual	2.02419E9	12	1.68682E8		
Total	4.65941E10	16			

Table 6. Results for the 17 apartments in Cluster 1

Observed Value y_i **(R\$)**	**Forecasted Value** $\hat{y}_i$ **(R\$)**	**Absolute Error** $y_i - \hat{y}_i$ **(R\$)**	**Percentage Error** **(**$y_i - \hat{y}_i$**) (%)**
130,000.00	128,715.00	1,285.00	0.98846
150,000.00	149,891.00	109.000	0.07267
120,000.00	117,273.00	2,727.00	2.2725
170,000.00	170,992.00	992.000	0.583529
250,000.00	244,826.00	5,174.00	2.0696
220,000.00	219,705.00	295.00	0.13409
200,000.00	198,514.00	1,486.00	0.743
250,000.00	249,772.00	228.00	0.0912
150,000.00	146,357.00	3,643.00	2.42867
120,000.00	127,476.00	7,476.00	6.23
250,000.00	249,481.00	519.00	0.2076
250,000.00	249,481.00	519.00	0.2076
115,000.00	114,222.00	778.00	0.67652
120,000.00	128,543.00	8,543.00	7.119167
140,000.00	142,864.00	2,864.00	2.045714
120,000.00	109,661.00	10,339.00	8.61583
210,000.00	217,228.00	7,228.00	3.441905

ATTACHMENT 1

List of attributes for apartments and their clusters

Attributes	Description	Categories	Apts. Cluster A	Apts. Cluster B	Apts. Cluster C
pbuilding	Identifies how the building is covered	1 = painting 2 = glazed covering 3 = ceramic 4 = marble / granite		X	X
level	Score related to the floor in which the apartment is located.	1 = 1st to 3rd floors 2 = 4th to 6th floors 3 = 7th to 9th floors 4 = 10th or higher	X	X	X
conservation	Identifies the estate's preservation conditions.	1 = bad 2 = regular 3 = good 4 = excellent	X	X	X
agereal	Score related to the building's chronological age (mirrors the technological state).	1 = more than 20 years 2 = 15 to 20 years 3 = 10 to 15 years 4 = 5 to 10 years 5 = 1 to 5 years 6 = up to 1 year	X	X	X
ageapparent	Score related to the apparent building's age.	(idem)	X	X	
dschool	Identifies distance from schools	1 = up to 500 meters 2 = 500 to 800 meters 3 = more than 800 meters	X	X	X
dhospitais	Identifies distance from hospitals	(idem)	X	X	X
dsmarket	Identifies distance from supermarkets.	(idem)	X	X	X
local	Classifies the district and other characteristics of where the residence is.	1 = valuing 0 = indifferent - 1 = devaluing			X
posapartam	Identifies the apartment's position in relation to the building (front, side or back).	1 = front 2 = side 3 = back	X	X	X
finquality	Identifies the several finishing levels.	1 = low 2 = normal 3 = high	X	X	X
nfloor	Number of floors the building has.	Quantity	X	X	X
elevator	Indicates the number of elevators in the building.	Quantity	X	X	X
area	Indicates the apartment's area expressed in square meters.	Area	X	X	X
vacancy	Indicates the number of vacancies for cars, available for the apartment.	Quantity	X	X	
nbedr	Indicates the number of bedrooms in the apartment.	Quantity	X	X	X
maidlod	Indicates the existence (or not) of maid lodgings.	0 = inexistent 1 = existent		X	X
suite	Indicate the presence (or not) of suites.	0 = inexistent 1 = existent			X
nsitrooms	Indicates the number of sitting rooms in the apartment.	Quantity	X	X	X
nrooms	Indicates the total number of rooms the apartment has.	Quantity	X	X	X
bath	Indicates the number of bathrooms in the apartment.	Quantity	X	X	X
Total Attributes	**21**		**17**	**19**	**19**

(source: Imobiliária Tapowik, Guarapuava, Paraná, Brazil)