ARGUMENTA OECONOMICA
No 2 (21) 2008
PL ISSN 1233-5835

Radosław Wiśniewski *

SELECTED ASPECTS OF THE USE OF ARTIFICIAL NEURAL NETWORKS FOR THE MASS APPRAISAL OF REAL ESTATES IN POLAND

Mass appraisal is a specific process of property valuation employing mass appraisal methods. The results of recent research and practical implementation show that property value estimation may be based on statistical methods. The aims of the present study were as follows: (1) to discuss the possibilities of an artificial neural network application (ANN MLP – Multi-Layer Perceptron), (2) to describe the specific characteristics of ANN, (3) to evaluate the obtained results, and (4) to develop the methodology of ANN application in the mass appraisal of real estate in Poland. The results obtained using multiple regression models were compared with those obtained using ANN MLP models.

Keywords: mass appraisal, artificial neural networks, regression, valuation of real estates

1. INTRODUCTION

The development of methods for property value estimation is accompanied by numerous problems, resulting primarily from the complex structure of market patterns and relations, but also from the unavailability of information about the actual system of value creation in the land market. Moreover, there exist no analytical forms of the above relations, which often makes it impossible to formulate algorithms for property value calculation.

The processes of property value creation are also complex and dependent on a variety of factors whose impact on the above value is difficult to determine, or can be only roughly estimated. Those processes are considered deterministic. However, an analysis of the factors affecting the market value of real estates provides many arguments for combining stochastic (i.e. uncertainty of transaction conditions, uncertainty of real estate behaviour) and deterministic methods (i.e. date of transaction, shape and area of a land parcel). In such a situation problems should be solved by means of specific research

* Department of Land Management and Regional Development, University of Warmia and Mazury in Olsztyn

procedures. Furthermore, such methods are expected to detect market relations, memorize them and then use this knowledge for solving certain problems.

The progress in research into the principles governing the processes of logical thinking enabled their application to highly advanced technology. Attempts at the development of mathematical notation of the processes of logical thinking, learning and memorizing lead to concrete technological solutions. As a result, the phenomena described artificially (in the form of algorithms) allow to look from a new perspective at the problems which have not been solved to date or whose solutions have been unsatisfactory.

The growing interest in the principles governing the land market and attempts at explaining them by analysis of actual transactions lead to the solutions in which many levels of the investigated phenomenon are taken into account. These levels are interrelated in a direct and indirect way. The direct interrelations, in contrast to the indirect ones, can be observed and measured. An analysis involving the use of artificial neural networks allows to find this factor and apply it as an element connecting the undetermined levels of phenomenon explanation. The above model permits the application of both theoretical and practical solutions.

Neural networks are in the center of interest stemming from the hope that part of the variation in the space of explanatory variables regarding association, inference, memorizing and generalization can be represented by artificial means.

2. REAL ESTATE MASS APPRAISAL

Mass appraisal may be defined as a systematic appraisal of property groups using standardized procedures. The accurate assessment of the value of a predefined set of properties, or one particular property, indirectly, using a model, for a given practical purpose, is the main target of those methodologies (Kauko 2007).The social-economic relevance of this topic cannot be over-stated if we consider that the main target of such a methodology is an accurate assessment of the value of a predefined set of properties, or one particular property, indirectly, using a model, for a given practical purpose. The importance of mass appraisal may be also seen from the perspective of the relationships between property value, property characteristics, and urban, social and economic problems.

Arguably, the standard multiple regression analysis (MRA) based on hedonic price models is not suitable for capturing all the necessary information involved in value formation, and literature devoted to the further

development of value modeling tools is evolving. Although the problems are highlighted, MRA remains at the moment the most important theoretical framework in mass appraisal (Kauko 2007).

Two related modeling traditions exist today, both of which deploy MRA for estimation, namely the model driven *hedonic approach* and the data driven *statistical approach*. Hedonic price models comprise the most frequently applied models in the valuation practice as well as in monitoring the housing market. In these models the variables are usually of two basic types: internal physical (i.e. house- and plot-specific, structural) and external locational. On top of that there may be additional variables, most notably some type of inflation control. The purpose of developing the hedonic price model was to enable an econometric analysis of large databases of price and other recorded information describing the nature of the property and its vicinity, and possibly some specific (other) circumstances of the transaction. A more practical or theoretical statistical, especially regression analysis-based value/price-modeling tradition has been applied in order to provide tools for valuation conducted by the public and private sectors in many countries with convenient land information infrastructure (i.e. readily available digital register information with the possibility of multiple spatial aggregation; Kauko 2007).

Artificial Neural Networks (ANN) are an alternative to traditional methods for property valuation that attempt to increase accuracy by reducing the impact of qualitative inputs, and by more accurately matching the underlying relationships within datasets.

When using MRA, the methodological problems of functional form misspecification, nonlinearity, multicollinearity and heteroskedasticity should be addressed. Multicollinearity does not affect the predictive ability of MRA or that of ANN because the inferences are made within the jointly defined region of the observations. Multicollinearity, however, does make it infeasible to disentangle the effects of the supposedly independent variables. Heteroskedasticity is normally present when cross-sectional data are used. In addition to the model's methodological problems, leaving out a relevant explanatory variable is another source of error when using MRA and ANN. This is often due to the unavailability of data (Nguyen Cripps 2001).

3. ANN THEORY

Artificial neural networks are adjusted to solving a given problem through learning, with the help of a series of typical stimuli and desirable

reactions corresponding to them (Cruse 2006a, 2006b). Similar opinions were also expressed by Rutkowski (2005). According to those opinions, ANNs are techniques employing stochastic algorithms of model fitting through learning. The cited authors also emphasize the fact (...) *that ANNs are based on noised numerical data, and that learning algorithms allow to build unidirectional or recurrent models of processes. Such models are characterized by the architecture of non-linear elements with a complex network of linear connections, often with local or global feedbacks* (...). Experts solving the problems of property value estimation, both in Poland and abroad, agree that attempts should be made at building a structure based on the above assumptions.

Another argument for the practical application of ANNs to the mass appraisal of real estate is the fact that they adapt functions describing actual models in the process of learning on the basis of data. The rules applied and the iterative learning process lead to the optimum use of intelligent, algebraic-logical models of outcome creation, subject to the assumptions made (Rutkowski 2005 p. 159; Cruse 2006a, 2006b). ANNs can reproduce behaviours from the learning sequence, generate conclusions, memorize them and put them to use. The ability of ANNs to make generalizations is also important.

A comparison of ANN models and multiple regression models shows that ANN models are distinguished by certain attributes which support their use in selected processes of the mass appraisal of real estates in Poland. Due to their architecture, mode of operation and range of applications, ANN models can accurately represent the complex structure of relationships observed in the land market and properties of this system. The following natural predispositions of ANNs can be used for the purpose of mass appraisal:

- the ability to imitate the functions of the land market system due to the application of flexible solutions generated at the learning stage; as a result, ANN models reflect the functions of the land market much better than the respective multiple regression models, based upon a fixed functional pattern;
- the ability to generalize results burdened with a high degree of variation, complexity and uncertainty; multiple regression models are sensitive to outlying and variable data, i.e. data coming from non-homogeneous sets of observations;
- the ability to model both individual and group behaviour, even under conditions of delayed responses which manifest themselves in the processes of variance grouping in the information structure – this is a distinctive feature of ANN models, which enables to create the outcome in a natural way, as observed in reality; the multiple regression models used in this study

have no such capacities, because they are based on a fixed functional pattern.

Other advantages of ANN models, weighing in favour of their use for the mass appraisal of real estates, compared to multiple regression models, include:

a) the increased ability to generalize, through division and redundance, the subspace of solutions;

b) the ability to process disjoint sets of variables with a high degree of redundance;

c) less strict requirements regarding the selection of the optimum model architecture;

d) unlimited possibilities of task structuring and of introducing functional relationships between modules;

e) the option of module specialization with respect to the processing of specified groups of cases, i.e. outlying cases;

f) the increased possibility of performing a sensitivity analysis.

4. ANN APPLICATION IN THE MASS APPRAISAL OF REAL ESTATES

One of the premises of using ANNs for property valuation is the large quantity of information that must be analyzed if the solutions obtained are to fulfill the conditions assumed. The processes taking place in the land market are difficult to observe and predict. This results, among others, from chaotic behaviour in the deterministic sense, non-systematic changes caused by randomness and the occurrence of gross errors. As a consequence, considerable amounts of information must be processed to confirm either stochastic or deterministic behaviour.

It is difficult to concentrate on many variables at a time. Therefore, specialists usually apply certain simplifications, making some assumptions or focusing on two to three factors only. This allows to solve problems, but only for the assumptions adopted. ANNs are structures of parallel and quick information processing, so they allow to analyze many variables and consider different levels of "value" for each of them (e.g. location: very good, good, poor, etc.).

Another advantage of artificial structures of data processing is that they do not display a tendency towards attaching too much weight to individual (expert) – often acquired – causes, while ignoring some other. Taking into account changing market conditions, those other causes may be of primary importance and may affect the ultimate solution to a high degree.

Many problems associated with the description of the real estate market,

resulting from its complex structure, cannot be solved by a cause-and-effect analysis. Those problems require an analysis based on interrelations and interdependences. Such an analysis involves a variety of interacting factors and overlapping sources (causes). Neural networks – capable of memorizing and associating great numbers of causal elements – can solve such problems at a much deeper level, beyond human understanding. ANNs permit a multiple analysis of large quantities of data, finding hidden regularities. As a result, single value-creating "signals" can be combined into bigger, more aggregated units. ANNs, acting as algorithms, are systematic and persistent. They allow to recognize seemingly non-existing relations in the land market.

Practical applications of artificial intelligence for the mass appraisal of real estates can be divided into three groups.

The first group comprises the processing of the acquired data and information, and their preparation for further analyses. The choice of techniques and methods of database analysis must be preceded by solving both theoretical and practical problems. The problems taken into account in the present study included:

1. Methods for quantification of property attributes.
2. Selection of solution-creating variables.
3. Procedures of outlying case elimination.
4. Analyses concerning database division (learning, testing and verifying cases).

The second group comprises the selection of multiple regression models and optimal architecture of artificial neural networks. In practice artificial networks are selected primarily by iterative approximations. In the case of multiple regression, those approximations concern mainly the choice of a model, whereas in that of ANNs iterative processes are used at each stage of testing and practical application of neural structures. Particular attention should be paid to the selection of the optimum structure of a neural network, being a tool employed to attain research objectives.

The third group comprises the indices of artificial intelligence model evaluation.

The proposal for the ANN model application, presented in this paper, refers to the local conditions, in particular to the existing sources of information, costs of data acquisition and economic aspects of property value estimation by artificial intelligence models. Figure 1 shows a schematic diagram of ANN application in the mass appraisal of real estates in Poland.

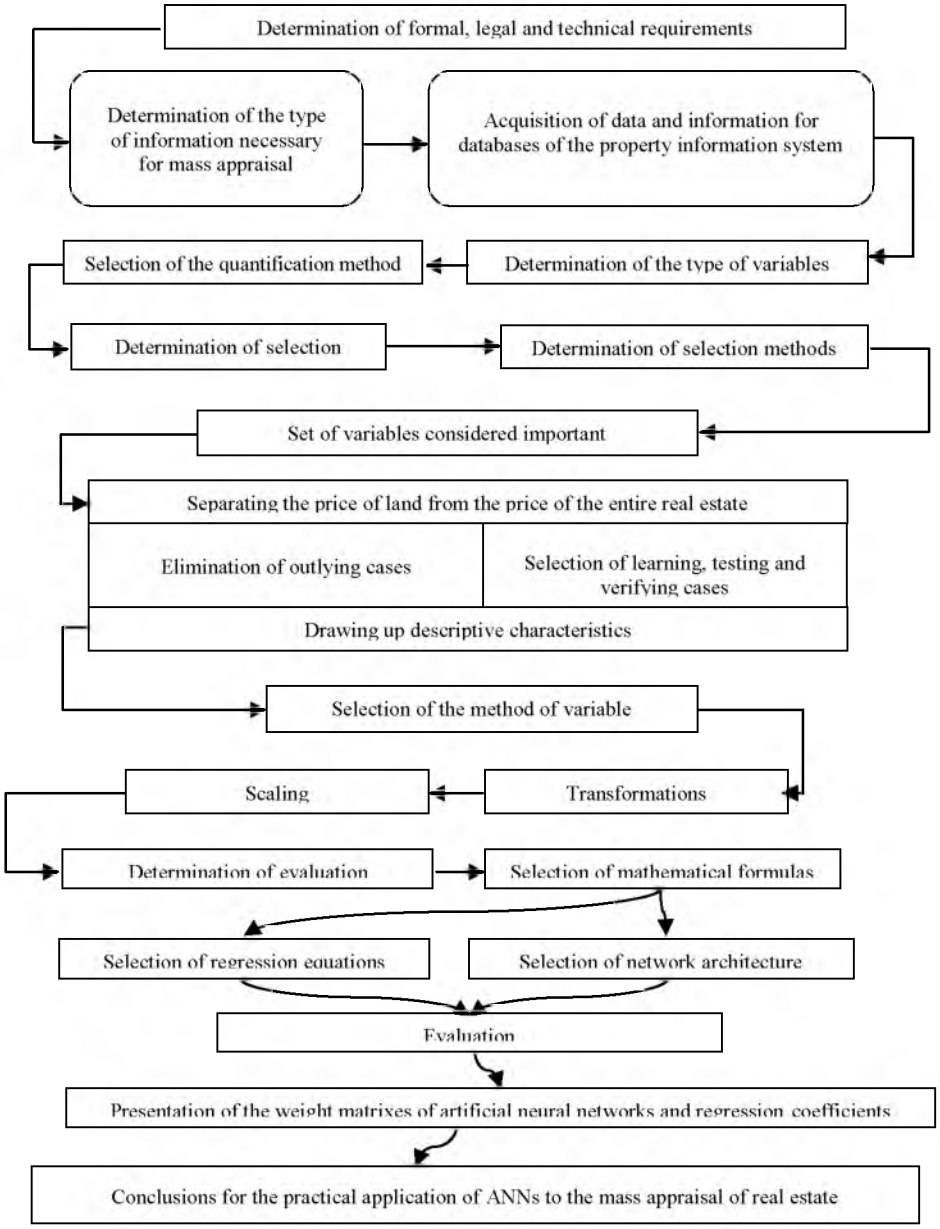

Figure. 1. Schematic diagram – methodology
Source: own study

According to the main hypothesis proposed in this paper, ANN models provide more reliable results than multiple regression models when applied in the processes of property value modeling for the purpose of the mass appraisal of real estates in Poland. According to the sub-hypothesis, ANN models may be used as an auxiliary tool in the processes of mass appraisal, in particular when it is necessary to apply generalized parameters reflecting trends in the land market, e.g. in the verification processes between taxation periods.

A thorough representative analysis involving the use of artificial neural networks for the purpose of mass appraisal of several types of real estates can be found in Wiśniewski (1998). This study is a continuation of previous research, launched in 1997. Recent alterations and amendments to the legal regulations concerning the mass appraisal of real estates in Poland have been taken into account. Due to the paper form and length limitations, the empirical materials and the obtained results should be regarded as illustratory only. However, it must be stressed that the presented example is representative of the results of investigations conducted systematically since 1997.

5. PRACTICAL APPLICATION OF ANN TO THE MASS APPRAISAL OF REAL ESTATES

The creative effect of the land market is reflected in the level of transaction prices for a given type of real estate. It is difficult to specify the factors determining the economic situation in the land market. However, an analysis of the rising tendencies observed each year in this market indicates that some factors can be referred to as value-creating. This group of factors is difficult to identify due to:

1. The specific character of the land market, including the law of supply and demand, and characteristic attributes of a real estate as an article of trade.
2. The principles governing the macroeconomic, political and social situation which – although present on a macro scale – can be also observed in local markets.

The above elements, combined with the subjective nature of purchase-sale transactions, affect market tendencies. They also influence real estate attributes, which in turn affect decision-making processes in the land market. Such a problems requires the use of procedures allowing to analyze the effects of explanatory variables on the explained variable (value).

The results presented below are based on own study conducted during the years 1998 – 2006, concerning an analysis of a non-built-up land property in the land market of the city of Olsztyn. The object of the study was selected due to:

- Location – Olsztyn is a big urban centre in north-eastern Poland.
- Representativeness – the investigated object performs a creative role in the local land market, which makes it representative and indicative of the general trends observed in the region.
- Availability of the necessary data.

5.1. Objective of the analysis and the applied models

The objective of the analysis was to determine the value of a non-built-up land property, using models of artificial neural networks and multiple regression (comparative analysis). The following models were applied:

a) unidirectional three- and four-layer sigmoid *neural networks* (multiple-layer perceptron) – parameters selected iteratively.

b) *multiple regression – power model*

$$V = \beta_1 * x_1^{\beta_2} * x_2^{\beta_3} * \ldots * x_p^{\beta_{p+1}} \tag{1}$$

where: V – value of land property; $P_U = LP - (P_W + P_T)$ model parameters; $x_1, \ldots, x_p$ – independent variables

c) *multiple regression – exponential model*

$$V = \beta_1 * \exp\left(\beta_2 x_1 + \beta_3 x_2 + \ldots + \beta_{p+1} x_p\right) \tag{2}$$

symbols as above.

The above models were selected following their practical application, verification and testing. The obtained evaluation parameters indicate that those models can be used for value estimation.

5.2. Data preparation

a) *Selection of variables.* The solutions applied to date in the mass and individual appraisal of real estates were used in the analysis. Cases of individual appraisal described in professional literature were taken into

account to widen the range of independent variables characterizing the investigated land property. A set of attributes determined in this way provided a basis for a complex characterization of the land property. A total of 25 attributes were selected for further analysis (Annex 1 – column 1).

b) *Quantification of land property attributes.* The process of selecting a method for data coding is directly interrelated with further data processing. The choice of a data quantification method is dependent on the measurement scale used for determining the variation of a given attribute. The use of an inappropriate scale may lead to the selection of an erroneous analytical method. As a result, the decision-making process concerning the significance of a given attribute, may also go in the wrong direction, or be too labourious. The application of a given measurement scale affects further transformations made on this scale (Cruse 2006b; Wiśniewski 2007, p. 141). Following the determination of measurement scales and transformations to be performed on these scales, a method for properties attribute coding was proposed (Annex 1 – column 3).

5.3. Elimination of outlying cases

Outlying observations, i.e. points which do not match the distribution pattern of the other data, are rare. They may reflect the actual properties of a given phenomenon or anomalies that should be disregarded in modeling. Outlying observations affect the slope of the regression line and, in consequence, the correlation coefficient. Even a single outlying observation may considerably change the slope of the regression line and the correlation coefficient. Outlying observations are believed to represent random errors which should be controlled. They may not only increase the value of the correlation coefficient, but also decrease the value of "real" correlation (StatSoft 2007).

Outlying cases should be considered untypical. Their occurrence results from the fact that market situations are characterized by certain randomness and that various disturbing factors are present in the market. From the perspective of population distribution, outlying cases should be treated as normal, within certain limits. The introduction of the term "hyper-outlying" (extreme) and the determination of a measure of outlying allow to systemize outlying and extreme cases. This procedure should be additionally based on elements of population representativeness (population classes).

It should be emphasized that neural network models are "resistant" to hyper-outlying cases. Data processing in ANN models enables to minimize

or even eliminate the effects of such cases (they vanish in the amount of data processed at a level of neurons). However, looking at the statistical parameters of estimation of the population distribution "normality", this procedure is burdened with an excessive load, which may be identified with error causing disturbances. Therefore, it seems that the simplest solution is to eliminate extreme values in order to reduce the risk of error.

At this stage of the study, outlying variables were selected using a combination of several methods: arithmetic means and standard deviations (for variables measured on a ratio scale), standardized residuals, eliminated residuals, Cook's distance, Mahalanobis' distance. According to the adopted criteria, four hyper-outlying cases were found for the investigated object, which accounted to 1.3% of the total number of cases (309).

In order to determine the effect of extreme case elimination, two regression models were analyzed: linear and non-linear exponential, in the form given by the formula in Masters (2005). A total of four models of multiple regression were investigated (two for 309 cases and two for 305 cases). Table 1 presents the results of this analysis.

Table 1

Effect of elimination of hyper-outlying cases

	309 cases		**305 cases**	
Model	linear	given by formula (2)	linear	given by formula (2)
R^2 * 100% [1]	81.66 %	93.20 %	45.38 %	49.96 %
AdjR^2	0.80	0.93	0.40	0.45
A [1]	75.47 %	54.48 %	46.94 %	47.36 %
% improvement	reference level – 100 %		27.81	13.07

[1] – *formulas (R^2 and A) are given in Annex 2.*

Source: own calculations

An improvement in the model adequacy was determined as follows: the level of model adequacy prior to the elimination of hyper-outlying cases was assumed as 100%, and then a decrease in the value of this index resulting from the elimination of hyper-outlying cases was calculated. In Table 1 R^2 * 100% decreased because of the elimination of hyper-outlying cases.

5.4. Selection of explanatory variables

Statistical analyses in the mass appraisal of real estates, based on many independent variables, involve the selection of the optimum subset of independent variables from among the examined attributes. The solution to this problem requires the elimination of unnecessary variables, whose absence does not decrease significantly the values of determination coefficients. This is done for practical reasons, since performing observations for a large number of variables in order to predict the value of the dependent variable is both expensive and time-consuming. Non-significant variables are also eliminated for theoretical reasons, because in the reduced model the estimators of regression coefficients are characterized by smaller mean squared errors (Fiedorowicz 1999). A subset of independent variables can be considered optimal if it allows to explain the variation of a dependent attribute to the same degree as the original model.

Due to a high number of parameters, neural networks are more sensitive to overfitting than other statistical methods. Overfitting can be prevented by applying a large number of learning cases. The size of the network affects the size of the learning set. When the number of learning samples is limited, we should refer to methods permitting the optimization of the number of input data (e.g. genetic algorithms). The elimination of a certain number of variables causes a decrease in the number of network parameters and, in consequence, in the number of observations. It is usually assumed that the number of learning cases should be two-fold bigger than the number of weights in the network. Masters (2005) recommends to double this number.

Variable selection was based on several methods, which allowed to find variables indispensable for learning and testing statistical models in the mass appraisal of real estates. The first two methods involved a stepwise regression analysis, the third method involved the optimization of a set of variables using genetic algorithms, while the fourth method involved polynomial selection. The polynomial selection method enabled to select significant variables within a set of variables, $N1, N1^2, N2, N2^2, \ldots, N25, N25^2$. The performed analysis allowed to reduce the number of property attributes (independent variables) to eight (Table 2).

Table 2

Types of variables and the ways of explanatory variable presentation

Real estate attribute	Symbol of variable	Variable	Type of variable	Way of presentation	Notes
1	**2**	**3**	**4**	**5**	**6**
Date of transaction	N1	Quantitative	Interval	1 neuron	Trans. NL [2]
Transport services	N2	Quantitative	Rational	1 neuron	Trans. NL
Distance to the city centre	N4	Quantitative	Rational	1 neuron	Trans. NL
Form of possession	N6	Qualitative	Interval	T [1]	-
Access	N12	Qualitative	Ordinal	T	-
Topography of the land parcel	N18	Qualitative	Ordinal	T	-
Area of the land parcel	N20	Quantitative	Rational	1 neuron	Trans. NL
Communications network	N24	Qualitative	Ordinal	T	-

[1] *T – presentation by the thermometric method (Masters 2005)*

[2] *Trans. NL – transformation by the natural logarithm function (i.e. Trans. N1L=ln(N1)).*

Source: own study

5.5. Preliminary analyses and the ways of variable presentation

The following analyses were performed for all variables selected at the previous stage of the study, prior to value estimation: characteristics of the variables, linearity, normality and correlations (Annex 3).

The following observations were made at this stage:

a) there were linear relationships between the explained variable and explanatory variables,

b) there were correlations between the explained variable and explanatory variables, and between the explanatory variables. It is difficult to say whether those relationships were of causal nature (the cause was reflected in the estimated outcome), or not.

c) there were significant differences between the explanatory variables, reflected in the coefficients of variation from 10% to 105%.

A given real estate attribute under analysis should be measured and coded in an appropriate way, and next it should be presented to the network (Table 2, column 5). Particular attention should be paid to the transformations which can make it easier for the network to process data and to eliminate elements related to the lack of data additivity. If data sets are additive, the network learns faster.

Quantitative variable showing no distribution symmetry, characterized by "big tails" or bimodality, were transformed using the natural logarithm function (Table 2, column 6). The models with variables subjected to logarithmic transformation provided more satisfactory results.

Qualitative ordinal variables were presented to the network by the thermometric method.

The set of observations was divided into three subsets, i.e. *learning* cases - $P_U = LP - (P_W + P_T)$, *testing* cases - $P_T = (LP - P_W) \times 20\%$ and *verifying* cases - $P_W = LP * 1\%$, LP – number of cases adopted for analysis (305).

5.6. Elimination of overfitting (network overtraining) and control over the learning process

Error control used a root mean squared (RMS) error histogram for the learning and testing sets – the admissible error was ± 20%.

The following measures were applied to select the optimum structure (Annex 2):

a) *Coefficient of determination* (R^2) – indicates the correlation between the predicted value and the observed value; this is the mean observed value,

b) *Mean-squared error (MSE)* – this is an absolute measure which shows the mean error of the estimated value of a real estate, compared with the observed value,

c) *Adequacy coefficient (A)* – indicates the mean relative error of value prediction (%),

d) *Coefficient of dispersion (COD)* – indicates the mean deviation (%) of the estimated value and the observed price ratio to the median of this ratio,

e) *Coefficient of variation (COV)* – unlike the coefficient of dispersion (COD), it is based on the mean ratio between the estimated value and the observed price.

5.7. ANN models – selected parameters

a) *Number of input neurons* – 8 explanatory variables.

b) *Selection of the number of neurons in the hidden layer of the network* – iterative, from the range ($1/2 * N : 2 * N$), where N is the number of input neurons. A network with one hidden layer containing 32 neurons was selected.

c) *Number of hidden layers* – 1 or 2. 1 hidden layer was selected.

d) *Number of output neurons* – 1 – property value.

e) *Method for weight initialization* – at random, in the range from –1 to 1.

f) *Activation function* – unipolar sigmoid.

g) *Scaling to the range of the activation function* – the variables were scaled to the range of 0.2 – 0.8 of the activation function.

h) *Network learning algorithms:* back-propagation of error with momentum (BP), conjugate gradient descent (CGD), Levenberg-Marquardt (LM). BP was selected (Masters 2005).

i) Based on the conducted experiments, a learning constant for the algorithm of error back-propagation was adopted at *a level of 0.4*. The following values were also analyzed 0.1, 0.2, 0.3, 0.5, 0.6, 0.7.

j) The momentum coefficient for the algorithm of error back-propagation was adopted at *a level of 0.3*.

k) *Number of epochs tested* – 10 000. Number of replications: 3.

l) Error function – a sum of the squares of differences between the actual network output and the set value.

5.8. Evaluation of results

All results obtained by each model tested in the experiment were recorded, focusing on the final sets of weights used by the networks to solve problems. The obtained parameters allow to use a given model for value prediction.

Among the three models tested, better results were obtained for the ANN model (Annex 4). For the investigated object, it was a *three-layer perceptron, with a hidden layer containing 32 neurons*, i.e. (according to Kolmogorov's theorem) theoretically the maximum number (Fig. 2). The network was taught by *error back propagation*.

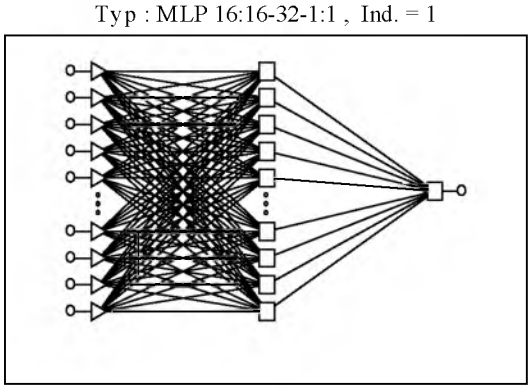

Figure 2. ANN configuration
Source: own study

The ANN models enabled to obtain the coefficient R^2 at a level of 67% for the learning set, and 74% for the testing set. The use of a considerable number of hidden neurons and a high level of variation explanation may indicate rather complicated relationships between the explained variable and explanatory variables. The ANN models provided better results than the regression models for the experimental object, i.e. the land market of the city of Olsztyn (Figures 3-6).

The adequacy coefficient (A) was 27% for both the learning and testing sets. This value was relatively high, which most probably resulted from the assumptions made with respect to case elimination, and from the absence of dataset segmentation. Similarly to the coefficient R^2, the adequacy coefficient was lower for the ANN models (approx. 28%) and higher for the multiple regression models (30% to 40%).

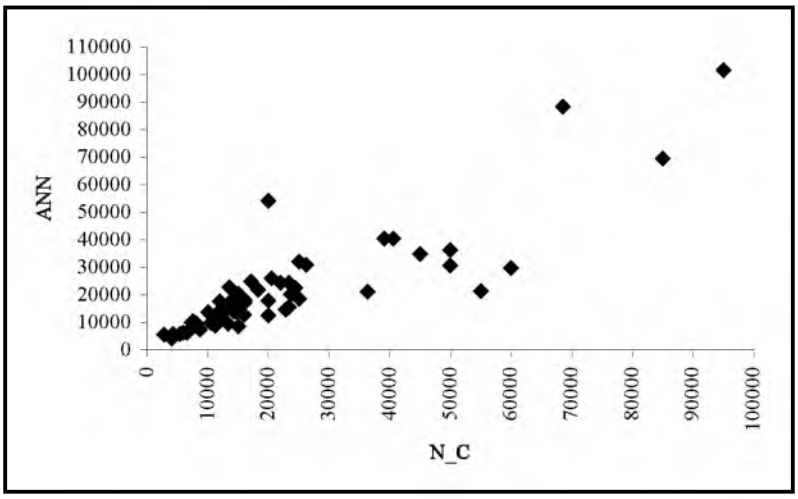

Figure. 3. Evaluation of ANN model fitting – Olsztyn
Source: own study

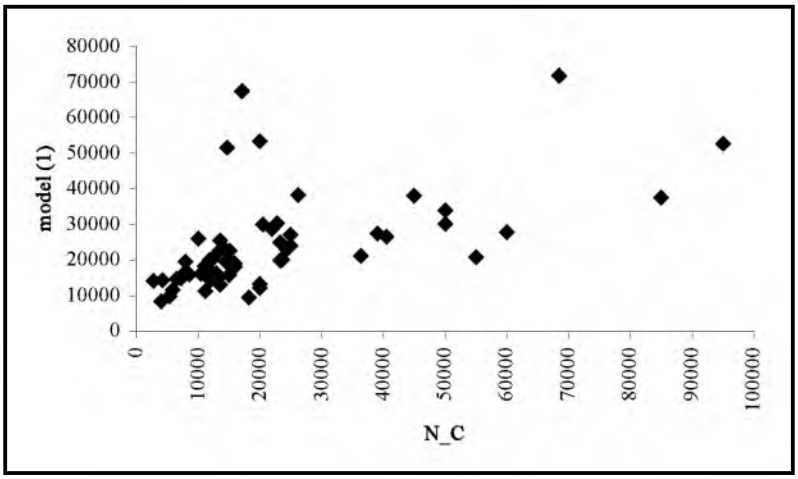

Figure. 4. Evaluation of model (1) fitting – Olsztyn
Source: own study

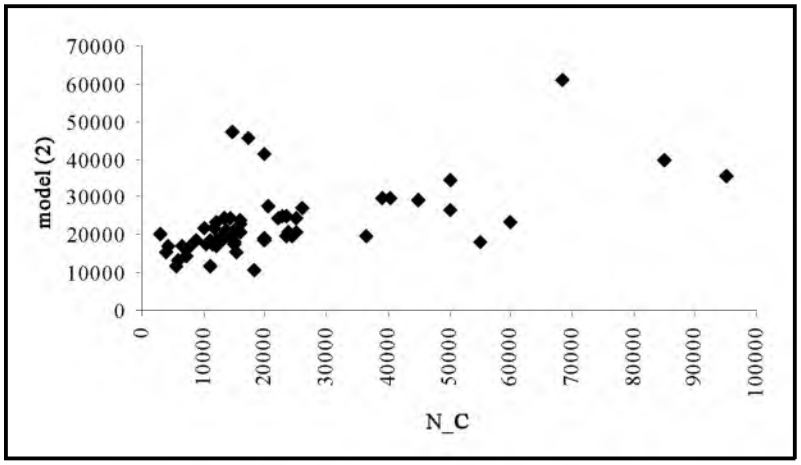

Figure. 5. Evaluation of model (2) fitting – Olsztyn

Source: Own study

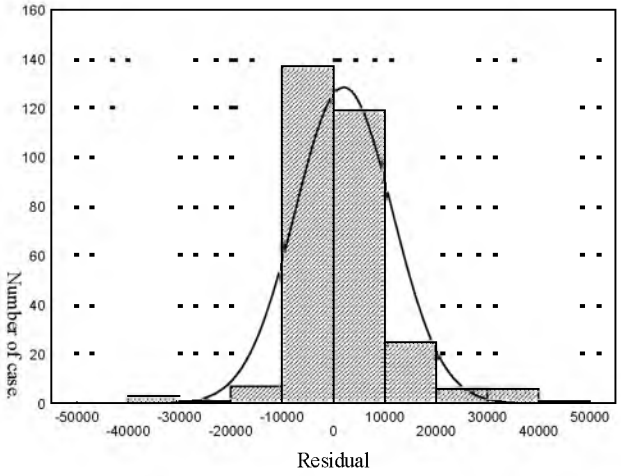

Figure. 6. Histogram of residuals (ANN model)

Source: own study

The mean squared error was approximately 9 600 for the city of Olsztyn, and the average price reached PLN 23 000. COD was at a similar level as

coefficient A. COV was higher (theoretically by 25%) than COD, reaching 45% for the learning set and 38% for the testing set.

The adequacy values calculated for the verifying cases differed from the mean values obtained for the learning and testing sets. An extreme value – 760% – was recorded for case 3. It was caused by the random selection of cases and by the fact that this case was characterized by one of the lowest prices – PLN 4 688.

The degree of learning set fitting was evaluated based on the histogram of absolute error (Fig. 7). It was found that 79 cases, i.e. 32.6% of 142 cases, remained within the range of error ± 10%.

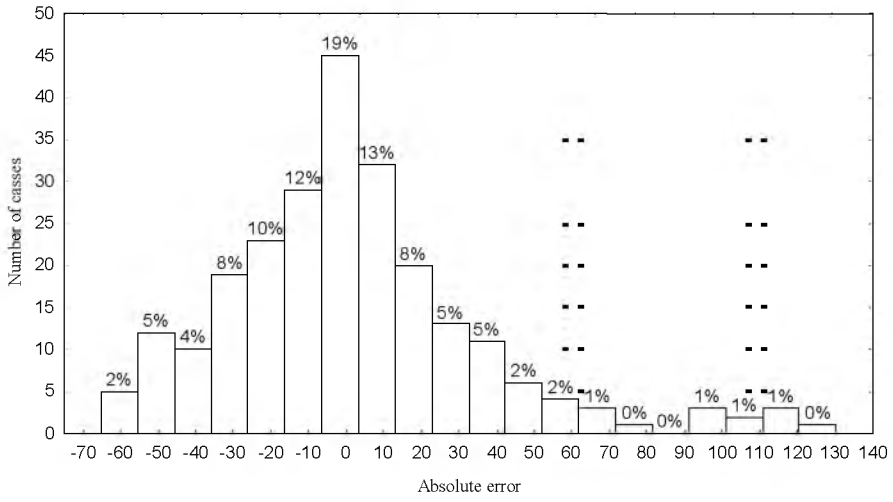

Figure. 7. Histogram of the absolute error of prediction for ANN – Olsztyn

Source: own study

6. SUMMARY AND CONCLUSIONS

The mass appraisal of real estates is governed by specific processes, including value creation. Taxation procedures should meet certain legal requirements, but their determination still arouses controversy. According to Polish legal regulations, the value of land should be determined taking into account zones distinguished due to similar factors affecting the market value of a given real estate. In the case of mass appraisal, the division into zones

allows to unify the effects of particular factors influencing the value of a property in a given zone. Therefore, a zone is a part of the appraised property, whose value is affected by the same factors.

Models of artificial neural networks may be used for zone creation and value determination in a given zone, because their weight matrixes permit value prediction. A "general model" for the investigated object, in the form of *weight matrixes*, reflects (with the use of previously selected attributes) the *value* of a given type of real estates in a given area. This allows to distinguish *zones* in which real estates are characterized by a similar value, fulfilling the condition of the *homogeneity of value-creating factors*, in this case expressed as the *value function*. The result of area division into zones is a *similar value* of real estates, and not a similar or identical way of value calculation. This general model may be also used for value determination within zones. This can be done in two ways. For smaller, homogenous objects, where the average number of transactions regarding real estate of the same type is 50 to 150 (within a time interval of up to 24 months), the general model can perform the role of a "zone model". In such a case the number of homogenous zones would be probably small, and the results obtained with the general model – satisfactory. For larger objects, differing in value-creating factors, with the average number of transactions higher than 150, zone models should be developed for particular zones. Zone models should be based on general models, typical of the entire object. However, the specific character of particular zones should be taken into account.

Models of artificial neural networks hold a special position, because the network architecture subject to the learning process, leading to the determination of certain values of weights, can be further taught. This great advantage of ANN models can be used for developing zone models. A general model should be further taught, using transactions concluded in a given zone. In this way neuronal connection weights can be adjusted to the specific character of the value-creating factors in this zone, and the model can reliably predict the value of real estates. From the practical perspective, zone models fulfill the conditions that must be satisfied by models used for property value estimation.

Many of the above problems can be solved using neural network models. Those structures are function approximators, so they can easily find hidden relationships and interdependences. ANNs accept non-linearity, are able to ignore random disturbances and generalize the discovered relationships. For these reasons they can be used by value appraisers for the mass appraisal of

real estates. ANNs may be employed at he following stages of this procedure:

a) Selection of significant (value-affecting) variables.
b) Modeling the way of variable presentation – analysis of sensitivity.
c) Selection of comparative or representative real estates.
d) Creation of taxation zones.
e) Development of taxation models.
f) Development of models of land unit value.
g) Introduction of corrections due to differences in the attributes of real estate.
h) Modeling of unit values at the points of contact between zones representing uniform values.

Furthermore, ANNs allow to:

i) find and eliminate outlaying and non-representative cases.
j) determine parameters related to the comparability of real estates.
k) carry out, in a quick and accurate way, the preliminary appraisal of real estates, e.g. to fix tax rates.

The above examples show that ANN models can be successfully used for the purpose of mass appraisal. However, their application requires looking at the legal and computational aspects of this procedure from a new perspective. This includes a different approach to the definition of representative real estates, or significant value-creating factors. All transactions observed in the land market should be considered in analyses, provided that they are homogenous. If necessary, it is also possible to valuate additional cases.

Statistical analyses enable to reduce the time needed for preliminary analyses, preceding the implementation of the adopted solutions. This is possible due to the use of software tools and statistical procedures, allowing to eliminate insignificant variables, recognize outlaying observations, employ appropriate methods for attribute quantification, verify the results of valuation, etc. Statistical methods are economical, which does not mean that the results obtained by those methods are less accurate. They allow to eliminate insignificant variables (whose acquisition would increase the costs of the entire process), maintaining the required adequacy and quality. From the technical perspective, data collection in modern integrated systems of information on real estates precedes the application of statistical methods, which constitutes the next step in the process of creating the outcome, i.e. the value.

REFERENCES

Cruse, H., *Neural Networks as Cybernetic Systems - Part I* (2nd and revised edition). Brains, Minds and Media, Vol. 2, 2006a.

Cruse, H., *Neural Networks as Cybernetic Systems - Part II* (2nd and revised edition). Brains, Minds and Media, Vol. 2, 2006b.

Fiedorowicz, J., *Laboratorium obliczeniowe zastosowań matematyki* [Computing platform of mathematics - User's guide]. Katedra Zastosowań Matematyki. ART. – Olsztyn, 1999.

Kauko, T., *Advances in mass appraisal methods – an international perspective*. ENHR International Conference. Rotterdam, 2007.

Masters, T., *Practical Neural Network Recipes in C++* (Paperback). Morgan Kaufmann; Book & Disk, 2005.

Nguyen, N., Cripps, A., *Predicting Housing Value: Comparison of Multiple Regression Analysis and Artificial Neural Networks*. "Journal of Real Estate Research", 22, 3, pp. 313–336, 2001.

Rutkowski, L., *Metody i techniki sztucznej inteligencji. Inteligencja obliczeniowa* [Methods and techniques of artificial intelligence. Computing intelligence]. Wydawnictwo Naukowe PWN, Warszawa, 2005.

StatSoft, Inc. *STATISTICA for Windows [Computer program manual]*. Tulsa, OK: StatSoft, Inc., 2007.

Wiśniewski, R., *Wielowymiarowe prognozowanie wartości nieruchomości* [Multidimensional forecasting of the real estate value]. Wydawnictwo UWM w Olsztynie, 2007.

Received: January 2008, revised: March 2008

ANNEX 1

Explanatory variables adopted for analysis and their quantification

Explanatory variable	**Symbol of variable**	**Quantification of variable**
1	2	3
Date of transaction	*N1*	Number of week – [0 – 128]
Transport services	*N2*	Crow-fly distance from the properties to the communication centre (transport junction) - [km]
Distance to the shopping-and-service centre	*N3*	Crow-fly distance from the properties to the shopping-and-service center – [km]
Distance to the city centre	*N4*	Crow-fly distance from the properties to the city centre – [km]
Land function in the local spatial management plan	*N5*	1.5 – single-family housing „ + ” (e.g. services causing no nuisance), 1.0 – single-family housing, 0.5 – single-family housing „ - ” (e.g. land management constraints).
Form of ownership	*N6*	1.0 – right to property, title, 0.6 – right to perpetual usufruct.
Form of transaction	*N7*	1.0 – concluded in the secondary market 0.5 – concluded in the primary market
Shape of the land parcel	*N8*	1.0 – regular (rectangular), 0.5 – irregular.
Frontage of the land parcel	*N9*	The length of the parcel boundary line adjacent to the road – [m].
Depth of the land parcel	*N10*	The length of a perpendicular (from the geometric centre of the frontage) connecting the parcel frontage with the line constituting the opposite boundary – [m].
Location of the land parcel	*N11*	0.5 – at the corner (e.g. at the cross-roads), 1.0 – other.
Access	*N12*	1.0 – poor (unsurfaced road, no bus service /no parking space), 2.0 – difficult (good-surfaced road, no bus service /parking space at the road), 3.0 – normal (hard-surfaced road, bus service / lay-byareas), 4.0 – good (arterial road, bus service / special parking places, e.g. in the courtyard), 5.0 – very good (thoroughfare, bus service / attended car-park).

Attractiveness of the parcel location, taking into account the vicinity of forest complexes and parks	*N13*	Crow-fly distance from the properties to the nearest forest complex – [km].
Attractiveness of the parcel location, taking into account the vicinity of recreation grounds and water bodies	*N14*	Crow-fly distance from the properties to the nearest water body – [km].
Neighbourhood nuisance: – motor roads	*N15*	1.0 – no nuisance, 2.0 – low degree of nuisance (e.g. location within a long distance from a service workshop), 3.0 – average degree of nuisance (e.g. location within a short distance from a cross-roads), 4.0 – high degree of nuisance (e.g. location within a short distance from a communication line), 5.0 – very high degree of nuisance (e.g. location within a short distance from a thoroughfare, main railway line, etc.),
Neighbourhood nuisance: – railway lines	*N16*	
Neighbourhood nuisance: - other (e.g. industrial areas)	*N17*	
Topography of the land parcel	*N18*	5.0 – flat ground (no falls or slopes) 4.0 – no slopes, undulated ground, 3.0 – sloping ground (18%), no undulation, 2.0 – sloping (18%), undulated ground, 1.0 – sloping (> 27%), undulated ground.
Number of land parcels	*N19*	Number of parcels being the object of one transaction.
Area of the land parcel	*N20*	[m^2]
Water-pipe network	*N21*	0.5 – no service connections or utilities 1.0 – service connections and utilities can be provided, 1.5 – service connections and utilities already provided.
Power network	*N22*	
Gas grid	*N23*	
Communications network	*N24*	
Sewerage system	*N25*	
VALUE OF THE REAL ESTATE	*N_C*	PLN

Source: own study

ANNEX 2

Measures of artificial intelligence model evaluation

No	Definition	Formula	Symbol
1	2	3	4
1	*Coefficient of determination* (R^2) indicates the correlation between the predicted values $(\hat{y}_i)$ and the observed (y_i), $(\overline{y}_i)$ values; this is the mean observed value	$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$	X.1
2	*Mean-squared error (MSE)* – this is an absolute measure which indicates the mean error of the estimated value of a real estate, as compared with the observed value	$SSq = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$	X.2
3	*Adequacy coefficient (A)* – indicates the mean relative error of value prediction (%)	$A = \frac{\sum_{i=1}^{n}\sqrt{\left(\frac{y_i - \hat{y}_i}{\hat{y}_i}\right)^2}}{n} \times 100\%$	X.3
4	*Coefficient of dispersion (COD)* – indicates the mean deviation (%) of the estimated value and the observed price ratio to the median of this ratio	$COD = \frac{\sum_{i=1}^{n}\left\lvert \frac{\hat{y}_i}{y_i} - \left(\frac{\hat{y}_i}{y}\right)_{median}\right\rvert}{n-1} \Big/ \left(\frac{\hat{y}_i}{y}\right)_{median} \times 100\%$	X.4
5	*Coefficient of variation (COV)* – unlike the coefficient of dispersion (COD), it is based on the mean ratio between the estimated value and the observed price [4]	$COV = \sqrt{\frac{\sum_{i=1}^{n}\left(\frac{\hat{y}_i}{y_i} - \frac{\overline{\hat{y}_i}}{y}\right)^2}{n-1}} \Big/ \frac{\overline{\hat{y}_i}}{y} \times 100\%$	X.5
Symbols: predicted value – $(\hat{y}_i)$; observed value – (y_i); mean observed value – $(\overline{y}_i)$; number of cases – (n)			

Source: own study

The optimum artificial intelligence model should be characterized by the maximization of the coefficient R^2 and the minimization of the other measures. Particular attention should be paid to the minimization of the coefficient A, which provides similar results as COD, but is calculated in a different way. The former is based on the estimated value and the observed price, whereas the latter – on the ratio $(\hat{y} / y)$ of those values.

ANNEX 3

Descriptive characteristics [denotation of variables as in Annex 1]

Number of observations n = 305 Number of variables g = 9

Significance level α = 0.050

EVALUATION OF PARAMETERS

EVALUATION OF CORRELATION MATRIXES

i \ j	N1	N2	N4	N6	N12	N18	N20	N24
N2	0.556*							
N4	0.445*	0.413*						
N6	-0.062	-0.113	-0.184					
N12	-0.062	-0.113	-0.184*	-0.060				
N18	0.218*	0.246*	0.144	-0.094	0.397*			
N20	-0.074	-0.051	0.044	0.165	-0.222*	-0.283*		
N24	0.188*	0.109	-0.074	-0.254*	0.403*	0.262*	-0.126	
N_C	0.057	-0.049	-0.111	0.0917	-0.081	0.109	0.418*	0.166

* - significant correlation

LIKELIHOOD RATIO TEST

Calculated value of the Q test = 1373.521*

Critical value Q alpha = 50.990

---------- *the hypothesis should be rejected*

Multidimensional skewness coefficient b1 = 46.956 (hypothetical = 0)
Multidimensional flatness coefficient b2 = 127.425 (hypothetical = 80)

MULTIDIMENSIONAL NORMALITY TEST

TEST FUNCTION	CRITICAL VALUE
For skewness: A = 2386.928*	146.567 (degrees of freedom = 120)
For flatness: B = 32.739*	1.960

---------- *the hypothesis should be rejected*

EVALUATION OF PARAMETERS

Symbol of variable	Min	Max	Mean	Median	Modal value	SD
N1	1	128	47.3	39	10.1	35.1
N2	3.49	7.42	4.65	4.23	4.18	0.93
N4	2.310	6.578	3.950	3.700	3.639	0.820
N6	0.6	1.0	0.9	1.0	1.0	0.1
N12	1	4	2.3	2	2	0.8
N18	0	5	3.9	4	5	1.3
N20	172.00	10841.00	1077.99	818.00	496.85	1100.02
N24	0.5	1.5	0.7	0.5	0.6	0.4
N C	1121.40	95070.00	22668.87	18000	16523.63	17200.74

RV – range of variation

SD – standard deviation

ANNEX 4

Results for ANN and regression models

	No of	Learning							Testing							Verifying		
		R^2	A			SSQ	COD	COV	R^2	A			SSQ	COD	COV	A		
			Av*	MIN	MAX					Av*	MIN	MAX				1	2	3
MLP 32	16	**0.67**	**28**	**0.1**	**187.0**	9632.8	27.2	45.8	**0.74**	**27**	**0.1**	**156.5**	**9627.4**	**24.7**	**37.7**	15.9	103.6	760.7
RP	8	0.46	38	0.5	193.7	12302.1	51.4	72.3	0.31	42	**0.1**	164.6	15669.3	46.1	78.7	36.3	**58.3**	**16.2**
RW	8	0.39	43	1.0	209.9	13099.5	61.9	99.8	0.31	46	1.8	205.5	15670.8	48.8	94.5	25.60	89.7	72.3

* Average