ISSN 1392 – 124X INFORMATION TECHNOLOGY AND CONTROL, 2006, Vol. 35, No. 2

# TRACKING OF DOUBTFUL REAL ESTATE TRANSACTIONS BY OUTLIER DETECTION METHODS: A COMPARATIVE STUDY

**Vilius Kontrimas, Antanas Verikas**

*Department of Applied Electronics, Kaunas University of Technology*
*Studentu St. 50, LT-51368 Kaunas, Lithuania*

**Abstract**. Doubtful real estate transactions, with the prices far away from the market prices, appear because of non commercial transactions or efforts in order to hide the taxes. To estimate the right values of parameters, such data must be removed from a data set or robust methods of parameters estimation are to be used, while developing a mass appraisal model. Such transactions are outlying observations, which can be detected and removed by outlier detection methods. The purpose of the work is to review outlier detection methods and to test the possibility of using them to solve the task. An overview of real estate market value, valuation methods and process of mass appraisal is made to introduce to real estate mass valuation. Overview of outlier detection method contains scaling and such methods: resampling by half means, the smallest half volume, the closest distance to the center, ellipsoidal multivariate trimming, minimum volume ellipsoid, minimum scatter determinant, analysis of projection matrix, principal components and residuals, also influence measures, robust regression, and classification methods. The reviewed methods were categorized; commonly used methods were selected and tested experimentally aiming to compare the effectiveness. Best results were achieved using the multilayer perceptron and the principal component analysis based technique.

## 1. Introduction

It is very important to use correct data when developing real estate mass valuation methods. Especially it is important in Lithuania and other countries that recently switched from planned economy to the market economy. Citizens' wealth accounting system does not function well; the proportion of black economy is high. So in efforts to hide real money gains and avoid taxes, the transactions are recorded with the price much lower then the market price. Obviously fictive transactions, having the value even lower then the registration cost, can be easily detected. Much more difficult is to detect transactions, where the value is lowered only to some extent or the value is fictious because of noncommercial nature of the transaction.

The research purpose is to review outlier detection methods, which can let detecting doubtful real estate transactions and to experimentally test their effectiveness.

Section 2 introduces to the definition of real estate market value, valuation methods, and mass appraisal of real estate. Section 3 presents overview of outlier detection methods. Experiments are described in the Section 4 followed by the Conclusions section.

## 2. Real estate mass appraisal

### 2.1. Real estate market value

The purpose of mass appraisal as individual appraisal is to determine the market value. It is very important to make the distinction between the market price and the value. The market price is formed, when curves of supply and demand intersect, it is influenced by many objective and subjective factors. The market price equals to the market value very rarely, because the market of real estate is not an ideal market. The market price of real estate reflects many subjective factors, so a real estate assessor must find the most objective, suitable for all value.

According to the Lithuanian Republic normative documents, market value is estimated money amount, for which property can be exchanged on valuation date between a willing buyer and a willing seller in arm's-length transaction after proper marketing, wherein the parties act knowledgeably, without compulsion and impact of other transactions and interests.

In international valuations standards 2005 (IVS), issued by International Valuation Standards Committee (IVSC), the market value is defined as the estimated amount of money for which a property should exchange on the date of valuation between a willing buyer and a willing seller in arm's-length transaction after proper marketing wherein the parties acted knowledgeably, prudently and without compulsion [22].

Nine non market values are defined in the international valuations standards. They are: value in use; investment value; going concern value; insurable value; assessed, rateable or taxable value; salvage

value; liquidation or forced sale value; special value; mortgage lending value [22].

Market value is most important and commonly used for the real estate valuation.

### 2.2. Market value valuation methods

There are three traditional real estate valuation methods: the sales comparison approach, income approach, and the cost approach [48].

In the case of the sales comparison method, value is determined comparing the subject with the other objects sold in the market. The value is adjusted according to differences, as real estate objects have differences. A difference up to 30-35 percent between exact object characteristics is acceptable. This method is very suitable for clear land. Reflections of the market price, quick and simple computations are the main advantages of this approach.

The income approach is based on the premise that the value is the present worth of future; the value is determined by discounting cache flows, generated by the object. It is very suitable for the objects that give incomes, for example, building with leased offices or flats, objects used for services, production. This approach is quite simple too and estimates the economic benefit from the object.

Value of the object is determined by construction costs minus depreciation in case of the cost approach. This approach can be applied only to buildings, and it is very suitable for such objects as schools, objects of engineering infrastructure and similar, which do not generate incomes and there are only a few objects to compare with. This method is often used to estimate the value of improvements.

### 2.3. Process of the real estate mass appraisal

According to the Lithuanian Republic law of property and business valuation basics, mass appraisal is such an appraisal method, when value of exact property is not determined, but ranges of value, covering value of the property being assessed are set by the analysis of collected information about that property [32]. Data are collected, analyzed and computations are made in a systematic approach. This valuation method is applied to property objects with many similarities. The individual appraisal is such an appraisal, when value of the exact object is determined according to all its individual characteristics.

Thus, mass valuation is a systematic valuation of a property objects group as of the given date, using standardized procedures and statistical methods. Individual valuation is designated for valuation of one object. Mass and individual valuations differ in market analysis and quality control, but have the same appraisal steps and principles such as: supply and demand, the highest and the best use, expectations, balance, changes, competition, integration, replacement, over profit, marginal utility.

Determining the variables influencing the value and their inclusion form into the model, are the most important issues to consider while building a valuation model. Different models are constructed for different valuation approaches (sales comparison, income, cost). Values of model parameters must be determined when the model type is chosen. Statistical software allowing to perform the statistical analysis automatically, is commonly used for that purpose. Model testing and assessing are the last step. All these steps are closely related and the process of model creation is recursive.

The developed model may not fit some specific real estate objects, therefore, there may be a need to develop an additional model. To assess the model quality, the coefficients of variation are computed for homogeneous groups of properties. Depending on the type of property, the coefficients must not exceed 10-30 percent. To check the modeling accuracy, the modeling results are compared with the actual prices. Models must be specified and calibrated again, if the differences exceed the allowable range, until the proper level of precision is reached. To reach the desirable level of precision, data must be correct and consistent.

## 3. Outlier detection methods

Observations inconsistent with majority are called outliers. They are “suspicious” data points [47]. Commonly, outliers have high influence onto model parameters, therefore, such data points must be detected and removed. However, it is very important to distinguish between outliers and high influence points, which are inconsistent with majority too, but they are important for correct estimation of model parameters. Some methods allow distinguishing between these two types of data, while most of techniques detect only outliers. Outlier detection methods can be categorized into four large groups:

1. Methods based on distance from a data center. They include techniques detecting observations outlying from majority of the data, techniques analysing the projection matrix, the principal component analysis based techniques [4, 7, 30, 47, 52].
2. Methods based on the difference between the predicted and actual values of a dependent variable. Residuals, graphical analysis of residuals, influence measures [3, 10, 15, 25, 26, 37, 45] constitute the group.
3. Robust regression – robust estimators are used instead of the common least square estimator. The least absolute deviation (LAD), M, the least trimmed squares (LMS), the least median of squares (LMS), S, minimum M (MM) estimators are the most known representatives of the group [ 6, 12, 17, 19, 29, 39, 52].

4. Classification methods. A classifier separates data into two classes: outliers and normal data [2, 5, 11, 24, 28, 43].

The robust regression reduces the influences of outliers onto regression parameters, other methods detect outliers.

### 3.1. Scaling

Scaling is used to increase effectiveness of outlier detection methods. Besides the increased effectiveness, scaling provides information about location of a data point in a data set [30]. There are two types of scaling: auto scaling and robust scaling. The mean $\bar{x}$ and standard deviation $s$ are used in auto scaling:

$$z_i = \frac{x_i - \bar{x}}{s} \quad (1)$$

If data are normally distributed, then [46]:

- approximately 68 percent of $z$ values are in the range (-1; 1);
- approximately 95 percent of z values are in the range (-2; 2);
- approximately 99 percent of z values are in the range (-3; 3);

In the case of non normally distributed data, the Chebyshev [42] rule is valid:

- at least 75 percent of z values are in the range $(\bar{x}-2s, \bar{x}+2s)$;
- at least 88 percent of z values are in the range $(\bar{x}-3s, \bar{x}+3s)$;

There are enough of these rules to detect and remove outliers, if there are only a few percent of outliers in the data of low variance and they are far from the data center.

Auto scaling is sensitive to outliers because of their influence on the mean and standard deviation, therefore, robust scaling is used.

Huber suggested a robust scaling method, where the median instead of the mean and the median of the standard absolute deviation from the median instead of the standard deviation are used [30]:

$$S_{MAD} = 1.4826 \mathrm{med}_i(|x_i - \mathrm{med}_j(\mathbf{x}_j)|), \quad (2)$$

where $\mathrm{med}_j(\mathbf{x}_j)$ (internal median) is the median of the j-th parameter, and the external median is the median of internal medians. Coefficient 1.4826 is required to make $S_{MAD}$ an unbiased estimate of the standard deviation for normally distributed data.

This scaling is effective even when data have 50 percent of outliers, but it takes a symmetric view on the variance, which may be ineffective for an asymmetric distribution. $S_n$ and $Q_n$ are two other estimates instead of $S_{MAD}$, which are more suitable for asymmetric distributions.

$$S_n = 1.1926 med_i(med_j(|x_i - x_j|)), \quad (3)$$

where the internal median $med_j(|x_i - x_j|)$ is the median of absolute pair-wise differences $|x_i - x_j|$, $j = 1,...n$, where $n$ is the number of observations. The external median is the median of internal medians, coefficient 1.1926 is required to make $S_n$ an unbiased estimate of the standard deviation for normally distributed data.

$S_n$ is an estimate of a typical distance between two observations and is effective for asymmetric distributions. However, $Q_n$ is even more effective:

$$Q_n = 2.2219\{|x_i - x_j|, i < j\}_{(k)}, \quad (4)$$

where $k$ is equal to $\binom{h}{2} = h(h\text{-}1)/2$, when $h=[n/2]+1$, where [ ] denotes the integer part. $k$ is approximately equal to $\binom{n}{2}/4$. Thus, $Q_n$ is the $k$-th order statistics of $\binom{h}{2}$ pair-wise data point differences. Coefficient 2.2219 is required to make $Q_n$ an unbiased estimate of the standard deviation for normally distributed data.

Scaling is recommended for all outlier detection methods based on distance from the data center. Chang et al [30] suggested using modified robust scaling, which is as robust to outliers as scaling with $S_n$ or $Q_n$, but a more accurate estimate of the standard deviation is obtained. First, the differences between each observation and their median are computed:

$$y_i = |x_i - x_{median}|, \quad (5)$$

and sorted in an ascending order. Then the standard deviation is computed using only a half of the smallest differences:

$$d_j = \sqrt{\frac{\sum_{i=1}^{(j+n)/2-1} (y_i - y_{mean})^2}{(j+n)/2-2}}. \quad (6)$$

When $j$ increases, the standard deviation $d_j$ increases too. While outliers are not reached, $d_j$ increases gradually. Rapid increase of $d_j$ is expected when outliers are included into the calculation. The variable $r_j$ can be used to avoid the graphical analysis:

$$r_j = \frac{d_{j+1}}{d_j}, \quad (7)$$

where the first rapid increase shows the beginning of outliers. Authors suggest using the fourth order standard deviation, which is even more sensitive to outliers, thus $d_j$ and $r_j$ increase even more rapidly. Having identified the normal data, the mean and

standard deviation are computed and used for scaling and outlier detection.

### 3.2. Methods based on the distance from the data center

It is very important to protect sample estimates from outliers influence. Below is the list of methods allowing to select observations, consistent with the majority. Sample estimates are then computed based on the selected data:

1. Resampling by half means (RHM).
2. The smallest half volume (SHV).
3. The closest distance to the center (CDC).
4. Ellipsoidal multivariate trimming (EMVT).
5. Minimum volume ellipsoid (MVE).
6. Minimum scatter determinant (MSD).

In order to start RHM, a sample of size *n/2* is randomly selected, where *n* is the number of observations, the mean and standard deviation of the sample are calculated and the data are scaled. Five percent of the data with the highest scaled values are then selected and marked as suspicious points. This procedure is repeated at least *2n* times. Then, the standard deviation and mean are calculated using the non marked points only. These estimates are used for scaling and outlier detection.

The data matrix **X** is auto-scaled and the pair-wise Euclidian distance is computed between all the observations when using SHV. Next, a $n \times n$ matrix of distances **D** is formed. Each column of **D** is sorted in an ascending order and the column of the smallest sum for the first *n*/2 distances is determined. These are the *n*/2 observations, closest to each other in the multivariate space. Outliers are inconsistent with the majority, so they stay in the other *n*/2 part. Estimates of the standard deviation and mean are calculated from the selected data and used for scaling and outlier detection, based on selecting $z_i$ values higher than the critical ones.

In CDC, outliers are selected according to the Euclidian distance between the observations and the mean. First, the data matrix **X** is scaled, then the Euclidian distance from each of the observations to the mean is calculated. Then, n/2 observations with the smallest distance to the mean are selected [30]. Estimates of the standard deviation and mean are calculated from the selected data and used for scaling and outlier detection.

In ellipsoidal multivariate trimming, the Mahalanobis distance is calculated for each observation:

$$d_m = (\mathbf{x} - \bar{\mathbf{x}})^T \times \mathbf{COV}^{-1} \times (\mathbf{x} - \bar{\mathbf{x}}), \quad (8)$$

where $\bar{\mathbf{x}}$ is the vector of means, and **COV** is the covariance matrix. At the beginning, these estimates are calculated using the whole sample. Then, *n*/2 observations with the smallest Mahalonobis distances are selected and new $\bar{\mathbf{x}}$ and **COV** are calculated. The Mahalonobis distance is then recalculated using the new estimates of the mean and covariance. The process is iterated until $\bar{\mathbf{x}}$ and **COV** stabilize. Estimates of the standard deviation and mean are then calculated from the remaining data and used for scaling and outlier detection.

MVE is based on seeking for the ellipsoid with the smallest volume, including at least h points of the data set. First, a sub-sample **K** of *m+1* observations is drawn from the data set **X,** where *m* is the dimensionality of **x**. Then, the sub-sample mean $\bar{\mathbf{x}}_K$ and the covariance matrix $\mathbf{COV}_K$ are computed. The parameter $?_K$ is used to inflate the ellipsoid size, thus the size of the ellipsoid is proportional to $|\mathbf{COV}_K|^{1/2} (l_K)^m$. The sampling process is iterated and results into the estimates of the MVE parameters $\bar{\mathbf{x}} = \bar{\mathbf{x}}_K$, and $\mathbf{COV} = \frac{l_K^2 \mathbf{COV}_K}{c_{n,0.5}^2}$, where $c_{n,0.5}^2$ adjusts the final covariance estimate to include all the good data points for the case of normally distributed data.

MCD is based on seeking for $h>n/2$ observations, with the smallest covariance matrix determinant. The mean and covariance matrix of these observations are used as the robust mean and covariance matrix estimates of the sample.

All types of scaling can be applied using these methods, scaling is called robust, when $S_{MAD}$, $S_n$, $Q_n$, $d_j$ or mean $\bar{x}$, obtained by MVE or MCD are used.

Influence observations, outlying from the majority in the **X** space, can be detected by employing analysis of the projection matrix **H** [36]. Distance from the data center is closely related to the values of the main diagonal elements of the projection matrix **H**. These values reflect the relative distance from the center due to the form of the data distribution. For example, if data points make an ellipse and $x_1$ is at the same distance from the centre as $x_2$, then the $H_{11}$ value calculated for $x_1$ is higher than the $H_{22}$ value of $x_2$, if $x_1$ is on a shorter diagonal of the ellipse. A $H_{ii}$ value can be found without forming the **H** matrix:

$$H_{ii} = \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i, \quad (9)$$

where $\mathbf{x}_i$ is the *i-th* observation or the *i-th* row of the matrix **X**. A $H_{ii}$ range is $1/c \geq H_{ii} \geq 1/n$, where *c* is the number of coincided observations. Influent points are those that have the value of $H_{ii} > 2p/n$, where *p*:

$$p = \sum_{i=1}^{n} H_{ii}. \quad (10)$$

The main task of the principal component analysis (PCA) is to reduce the data dimensionality. Besides that, PCA can also be used to detect outliers [7].

In the matrix form, the principal component analysis can be expressed as:

$$\mathbf{X} = \mathbf{T}\mathbf{Q}^T + \mathbf{E}, \quad (11)$$

where **X** is the $n \times m$ centered or scaled data matrix, **T** is the $n \times a$ matrix of scores, where $a$ is the reduced number of dimensions, $\mathbf{Q}^T$ is the $a \times n$ loading matrix, and **E** is the $n \times m$ matrix of residuals. The transformed data are stored in the matrix of scores with columns orthogonal to each other. The contribution of the variables to the scores is seen in the loading matrix. Because the scores are normally distributed, the Student t-test can be applied. For that purpose, the so-called $T^2$ values are computed according to:

$$T_i^2 = \sum_{z=1}^{a} \frac{t_{iz}^2}{s_z^2} . \quad (12)$$

where $t_i$ is the $i$th row of the matrix of scores and $s_z^2$ is the variance. The random variable

$$T_i^2 \times n(n-a)/a(n^2-1) , \quad (13)$$

is $F$-distributed with $a$ and $a$-$n$ degrees of freedom. Thus, outliers can be detected using the F test:

$$T_i^2 > a(n^2-1)/n(n-a) \times F_{(a)}(a, n-a) . \quad (14)$$

### 3.3. Methods based on the difference between predicted and actual values of a dependent variable

Outlier detection methods of the second large group are based on the difference between predicted and actual values of a dependent variable. Commonly, the ordinary least squares technique is used to estimate regression parameters used for prediction [36]. The methods can be divided into the following subgroups:

1. Methods, based on residual analysis:
   a. Residual cut-offs,
   b. Residual plots.
2. Methods, based on influence measures.

#### 3.3.1. Residual analysis

The true error e is a normally and independently distributed random variable with the mean $\mu$=0 and variance N(0,$\mathbf{I}d^2$). The residual $e$ is the estimate of $e$, which shows the difference between the actual value $y_i$ and the regression result $\hat{y}_i$

$$e_i = y_i - \hat{y}_i . \quad (15)$$

The residual vector **e** can be calculated using the residual projection matrix **M** (this matrix equals to the difference between the identity matrix **I** and the projection matrix **H**) and the **y** vector [27]:

$$\mathbf{e} = \mathbf{My} = (\mathbf{I} - \mathbf{H})\mathbf{y} . \quad (16)$$

When used to find outliers, the residuals are scaled:

$$e_{Ni} = \frac{e_i}{s} , \quad (17)$$

where $s$ is the standard deviation of the residuals. The $3s$ rule is usually applied to the scaled values. But internally studentized residuals are more sensitive to outliers:

$$e_{Si} = \frac{e_i}{s \times \sqrt{1 - H_{ii}}} , \quad (18)$$

where $H_{ii}$ is the diagonal element of the projection matrix **H**. However, more often the externally studentized or jackknife residuals are used:

$$e_{Ji} = e_{Si} \sqrt{\frac{(n-m-1)}{(n-m-e_{Si}^2)}} , \quad \mathbf{(19)}$$

where ($n$-$m$-$1$) is the degree of freedom of the Student's $t$-distribution, $n$ is the number of observations and $m$ is the number of variables. These residuals can be computed using the standard deviation $s_i$, estimated when the $i$-$th$ observation is omitted.

$$e_{Ji} = \frac{e_i}{s_i \times \sqrt{1 - H_{ii}}} . \quad (20)$$

These externally studentized residuals are $t$-distributed, so they are the type of residuals most often used for outlier detection; commonly a significance level of 0.95 is utilized.

The predicted residuals are residuals of one more type used for outlier detection:

$$e_{Pi} = \frac{e_i}{1 - H_{ii}} . \quad (21)$$

Externally studentized or predicted residuals are commonly used for outlier detection.

The model correctness can be easily assessed by drawing a plot with residuals or standardized residuals on the ordinate axis and predicted or actual $y$ values on the abscissa axis [27]. Figure 1 presents examples of such plots.

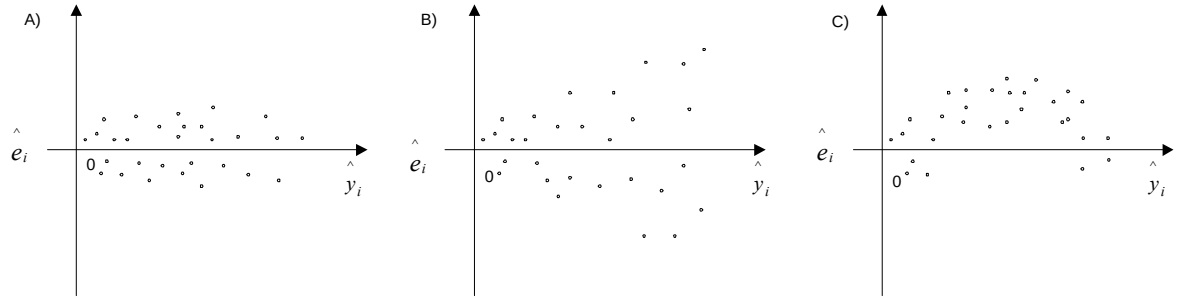


**Figure 1.** Examples of plots used to assess a model

Residuals of a correct model are given in plot A. The growing variance of residuals, shown in plot B indicates some deficiencies in the model used. The shape created by residual points shown in plot C indicates that some independent variable is missing or maybe the square of present variable omitted. It is important to notice that an observation outlier not necessarily produces a residual outlier, thus such an analysis is not suitable for outlier detection. More complicated graphical analysis is used for that purpose.

The Williams graph plots the externally studentized residuals on the ordinate axis and diagonal elements of the **H** matrix on the abscissa axis. Parallel to the abscissa axis the boundary line $y = t_{0.95}(n-m-1)$ is drawn for outliers, where $t$ stands for the Student distribution, 0.95 is the significance level and *(n-m-1)* is the number of degrees of freedom. Parallel to the ordinate axis the line $y = 2m/n$ is drawn for high-leverages [16].

The Pregibon graph plots the squared standardized residuals on the ordinate axis and diagonal elements of the **H** matrix on the abscissa axis. Two lines are drawn: $y = -x + 2(m+1)/n$ and $y = -x + 3(m+1)/n$. If a point is between them, so it is an influential point. If a point is above the upper line, it is a high influential point. An influential point can be a high-leverage point or an outlier [13].

The McCulloh and Meeter graphs plot the natural logarithm of the squared internally studentized residuals on the ordinate axis and $\ln(H_{ii}/(m(1-H_{ii}))$ on the abscissa axis. Parallel to the abscissa axis the line $y = -x - \ln F_{0.9}(n-m, m)$ is drawn for outliers. Parallel to the ordinate axis the line $y = \ln(2/(n-m) \times (t^2{}_{0.95}(n-m-1)$ is drawn for high-leverages [9].

The Gray L-R graph plots the standardized squared residuals on the ordinate axis and diagonal elements of the **H** matrix on the abscissa axis. The hyperbolic line $y = (2x - x^2 - 1)/(x(1-K)-1)$ is drawn for influential points, where $K = n(n-m-1)/c^2 m$ and c is usually equal to 2,4 or 8 [26].

### 3.3.2. Influence measures

There are numerous influence measures used for outlier detection. Cook's *D* measure shows the difference between the regression coefficients estimated, when the *i-th* observation is included and omitted [27].

$$D_i = \frac{(\mathbf{b}_i - \mathbf{b})^T (\mathbf{X}^T \mathbf{X})(\mathbf{b}_i - \mathbf{b})}{ps^2}, \qquad (22)$$

where $\mathbf{b}_i$ is the vector of regression coefficients computed without the *i-th* observation, $p$ is the sum of diagonal elements of the **H** matrix and $s^2$ is the variance.

A point is influential if $D_i$ exceeds $F_{(a)}(p, n-p)$, when $a$ equals to 0.5. A more simple rule can be used: an observation is influential if $D_i$ exceeds *4/n* [47], where $n$ is the number of observations.

The Welsh and Kuh measure $WK_i$ is very similar to $D_i$, with the difference that the variance estimate $s_i^2$, computed omitting the *i-th* observation, is used instead of $s^2$:

$$WK_i = \frac{(\hat{\mathbf{y}}_i - \hat{\mathbf{y}})^T (\hat{\mathbf{y}}_i - \hat{\mathbf{y}})}{ps_i^2}. \qquad (23)$$

DFFITS reveals the impact of the *i-th* observation on the predicted value $\hat{y}$

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{s_i \sqrt{H_{ii}}} = e_{Ji}^2 \times \frac{H_{ii}}{1-H_{ii}}, \qquad (24)$$

where $\hat{y}_{i(i)}$ is the predicted $\hat{y}_i$ value and $s_i$ is the standard deviation, computed without the *i-th* observation. It is assumed that outliers are observations with *DFFITS* exceeding 2 or more. The cut-off value, dependent on the sample size is $2\sqrt{p/n}$.

The Atkinson measure is closely connected with *DFFITS*.

$$A_i = |e_{ji}^2| \sqrt{\frac{n-p}{p} \times \frac{H_{ii}}{1-H_{ii}}} = \\ = |DFFITS_i| \sqrt{\frac{n-p}{p}}. \qquad (24)$$

The cut-off value is $2\sqrt{(n-p)/n}$.

The measure D, DFFITS, and the Atkinson measure are very similar, therefore, usually only one of them is used.

DFBETAS reveals the impact of the *i-th* observation on separate regression coefficients:

$$DFBETAS_{ji} = \frac{b_j - b_{j(i)}}{s_i \sqrt{(\mathbf{X}^T \mathbf{X})_{jj}^{-1}}}, \qquad (25)$$

where $b_j$ is the estimation of the *j-th* coefficient, $b_{j(i)}$ and $s_i$ are the estimation of the *j-th* coefficient and the standard deviation, when the *i-th* observation is omitted, and $(\mathbf{X}^T \mathbf{X})_{jj}^{-1}$ is the *j-th* element of the main diagonal of $(\mathbf{X}^T \mathbf{X})^{-1}$.

COVRATIO determines the influence of the *i-th* observation on the covariance matrix determinant:

$$COVRATIO_i = \frac{\det\left[s_i^2 (X_i^T X_i)^{-1}\right]}{\det\left[s^2 (X^T X)^{-1}\right]}. \qquad (26)$$

Observations with the COVRATIO value about 1 are non-influential, for influential observations COVRATIO is out of the following range $(1-3m/n,\ 1+3m/n)$.

The Andrews-Pregibon measure expresses the influence of the *i-th* observation on the volume of the confidence ellipsoid [36]:

$$AP_i = 1 - H_{ii} - e_{Ni}^2 . \quad (27)$$

Observations are influential if $1 - AP_i > 2(m+1)/n$.

The Cook-Weisberg likelihood measure is given by the difference between the logarithm of the likelihood function maximum value estimated using all data points and the corresponding value obtained when the *i-th* observation is omitted [15].

$$LD_i = 2(L(\hat{?}) - L(\hat{?}_i)) . \quad (28)$$

This measure allows examining the observation influence onto regression coefficients, or the variance of residuals, or both. Thus, the vector $\hat{?}$ can contain regression coefficients or variance values. The cut-off for influential points is $LD_i > c^2_{1-a}(m+1)$, where $c^2$ is the chi square distribution.

### 3.4. Robust regression

The approaches analyzed so far detect and remove outliers, while the robust regression reduces their influence.

Even one outlier can have a significant impact on regression coefficients estimated by the ordinary least squares. The robust regression is one of means to get more reliable estimates. M estimates, the high breakdown value estimates and their combination are the most known techniques [6].

The standard regression model is given by:

$$\mathbf{y} = \mathbf{X}ß + \mathbf{e} , \quad (29)$$

where $\mathbf{y}$ is a $n \times 1$ vector of dependent variable values, $\mathbf{X}$ is a $n \times m$ matrix containing values of independent variables, $ß$ is a $m \times 1$ vector of regression coefficients, and $\mathbf{e}$ is a $n \times 1$ vector of true errors with standard deviation $s$. The estimate $\mathbf{b}$ of $ß$ is obtained as a solution to:

$$\min_b Q_{OLS}(\mathbf{b}) , \quad (30)$$

where $Q_{OLS} = \sum_{i=1}^{n} e_i^2$ is in the ordinary least squares case.

One of the estimates, classified as robust, is the least absolute deviation (LAD), minimizing the sum of absolute values of residuals:

$$Q_{LAD} = \sum_{i=1}^{n} | y_i - \mathbf{x}? | , \quad (31)$$

where $\mathbf{x}_i$ is the i-th row of $\mathbf{X}$.

However, LAD is also very sensitive to outliers.

M estimate, introduced by Huber, is one of the most important robust estimates [19]. M estimates belong to the class of generalized maximum likelihood estimators. Instead of minimizing the sum of squared residuals, M estimator minimizes the sum of less rapidly increasing function of residuals. The less rapidly increasing cost or loss function p, which is, symmetric with a unique minimum at zero, is used:

$$Q_m = \sum_{i=1}^{n} \mathrm{p}(e_i) . \quad (32)$$

In the ordinary least squares case, $\mathrm{p}(e) = e^2$. Differentiating this expression with respect to the regression coefficients yields:

$$\sum_{i=1}^{n} ?(e_i)\mathbf{x}_i = \mathbf{0} , \quad (33)$$

where $? = \mathrm{p}'$.

The solution is not invariant to scaling. Therefore, residuals are scaled:

$$\sum_{i=1}^{n} ?(e_i / d)\mathbf{x}_i = \mathbf{0} , \quad (34)$$

where $d$ is some robust estimator of scale, for example, $S_{MAD}$.

M estimators are vulnerable to high-leverage points. It is why the high breakdown value estimates not sensitive to outliers were proposed. For example, the breakdown point of mean is 0, because even one outlier changes its value, the breakdown value of median is 0.5, since its value remains unchanged until the number of outliers reaches 50% of the sample size. The breakdown value of M estimator is *n/m*, where *n* is the number of observations and *m* is the number of parameters. The least median of squares (LMS) and the least trimmed squares (LTS) are one of the first robust estimates with a high breakdown value. The LMS estimate minimizes the median of squared instead of the sum of squared residuals:

$$Q_{LMS} = e_h , \quad (35)$$

where h is determined by $(n+1)/2 < h < (3n+m+1)/4$ and $m$ is the number of variables. When computing this estimate, 50% of largest values are eliminated and the remaining are used for the minimization. The breakdown value is *(n-h)/2*.

In the case of LTS, only a fraction *h* of the squared errors is used to estimate the regression parameter vector $\mathbf{b}$.

$$Q_{LTS} = \sum_{i=1}^{h} e_i^2 . \quad (36)$$

The breakdown value *(n-h)/2* is the same as for LMS. However, because of the higher convergence rate and smoother objective function the method is more often used.

S estimator is a robust estimator with a high breakdown value. It generalizes the LTS and LMS estimators. Regression coefficients are computed by minimizing variance of the coefficients:

$$b = \arg\min_{b} s_S , \tag{37}$$

where $s_s$ is the variance obtained from the following equation:

$$\left(\frac{1}{n-p}\right)\sum_{i=1}^{n} c\left(\frac{e_i}{s_S}\right) = K , \tag{38}$$

where $K$ is a constant equal to $\int c(s)d\Phi(s)$ and Tukey's bisquare or Yohai function is usually used as the ? function [6].

By combining a high breakdown value estimator and M estimator a MM estimator is obtained. First, the parameters $\mathbf{b}_{S,LTS}$, $s_{S,LTS}$ are obtained by applying a high breakdown value estimator, usually S or LTS. Then, M estimator is applied, while scaling with $s_{S,LTS}$.

$$Q_{MM} = \sum_{i=1}^{n} p\left(\frac{e_i}{s_{S,LTS}}\right). \tag{39}$$

If $Q_{MM}$ has many solutions, the one with the minimum variance $s(\mathbf{b})$ is picked.

### 3.5. Classification

A classification task is to find a mapping $f : X \to C$, where each data vector $\mathbf{x}_i$ is assigned to one of $C_i$, where $C = \{C_1, C_2, ...C_Q\}$, Q is the number of decision classes. There are only two classes: normal data and outliers in the outlier detection case. There is a large variety of classification algorithms [11]. Five big groups can be distinguished:

1. Statistical methods: the Bayese rule, Parzen classifier, methods based on discriminant analysis, logistic regression, nearest neighborhood methods [28].
2. Decision trees [44].
3. Kernel methods [2,24].
4. Artificial neural networks [2, 5, 43].
5. Combination of various classifiers [31].

In this study, only a brief review of the kernel methods is presented.

A transformation function $f$ maps original data into a new space, where a nonlinear decision boundary appears linear, as illustrated in Figure 2.

The new space, wherein the data are mapped, is called a feature space and denoted as $H$.

$$f : X \times X \to H . \tag{40}$$

Inner products $(\mathbf{x}, \mathbf{x})$ are often used in the classification process. The kernel function $k(\mathbf{x}, \mathbf{x})$ computes the inner product in the feature space directly, without explicitly computing the mapping $f$

$$k(\mathbf{x}, \mathbf{x}) = (f(\mathbf{x}), f(\mathbf{x})) . \tag{41}$$

The kernel matrix, also called the Gram matrix, contains kernel function values:

$$\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) . \tag{42}$$

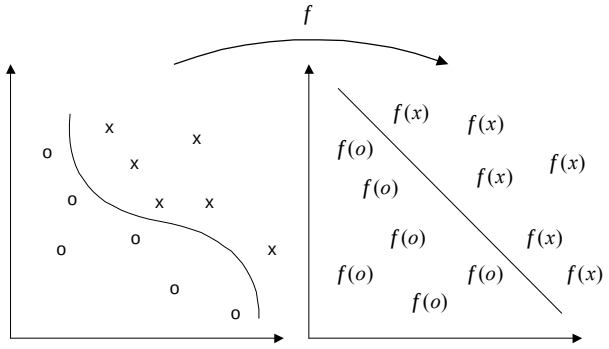


**Figure 2.** Transformation function $f$ maps the original data into a new space

Kernel methods based outlier detection can be performed by finding an appropriate hypersphere. Data points falling outside the hypersphere are then defined as outliers. The task is to find the radius and the center of the hypershere. To increase stability of the solution, usually it is allowed for some "good" data points to fall outside the hypersphere. Thus, the task is to minimize the radius of the hypersphere $r$ and the number of data points outside [24]:

$$\min_{\mathbf{c},r,x} \quad r^2 + C\sum_{i=1}^{n} x_i , \tag{43}$$

where $c$ is the center of the hypersphere, $n$ is the number of observations, $x_i$ is the so-called slack variable equal 0 for data points inside and measures the degree to which the distance squared from the center exceeds $r^2$ for points outside. The parameter $x_i$ allows to leave some data points outside, and the constant $C$ controls the trade-off between the radius $r$ and the slack variables. The task is solved by maximizing the Lagrangian function:

$$W(a) = \sum_{i=1}^{n} a_i k(\mathbf{x}_i, \mathbf{x}_i) - \sum_{i,j=1}^{n} a_i a_j k(\mathbf{x}_i, \mathbf{x}_j) \qquad \sum_{i=1}^{n} a_i = 1;\ 0 \le a_i \le C;\ i = 1,...,n. \tag{44}$$

A data point $\mathbf{z}$ is an outlier if:

$$k(\mathbf{z}, \mathbf{z}) - 2\sum_{i=1}^{n} a_i k(\mathbf{z}, \mathbf{x}_i) + + \sum_{i,j=1}^{n} a_i a_j k(\mathbf{x}_i, \mathbf{x}_j) \ge r^2. \tag{45}$$

The multilayer perceptrons are often used for classification. Due to numerous references available [2, 5, 43] the approach is not reviewed here.

## 4. Experiments

Sale transactions of land lots from one homogeneous unit, dated from 2003 till 2004, were chosen

for the experiment. The Register center is the institution registering the transactions and developing the value maps, it assumes this unit as one value zone. The unit is a suburb of one city. Parameters of the sample are as follows: the sample size is 67, the average price per acre is 2804.09 Lt (812.79 €), the median of the price is 2175.49 Lt, the standard deviation is 2517 Lt, the minimum and the maximum prices are 3.34 Lt and 11235 Lt, respectively. The parameters already show that there are some outliers between the sale transactions. Independent variables chosen for modeling are: the size of land lot, the distance to the center of the city, and the level of communications. Parameters of the linear regression equation, estimated by the ordinary least squares using all the data, are given below:

$$\hat{y} = x_1 \times 6584.76 + x_2 \times -10.39 + x_3 \times 1489.24 , \quad (46)$$

where $y$ is the price of the land lot, $x_1$ is the size, $x_2$ is the distance to the center and $x_3$ stands for the level of communications. The Student t values are: $t_1 = 20.06, t_2 = 10.21, t_3 = 6.01$, thus all exceed the cut-off value $t_{1-l/2}(67-3) = 1.99$, $l = 0.95$.

The matrix of the correlation coefficients is equal to:

$$\mathbf{R} = \begin{bmatrix} y & x_1 & x_2 & x_3 \\ 1 & 0.9 & -0.02 & 0.39 \\ 0.9 & 1 & -0.004 & 0.16 \\ -0.02 & -0.004 & 1 & 0.12 \\ 0.39 & 0.16 & 0.12 & 1 \end{bmatrix}. \quad (47)$$

The correlation coefficient values between the independent variables are small, the highest value is 0.16 between the size and the level of communications.

The value of the Durbin-Vatson coefficient is 1.7983, thus there is no autocorrelation between the observations and there is no need to investigate the impact of time. This can be explained by a short period of sale transactions. The variance inflation factors are: $VIF_1 = 1.00, VIF_2 = 1.00, VIF_3 = 1.04$, so there is no multi-collinearity.

One method from each group has been tested in the outlier detection task. The CDC method was used from the distance-based group. The method was applied with auto an robust scaling (equations (1), (5), (6) and (7)) and is referred to as $CDC_{AS}$ and $CDC_R$, respectively. The projection matrix **H** based analysis (equations (9) and (10)), the principal component based analysis, the externally studentized residuals (20), the Cook distance (22), COVRATIO (27), the M estimator based robust regression (35), the kernel methods (equations (44), (45) and (46)), and the multilayer perceptron are the methods used. In the PCA case, a linear discriminant function was constructed in the space of the first two principal components. The multilayer perceptron was trained using Bayesian regularization, which prevents data over-fitting. The perceptron was tested using the leave-one-out approach.

The expert of real estate has selected 42 sale transactions, where the sale price was consistent with the market price. These transactions can be fully trusted, other 25 transactions are doubtful, but only 17 of them can be clearly identified as outliers. After some analysis, it was decided to consider the remaining eight doubtful transactions as normal. Parameters of the regression equation computed without the outliers are given below:

$$\hat{y}_{Expert} = x_1 \times 6952.55 + x_2 \times -8.45 + \\ + x_3 \times 1111.39 . \quad (48)$$

This regression equation is used to assess the effectiveness of the outlier detection methods. The data points identified as outliers by a particular method are excluded and the remaining data are used to estimate the model parameters. Having the parameters, the prediction $\hat{y}$ is made and compared with $\hat{y}_{Expert}$ by computing the following distance:

$$Dif = \frac{1}{exp\_n} \sum_{i=1}^{exp\_n} | \hat{y}_{Expert(i)} - \hat{y}_i | , \quad (49)$$

where $exp\_n = 50 - FP$ with $FP$ being the number of false positives (the number of normal observations denoted as as outliers by the particular method).

Two more measures have been used in the evaluation process. The correct classification rate $Perf$ is one of them:

$$Perf = \frac{n - (FN + FP)}{n} , \quad (50)$$

where $n$ is the number of observations and $FN$ stands for false negatives – undetected outliers.

The determination coefficient $R^2$ was the second measure. It shows how large fraction of the variance of the dependent variable is explained by the independent variables.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} , \quad (51)$$

where $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$ is the total sum of squares, $SSR = \sum_{i=1}^{n} (\hat{y}(x_i) - \bar{y})^2$ is the sum of squares due to regression and $SSE = \sum_{i=1}^{n} e_i$ stands for the sum of squared residuals (errors). The $R^2$ measure must be adjusted, when the number of observations is close to the number of variables. This is not the case in our study.

The observations denoted as outliers by the methods are given in Table 1.

**Table 1.** Observations denoted as outliers by the different methods

| **Method** | **Results (outlier, influential point)** |
|---|---|
| $CDC_{AS}$ | 3, 40 |
| $CDC_{R}$ | 40 |
| Analysis of the **H** matrix | 3, 40, 64, 66 |
| Externally studentized residuals | 40, 64, 66 |
| Cook's distance | 3, 40, 64, 66 |
| COVRATIO | 3, 64 |
| Kernel method | 3, 12, 34, 39, 40, 44, 45, 46, 50, 51, 57, 58, 59, 62, 65 |
| PCA | 23, 39, 42, 45, 46, 47, 50, 51, 52, 57, 59, 66 |
| Multilayer perceptron | 23, 30,31, 38, 39, 42, 43, 45, 46, 47, 50, 51, 52, 57, 59, 60, 64, 66, 67 |
| Outliers, selected by expert | 23, 30, 31,39, 42, 43, 45, 46, 47, 50, 51, 52, 57, 59, 60, 64, 66 |

**Table 2.** Comparison of the outlier detection methods

| **Method** | **Parameters** | | | | | |
|---|---|---|---|---|---|---|
| | $x_1$ | $x_2$ | $x_3$ | $R^2$ | ***Dif*** | ***Perf*** |
| $CDC_{AS}$ | 3246.79 | -5.08 | 1379.45 | 0.563 | 13465.11 | 0.71 |
| $CDC_{R}$ | 6627.69 | -10.27 | 1363.31 | 0.893 | 11089.87 | 0.73 |
| Analysis of the **H** matrix | 4096.15 | -6.08 | 1347.04 | 0.612 | 10854.42 | 0.74 |
| Externally studentized residuals | 6837.25 | -10.01 | 1290.49 | 0.920 | 8427.38 | 0.76 |
| Cook's distance | 4096.15 | -6.08 | 1347.04 | 0.612 | 10854.42 | 0.74 |
| COVRATIO | 3631.87 | -5.89 | 1550.73 | 0.594 | 11383.23 | 0.74 |
| Kernel method | 1896.07 | -1.5 | 676.88 | 0.456 | 17762.44 | 0.80 |
| Robust regression | 6548.83 | -10.40 | 1489.20 | 0.885 | 9120.62 | - |
| PCA | 6899.89 | -8.55 | 985.48 | 0.909 | 5002.70 | 0.92 |
| Multilayer perceptron | 6952.59 | -8.39 | 1105.42 | 0.935 | 322.28 | 0.97 |

Observe that the M estimator based robust regression approach does not identify outliers.

Table 2 provides values of the three measures used to assess the methods as well as the regression equation parameter values computed in the way explained above.

The obtained results show that the multilayer percpetron is the best technique for categorizing the data, followed by the PCA based approach. This is expected, since both methods utilize supervised learning meaning that training data must be analyzed and labeled by an expert. The method based on externally studentized residuals provided the best performance amongst the techniques trained without supervision. The kernel-based approach detects many outliers, but also erroneously denotes many normal observations as outliers, so it is not as effective as the externally studentized residuals. Moreover, a teacher is often utilized to set the hyper-parameters in the kernel based approach. Therefore, the externally studentized residuals based technique is preferred, if there is no possibility of using expert knowledge.

## 5. Conclusions

Outliers are observations, inconsistent with the majority, which corrupt the parameters of a model of mass valuation. Therefore, outliers must be detected and removed, or parameters must be estimated using robust techniques.

Outlier detection methods can be categorized into four large groups, namely methods evaluating the distance from the observation to the data center, methods based on the difference between the actual and predicted values of the dependent variable, robust estimators of regression model parameters and data classification into the outlier and inlyer classes based techniques.

Most of these techniques can be trained without supervision. However, the best outlier detection results were obtained using the supervised learning based approaches, namely the multilayer perceptron and the PCA based technique constructing a linear discriminant function in the space of the first two principal components. Notwithstanding the need of using expert knowledge, the supervised training based approaches are of great interest, since when training is completed, the techniques can be used without any additional expert interruption. The externally studentized residuals based method is preferred, if there is no possibility of exploiting expert knowledge.

Further research topic will be techniques for fusing analysis results obtained from the different outlier detection methods.

## References


**[1]** **A. Damodaran.** Investment Valuation: Tools and Techniques for Determining the Value of Any Asset. *Second Edition, Wiley*; 2*nd edition*, 2002.

**[2]** **A. Verikas, A. Gelžinis.** Neuroniniai tinklai ir neuroniniai skaiciavimai, Technologija, Kaunas, 2003, 175.

**[3]** **A.K.F. Siu.** Applications of influence analysis. *Hong Kong Economics Papers*, 18, 1987, 23-42.

**[4]** **B. Walczak, D.L. Massart.** Multiple outlier detection revisited. *Chemometrics and Intelligent Laboratory Systems*, 41, 1998, 1-15.

**[5]** **C. Bishop.** Neural networks for pattern recognition. *Oxford: Oxford University Press*, 1996

**[6]** **C. Chen.** Robust regression and outlier detection with the ROBUSTREG procedure. *SAS Institute Inc., Cary, www2.sas.com/proceedings/sugi27/265-27.pdf*, 2005.

**[7]** **C. Wikstrom, Ch. Albano, L. Eriksson, H. Friden, E. Johansson, A. Nordahl, S. Rannar, M. Sandberg, N. Kettaneh-Wold, S. Wold.** Multivariate process and quality monitoring applied to an electrolysis process. *Part* I. *Process supervision with multivariate control charts, Chemometrics and Intelligent Laboratory Systems*, 42, 1998, 221-231.

**[8]** **C. Zu, H. Kitagawa, S. Papadimitriou, Ch. Faloutso.** OBE: outlier by example. *Proceedings of PAKDD*, 2004.

**[9]** **C.E. McCulloch, D. Meeter.** Discussion of outliers by R.J. Beckman and R.D. Cook. *Technometrics*, 26, 1984, 197-208.

**[10]** **C.G. Lalor, Ch. Zhang.** Multivariate outlier detection and remediation in geochemical databases. *The Science of the Total Environment*, 281, 2001, 99-109.

**[11]** **Ch-L. Liu, H. Fujisawa.** Classification and Learning for Character Recognition: Comparison of Methods and Remaining Problems in Neural Networks and Learning in Document Analysis and Recognition. *First IAPR TC3 NNLDAR Workshop, Seoul, Korea*, 2005, 1-7.

**[12]** **D. Pena, J.F. Prieto.** Multivariate outlier detection and robust covariance matrix estimation. *Technometrics*, 2001, *Vol*.43, *No*.3.

**[13]** **D. Pregibon.** Logistic regression diagnostics. *Annual Statistics*, 9, 1981, 45-52.

**[14]** **D.E. Ramirez.** Noncentral generalized F distributions with applications to joint outlier detection. *Proceedings of Computional biology and bioinformatics, Symposium on the interface: computer science and statistics*, 2004.

**[15]** **D.R. Jensen, D.E. Ramirez.** Bringing order to outlier diagnostics in regression models. *Proceedings of Recent advances in outlier detection, summer research conference in statistics/ASA*, 2001.

**[16]** **D.X. Williams.** Letter to the editor. *Applied Statistics*, 22, 1973, 407-408.

**[17]** **E. Hund, L.C. Massart, J. Smeyers-Verbeke.** Robust regression and outlier detection in the evaluation of robustness tests with different experimental designs. *Analytica Chimica Acta*, 463, 2002, 53–73.

**[18]** **E.C. Malthouse.** Ridge regression and direct marketing scoring models. *Journal of Interactive Marketing, Vol*.13, *No*.4, 1999, 10-23.

**[19]** **F. Hampel.** Robust statistics: a brief introduction and overview. *Research report* 24, *Seminar for statistics*, 2001.

**[20]** **G.C. Lalor, Ch. Zhang.** Multivariate outlier detection and remediation in geochemical databases. *The Science of the Total Environment*, 281, 2001, 99-109.

**[21]** **Y. Chen.** Outliers detection and confidence interval modification in fuzzy regression. *Fuzzy Sets and Systems*, 119, 2001, 259-272.

**[22]** International Valuation Standards Seventh edition, Appraisal Inst, 2005.

**[23]** **J. Hardin, D.M. Rocke.** Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. *Computational Statistics & Data Analysis*, 44, 2004, 625 -638.

**[24]** **J. Shave-Taylor, N. Christianini.** Kernel methods for pattern analysis. *Cambridge University Press*, 2004, 462.

**[25]** **J.A.F. Pierna,F. Wahl, O.E. Noord, D.L. Massart.** Methods for outlier detection in prediction. *Chemometrics and Intelligent Laboratory Systems*, 63, 2002, 27– 39.

**[26]** **J.B. Gray.** Graphics for regression diagnostics. *Proceedings of Statistical Computing Section*, 102-107, 1985.

**[27]** **J.O. Rawlings, G.S. Pantula, D.A. Dickey.** Applied regression analysis. *A research tool, Secaucus, NJ, USA: Springer-Verlag New York*, 1998, 657.

**[28]** **K. Fukunaga.** Introduction to Statistical Pattern Recognition. 2*nd edition, Academic Press*, 1990.

**[29]** **K.A. Hoo, K.J. Tvarlapati, M.J. Piovoso, R. Hajare.** A method of robust multivariate outlier replacement. *Computers and Chemical Engineering*, 26, 2002, 17–39.

**[30]** **L.H. Chiang, R.J. Pell, M.B. Seasoltz.** Exploring process data with the use of robust outlier detection algorithms. *Process Control*, 2003, 13, 437-449.

**[31]** **L.I. Kuncheva.** Combining clasifiers: soft computing solutions. *Patterns recognition: from classical to modern approaches, World Scientitific*, 2001, 427-451.

**[32]** Lietuvos Respublikos turto ir verslo vertinimo pagrindu istatymas Nr. VIII-1202. 1999-05-25, *Žin*. 1999, *Nr*.52.

**[33]** Lietuvos Respublikos turto ir verslo vertinimo pagrindu istatymo 8 straipsnio pakeitimo istatymas, Nr. IX-1428, 2003-04-03, *Žin*. 2003, *Nr*.38.

**[34]** Lietuvos Respublikos vyriausybes nutarimas Del Nekilnojamojo turto vertinimo taisykliu patvirtinimo, 2005 09 29, *Žin*, 2005, *Nr*.117.

**[35]** Lietuvos Respublikos žemes istatymas, Nr. IX-1983, 2004-01-27, *Žin*. 2004, *Nr*.28-868.

**[36]** **M. Meloun, J. Militky.** Detection of single influential points in OLS regression model building. *Analytica Chimica Acta*, 2001, *No*.439, 169-191.

**[37]** **M.M. Brening, H-P. Kriegel, R.T. Ng, J. Sander.** LOF: identifying density-based local outliers. *Proceedings ACM Sigmod*, 2000.

**[38]** **M.R. Linne, S.M. Kane, G. Dell.** A Guide to appraisal valuation modeling. 2000, 200.

**[39]** **N. Billar, A.S. Hadi, P.F. Velleman.** BACON: blocked adaptive computationally efficient outlier nominators. *Computional Statistics & Data Analysis*, 34, 2000, 279-298.

**[40]** **N. Nguyen, A. Cripps.** Predicting housing value: a comparison of multiple regression analysis and artificial neural networks. *Journal of real estate research*, *Vol*.22, *No*.3, 2001.

**[41]** **P. Filzmoser.** A multivariate outlier detection method. *Proceedings of the Seventh International Conference on Computer Data Analysis and Modeling, Belarusian State University*, 2004, 18 - 22.

**[42]** **R.J. Gloudemans.** Mass appraisal of real property. 1999, 429.

**[43]** **R.O. Duda, P.E. Hart, D.G. Stork.** Pattern Classification. *Second edition, John Wiley & Sons*, 2001.

**[44]** **T. Mitchel.** Machine learning. *McGraw Hill*, 1997.

**[45]** **T. Zewotir, J. S. Galpin.** Influence Diagnostics for Linear Mixed Models. *Data Science* 3, 2005, 153-177.

**[46]** **V. Cekanavicius, G. Murauskas.** Statistika ir jos taikymas I. *TEV, Vilnius*, 2003, 238.

**[47]** **V. Cekanavicius, G. Murauskas.** Statistika ir jos taikymas II. *TEV, Vilnius*, 2004, 268.

**[48]** **V. Kontrimas.** An application of hybrid computational methods to real estate markets. *Proceedings of Information technologies* 2005, *Technologija, Kaunas*, 2005, 141-144.

**[49]** **V. Roth.** Sparse kernel regressors. *Artificial Neural Networks-ICANN* 2001, *Springer, LNCS* 2130, 2001, 339-346.

**[50]** **V. Sakalauskas.** Statistika su Statistica. *Margi raštai, Vilnius*, 1998, 223.

**[51]** **W. Zhao, D. Chen, Sh. Hu.** Detection of outlier and a robust BP algorithm against outlier. *Computers & Chemical Engineering, Vol*.28-8, 2004, 1403-1408.

**[52]** **W.J. Egan, S.L. Morgan.** Outlier detection in multivariate analytical chemical data. *Analytical chemistry*, 1998, 70, 2372-2379.

Received March 2006.