ATHENS, 26 may 2004

Criteria of database quality appraisement and choice stochastic models in prediction of real estate market value

PhD Anna Barańska
Terrain Information Department
Faculty of Mining Surveying and Environmental Engineering
University of Science and Technology
Krakow, POLAND

INTRODUCTION

Creation of mathematical pricing model for a chosen market is a problem extremely complex, because it involves an appropriate preparation of modelling database as regards its completeness and reliability, taking into consideration an optimum selection of variables. To find the "best" model, it is necessary to examine a great amount of information in database. Therefore, it is necessary to formulate some criteria of choosing an optimal model, as well as the criteria allowing the estimation of the quality of modelling database form. It concerns mostly relation between the quantity of variables and the base size.

Definitions of invariant parameters

1. Coefficient of determination R^2

$$R^2 = 1 - \frac{\det K}{\det K_0}, \tag{1}$$

where:

$$K = \begin{pmatrix} 1 & r_{01} & r_{02} & \dots & r_{0k} \\ r_{10} & 1 & r_{12} & \dots & r_{1k} \\ r_{20} & r_{21} & 1 & \dots & r_{2k} \\ \dots & \dots & \dots & \dots & \dots \\ r_{k0} & r_{k1} & r_{k2} & \dots & 1 \end{pmatrix} \tag{2}$$

K – correlation matrix,
$\det K$ – determinant of correlation matrix,
$\det K_0$ – determinant of submatrix, created by cancelling of the first row vector and the first column in matrix K, i.e. correlation coefficients concerning dependent variable (price).
R – coefficient of linear multiple correlation, determining the degree of matching the hyperplane to the point pattern representing prices and attributes of individual real estates

2. Parameter settled on the basis of covariance matrix trace:

$$\sigma_{tr} = \frac{1}{W_{śr}} \sqrt{\frac{tr\{Cov\{W\}\}}{n}}, \tag{3}$$

where:

$tr\{Cov\{W\}\}$ – trace of covariance matrix for predicted market values of real estates establishing a pricing model,
$W_{śr}$ – mean value of predicted real estates market values establishing a pricing model,
n – number of real estates used to estimate a model.

3. Prameter calculated from the covariance matrix determinant:

$$\sigma_{det} = \frac{1}{W_{śr}} \sqrt[u^2]{det\{Cov\{W\}\}}, \tag{4}$$

where:

$det\{Cov\{W\}\}$ – determinant of covariance matrix for predicted market values of real estates establishing a pricing model,
$W_{śr}$ – mean value of predicted real estates market values establishing a pricing model,
u – number of independent variables occurring in a model.

The analysis of formulas (3) and (4) indicates that these quantities σ_{tr}, σ_{det} constitute some kind of measure of dispersion round the mean model value, so they can be objective parameters applied to formulate criteria of reliability of databases used in modelling of market value.

METHODOLOGY OF INVESTIGATION

Information on land properties for housing as the object of commercial traffic was used as starting material for investigation. It comes from 10 different local markets of properties. Great differences of prices and variable dynamics of transactions are characteristic features of analysed markets. Gathered databases contain 20 to 130 properties. Information includes totally 530 land properties.

The market of estates is created for every town (commune) separately. For big towns, separate estate markets are created for particular quarters.

Achieved data have been subject of a pre-treatment aiming to prepare pricing models to the tests. Preliminary analyses permitted to define methods of grouping achieved data, isolated features influencing estate value and imagined influence rate of respective features. By this treatment, databases are brought to connectivity allowing to acquire more reliable results of investigation.

Modelling of real estate market value

In modelling process of real estates market, first, in every database, a multidimensional linear model as linear multiple regression has been tested:

$$F(X_i, a) = w_i = a_0 + \sum_{k=1}^{u} X_{ik} * a_k \qquad (5)$$

where:

w_i – model value of i-th estate in a given database,
X_i – attributes value vector for i-th estate ($1 \times u$),
X_{ik} – value of attribute k for i-th estate,
a_0 – free term in the model,
a – vector of multiple regression coefficient ($(u+1) \times 1$),
a_k – coefficient of regression standing by attribute k.

If the analysed market of real estates was steady, variability of prices in relation to respective attributes would be linear. However, the market of real estates in Poland being not stable, variability of prices often is not linear and sometimes changes are even abrupt. Consequently, in cases where linear model proved unreliable, estimation of non-linear model parameters has been done.

The following elementary functions have been considered: **linear**, **polynomial**, **power**, **logarithmic**, **exponential**, **irrational**. These functions describe variability of market prices in relation to the established attribute.

Reliability measure for regression model

$$q^2 = 1 - \frac{\sum_{i=1}^{n} [y_i - g(X_k)]^2}{\sum_{i=1}^{n} [y_i - \hat{y}]^2} \qquad (6)$$

where:

y_i – unit price of i-th estate in database,
$\hat{y}$ – mean value (arithmetic mean) of estate price from a database,
$g(X_k)$ – predicted unit price for attribute k of a real estate, resulting from admitted non-linear pricing model.

This coefficient value may be a criterion for selecting suitable form of function g.

Selected forms of function g were used to create multidimensional models, i.e. global non-linear functions F for respective databases:

$$W = F(X, a) \qquad (7)$$

where:

W – set of predicted prices generating a pricing model,
X – multidimensional variable representing real estate attributes,
a – vector of model parameters.

Each of estimated non-linear models has been reduced to a linear model with the assistance of Taylor series expansion.

Reliability of estimated models has been verified by testing the hypothesis on variance equality of the part explained by regression model and of the part unexplained. Fisher-Snedecor test has been used here, at the significance level $p = 0.05$, for which test statistics has the following form:

$$F = \frac{R^2}{1 - R^2} * \frac{n - m}{m - 1} \qquad (8)$$

where:

R – coefficient of multiple linear correlation,
n – size of trial (quantity of real estates in a database),
u – quantity of independent variables in a model,
$m = u + 1$ – quantity of estimated parameters of a pricing model.

The above test not only examines the absolute variance ration of explained and unexplained parts, but also takes into account the necessity to retain in a model right proportions between the quantity of cases and the quantity of unknowns.

Determination of covariance matrix for model values of estates

$$\hat{\sigma}_0^2 = \frac{Y^T Y - \hat{a}^T X^T Y}{n-u-1}$$

$$Cov(\hat{a}) = \hat{\sigma}_0^2 \cdot (X^T X)^{-1}$$

$$Cov(W) = \hat{\sigma}_0^2 \cdot X^T (X^T X)^{-1} X$$

$\hat{\sigma}_0^2$ – unloaded estimator of remainder variance (determining inaccuracy of model parameters estimation),
X – matrix containing independent variables (attributes),
a – vector of multiple linear regression coefficients,
Y – vector of dependent random variable (prices of estates).

Determination of invariants values

In sum, on all estates databases, many different models determining estate market value have been tested. First, all models with multiple correlation coefficient less than 0.60 have been eliminated. For each of remaining 97 models, after the correlation matrix for variables occurring in model and covariance matrix for estate model values establishing a pricing model have been determined, the values of three defined invariants were calculated. Thereby, 97 sets of these three numbers, were achieved.

Besides the values of invariants, the following table contains: the name of the town where the data used to modelling were acquired, the type of model and constant values describing the database and the model.

VALUES OF INVARIANTS

TOWN	model	*n*	*u*	*m*	*k*	σ_0^2	R^2	σ_{tr}	σ_{det}
Bolesław	linear	18	7	8	10	5,206	,702	,172	,0001
Bolesław	non-linear	18	8	11	9	0,986	,949	,077	,0005
Busko Zdrój	linear	31	15	16	15	37,875	,810	,090	,0030
Busko Zdrój	non-linear	31	15	16	15	10,212	,949	,047	,0025
Krowodrza	non-linear	38	13	14	24	2324,053	,695	,342	,0003
Krowodrza II	linear	131	12	13	118	779,711	,919	,040	,0000
Krowodrza II	non-linear	121	16	17	104	343,562	,962	,033	,0000
Nowy Sącz	linear	28	12	13	15	11,458	,898	,105	,0015
Nowy Sącz	non-linear	30	16	17	13	21,942	,859	,151	,0095
Proszowice	linear	60	9	10	50	18,913	,923	,197	,0025
Proszowice	non-linear	57	8	10	48	9,560	,963	,121	,0000
Przeworsk	non-linear	30	18	19	11	1,158	,643	,038	,0136
Rzeszów	non-linear	48	13	14	34	33,640	,824	,059	,0004
Świdnik	linear	41	11	12	29	25,487	,676	,088	,0000
Świdnik	non-linear	40	15	16	24	15,515	,833	,079	,0012
Trzyciąż	non-linear	50	14	15	35	1,289	,828	,053	,0002

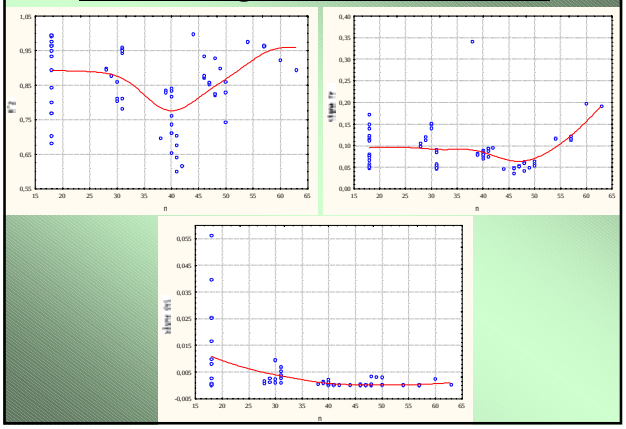
Study on dependence of invariants on quantities describing the database of real estates and the pricing model

In order to formulate criteria for real estates database and model reliability estimation, several scatter diagrams of invariants dependence on constant quantities describing a database and applied pricing model have been made. On each of these diagrams, a trend line, estimated with least squares method, has been put on. It enables to confirm the occurrence or non-occurrence of any relations between these quantities and, in consequence, to draw conclusions concerning the model or the database.

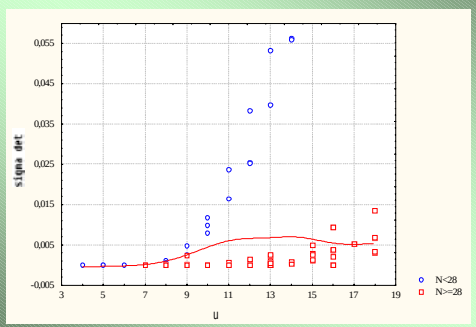
Diagrams of invariants dependence on following parameters have been made:

- number of real estates in a database – n,
- number of independent variables in a model – u,
- number of degrees of freedom – $k = n - m$,
- remainder variance – σ_0^2 .

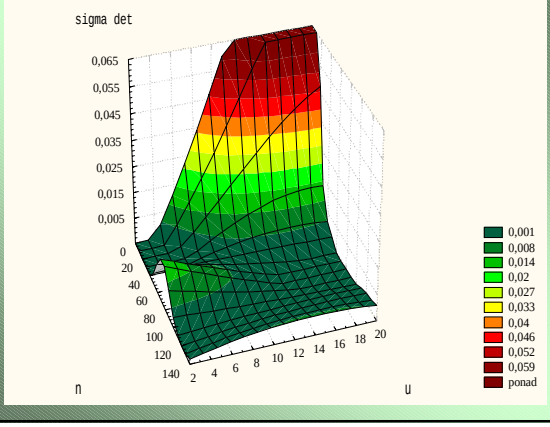
Selected diagrams: *invariants* with *n*



Scatter diagram *(third invariant with number of independent variables 'u')*



Surface diagram 3D *sigma det (n, u)*



CONCLUSIONS

The main purpose of the study was the presentation of possibilities for valuation of a database and a model applied to estimate the real estate market value, using defined covariance matrix transformation parameters – invariants for model values of estates. After the parameters of about a hundred models, tested on 10 databases of different size (18 to 132 estates) and of different amount of features taken into account (4 to 18), have been estimated, and after the values of three invariants: R^2, σ_{tr}, σ_{det} - have been determined, on the basis of these invariants scatter diagrams in conjunction with characteristics of databases and models, the following conclusions can be formulated:

1. Optimal quantity of real estates in a database for modelling the estates values should correspond to threefold number of determined parameters of a pricing model. It must be pointed that this number should be, at least, twice as large as the number of determined model parameters, and that its enlarging (more than quadruple number of parameters) in most cases does not bring improvement of the model.
2. Maximum number of independent variables should not exceed 14 parameters. Optimum number of parameters (attributes) should be determined by the preliminary analysis of real estates market.
3. The number of degrees of freedom in a pricing model should be contained between 28 and 42.
4. The selection of a database for estimation of a pricing model can be acknowledged to be optimal if the value of invariant σ_{det} is situated in the interval: $\sigma_{det} \in (0.0008; 0.0055)$.
5. The selection of model for estimation of real estates values can be acknowledged as satisfying when the value of determination coefficient R^2 fulfils the inequality: $R^2 \geq 0.837$.
6. Using the parameter σ_{tr}, criteria of selecting a database for estimation of a pricing model can not be formulated, because the value of this parameter shows a strong fluctuation.
7. From the comparison of conclusions 4 and 6 it may be observed that for selection of a correct pricing model, consideration of nothing but variances of model values is not sufficient. It is necessary to take into account the whole covariance matrix for model values $Cov(W)$.