Use of Collocation Method as an Improved Sales Comparison Approach

Sebastian ZADDACH and Hamza Alkhatib, Germany

Key words: real estate valuation, regression analysis, collocation

SUMMARY

The sales comparison approach, which is used in real estate valuation, is from a mathematical-statistical point of view based on the multiple linear regression analysis. Since decades it has been considered as a standard procedure for analysing the real estate market and to determine the current market value of real estates. Nevertheless, since its introduction the method has not been enhanced significantly. The estimated comparative value is in particular depending on the number and the type of value influencing variables which are considered within the regression equation. However, the multiple regression analysis suffers from the facts, that on the one hand complex interactions cannot be expressed by a functional relationship, on the other hand, a large number of influencing factors and their interaction on the purchase price are not known or cannot be imposed.

The aim of this contribution is to enhance the use of regression analysis in real estate valuation by introducing a collocation approach. The research approach extends the system equation of the regression by a signal component. This additional component contains the mentioned value effecting influences, which cannot be modeled in an adequate way and thus the collocation is able to exhaust additional information. The results of the regression analysis as a linear trend model are improved; the extended approach is able to give a higher explanation to the real purchase prices. The methodology is tested on a real data set and the results are validated by comparison of the predicted market values with real purchase prices.

Use of Collocation Method as an Improved Sales Comparison Approach

Sebastian ZADDACH and Hamza Alkhatib, Germany

1. INTRODUCTION

The general aim of a real estate valuation is the determination of the market value. In Germany, the Federal Building Code and the released legal rules make sure, that the valuation experts determine market-conform and generally accepted values for real estates. The focus of this research is based on the sales comparison approach, in which the market value is determined by comparing the valuation object with purchase prices of properties that match the valuation object in a sufficient manner concerning the value influencing characteristics.

As a mathematical-statistical method for this valuation purpose the regression analysis was already adapted in the early 1920 for appraisal purposes and can be considered as a standard since the 1980s in Germany. For random samples in which many purchase cases are given, it can be considered as the valuation method with the highest marketability. Within this approach a dependent variable (e. g. the purchase price) is explained by a linear (if applicable non-linear) function representing a general trend model. The model of the linear multiple regression analysis assumes that there is a deterministic context between the purchase price and its value influencing characteristics.

The aim of this contribution is based on the fact, that on the one hand in real estate market analysis influence quantities occur, which cannot be expressed by a functional relationship (e. g. location), on the other hand a large number of influencing factors and their interaction on the purchase price are not known or cannot be imposed (e. g. characteristics of the purchaser). More differentiated value relations cannot be expressed by easy functions. The deterministic model therefore cannot be formed up fully and correctly. In order to take account of an information content, which is not used up to now, it is possible to take advantage of the collocation approach based on Ordinary Least Squares (OLS). This approach was originally established to solve problems in physical geodesy; first attempts were already developed at the end of the 1970's (e. g. Moritz 1980). Shortly after its introduction, the theory was applied successfully for questions of valuation purposes (e. g. Pelzer 1978, Ziegenbein and Hawerk 1978, Uhde 1982). By introducing the collocation, the methodology of the regression analysis shall be extended. One of the main aims is to improve the modeling of uncertainty and the exhaustion of information, which has not been not considered up to now. The collocation approach extends the system equation of the regression by a signal component. This additional component contains the mentioned value effecting influences, which cannot be modeled in an adequate way. The amount of signal obtained for each purchase data set is composed of a linear combination of irregular and systematic differences that consist of the non ascertainable characteristics, so that this additional component involves parts of the false specifications. In order to recognize and quantify the additional information it can be assumed that the residuals of the purchases, which are associated the closest to the valuation object, have a similar explainable proportion of value effecting influences. From this information an autocovariance function can be derived, which describes the behavior of the value between the objects.

2. USE OF REGRESSION ANALYSIS FOR VALUATION PURPOSES

The use of the regression analysis can be regarded as a standard in real estate valuation since the researches accomplished by Ziegenbein (1977) approx. 30 years ago. Since the adaption, the regression analysis has proved itself for multifunctional purposes; thus it is not only used to determine the market value as it is defined in the Federal Building Code for developed and undeveloped real estates, but also to deduce relevant data for real estate valuation (e. g. conversion factors, index series). Simply, the multiple linear regression analysis is a model which enables to explain a dependent variable (observation) by several value influencing characteristics, the independent variables. In order to explore such issues, data on the underlying variables of interest are assembled and the regression analysis is used to estimate the quantitative effect of the causal variables upon the variable that they influence.

2.1 Theory of Multiple Regression Analysis

The regression analysis as an approximation method gives the possibility to explain variations of the dependent variable by the variability of the independent variables to indicate a functional connection. Besides, the connections between the dependent variable and the independent variables are often known only vaguely, knowledge about an exact function is not given. For the applications in valuation purposes this means that the market value of a real estate is compared to the purchase prices of other properties that match the valuation object in a sufficient manner concerning the value influencing characteristics as independent variables. From a mathematical-statistical point of view a model ist put up, in which the observation y_i (e. g. purchase price per m²) is expressed regarding to m independent variables:

$$y_i = b_0 + b_1 x_{i1} + \ldots + b_m x_{im} + u_i = b_0 + \sum_{j=1}^{m} b_j x_{ij} + u_i, \qquad [1]$$

where $i = 1, \ldots, n$ and n = number of data sets. y_i is the dependent variable, $x_{i1},\ldots ,x_{im}$ are the independent or explanatory variables, and u_i is the residual or error term. The residuals contain the difference between the real observations and the estimated value. The residuals are characterised as the stochastic component of the regression model, whereas the functional connection is characterised as the systematic component (Fahrmeir et al. 2009). The goal of regression analysis is the estimation of the systematic component based on the given data set for the observation and its value influencing parameters and to separate it from the residuals. In Eq. [1] $b_0,\ldots, b_m$ indicate the regression coefficients. The coefficient b_0, so called intercept or offset, describes a constant term. It is slightly evident that in Eq. [1] every independent variable has an linear effect on the dependent variable. In case of matrix-vector notation, the dependent variable can be noted as $\mathbf{y} = [y_1 \quad y_2 \quad \ldots \quad y_n]'$, the unknown parameters as $\mathbf{b} = [b_0 \quad b_1 \quad \ldots \quad b_m]'$ and the residuals as $\mathbf{u} = [u_1 \quad u_2 \quad \ldots \quad u_n]'$. The independent variables are combined in the full ranked design matrix $\mathbf{X}$ with dimension $[n \times m + 1]$:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{bmatrix}. \qquad [2]$$

Eq. [1] can now be represented by

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{u}. \quad [3]$$

The functional model is complemented with the stochastic model which is given by the variance-covariance matrix (VCM) of the observation

$$\mathbf{\Sigma}_{\mathbf{yy}} = \sigma^2 \mathbf{P}^{-1}, \quad [4]$$

where σ^2 is the unknown variance of unit weight, and **P** is the known weight matrix of the observations. Concerning valuation approaches, the single observations are usually regarded as independent from each other with equal variance. Then the weight matrix can be formulated as **P = I**, where **I** is the identity matrix (Koch 1995). The optimal estimates of the unknown regression coefficents **b** can be given by means of OLS , i. e. the minimization of the residuals sum of squares:

$$\hat{\mathbf{b}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{y}\right). \quad [5]$$

The residuals can now be estimated as follows:

$$\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\mathbf{b}}. \quad [6]$$

2.2 Prediction

The OLS technique ensures the estimation of the unknown parameters, which fit the sample of data sets best in the specific sense of minimizing the sum of squared residuals. After the whole model is tested on its validity and the significance of the coefficients, the regression function can be used to predict new values, usually in real estate valuation only within the borders of the data sample. The prediction is calculated by

$$\hat{\mathbf{y}}_{\mathbf{r}} = \mathbf{X}_{\mathbf{r}}\hat{\mathbf{b}}. \quad [7]$$

In this equation, $\hat{\mathbf{y}}_{\mathbf{r}}$ includes the objects, which have to be predicted, and $\mathbf{X}_{\mathbf{r}}$ the value influencing variables for each object. The use of the regression analysis for valuation purposes presents itself as an iterative process which will run through until an optimum regression function is found. The definition as optimum regression function is a matter of fulfilling different conditions, the possibilities to deal with single defects are not topic of this contribution and can rather be found in Ziegenbein (1977) or Fahrmeir et al. (2009).

3. USE OF COLLOCATION METHOD FOR VALUATION PURPOSES

Using the described application of regression analysis, the value of an object y_i is estimated. The real causal relationship between dependent variables and value influencing variables can only be given as a rough approximation. When comparing the function values with an actual purchase price, a difference remains which reflects the imperfect modeling. The principle of the collocation approach is based on the consideration that each (originally geodetic) observation can be split into a regular systematic component (*trend*), an irregular systematic component (*signal*) and an irregular random component (*noise*). In the context of this distribution the collocation is given as a Gauss-Helmert model; the functional model is discussed in section 3.1, the stochastic model in section 3.2. For a more in-depth look at the adjustment approach of the Gauss-Helmert model see Niemeier (2008).

3.1 Functional Model

According to Eq. [3] the classical regression analysis represents a functional trend, which is overlain by a random white noise. The functional trend gives the explanation of the purchase price by the most dominant value influencing characteristics. In analogy to the proposed method of regression analysis, the functional model of Eq. [3] is extended to the separation of the residuals **u** into a signal part **s** and a random noise component **v,** so in case of a simple collocation the equation can be split into:

$$\mathbf{y} = \underbrace{\mathbf{Xb}}_{trend} + \underbrace{\mathbf{s}}_{signal} + \underbrace{0 \cdot \mathbf{s_p}}_{prediction\ of\ signal} + \underbrace{\mathbf{v}}_{noise}. \qquad [8]$$

The component $0 \cdot \mathbf{s_p}$ describes the predicted signal which will be part of the practical examinations of this contribution. The modeling of the trend **Xb** can be done (depending on the task), for example by estimating lower or higher order polynomials, splines or harmonic functions. For applications in the context of this research initially a simple approach based on linear functions will be used. The irregular random noise occurs only in the measured values, which is different to the stochastic signal. In the collocation approach, the trend parameters are determined through an adjustment based on OLS, so that the trend function **Xb** adapts the measuring points in an optimal way. The observations (in this case: the purchase prices) are improved by the signal values of **s**; the noise **v** has to be filtered out in order to exhaust the regular and irregular-systematic information. Finally, at locations where no original observations $\mathbf{y_p}$ are available, the values of the trend component $\mathbf{X_p b}$ and the predicted signal $\mathbf{s_p}$ can be used to calculate predicted observations.

3.2 Stochastical Model

In particular the prediction mentioned in the last section plays an important role in the estimation of the target size. In a mathematical sense this is an approximation method, in which a value is determined at any point from the existing, surrounding values. For this purpose stochastic information of the observations can be used in the collocation. Contrary to the classical regression analysis, the prediction in the collocation is performed by exhausting the dependences of stochastically neighboring observations, which are described by their covariances. For any random observation vector **y** the VCM is given by

$$\mathbf{\Sigma_{yy}} = \begin{bmatrix} \sigma_1^2 & \cdots & \rho_{1n}\sigma_1\sigma_n \\ \vdots & \ddots & \vdots \\ \rho_{n1}\sigma_n\sigma_1 & \cdots & \sigma_n^2 \end{bmatrix}. \qquad [9]$$

Here, σ_i denotes the standard deviation of *i*-th observation and ρ_{ik} the correlation coefficient between the *i*-th and the *k*-th observation. In order to simplify the approach for the application in the context of the valuation, it can be assumed, that $\mathbf{\Sigma_{yy}} = \mathbf{I}$. Similarly, the VCM can also be established for the signal, postulating the distribution $\mathbf{s} \sim N(\mathbf{0}, \mathbf{\Sigma_{ss}})$. The cross-correlations of the signal and the VCM of the observations are neglected:

$$\mathbf{\Sigma_{ys}} = \mathbf{\Sigma_{sy}} = \mathbf{0}. \qquad [10]$$

Regarding these conditions and including the predicted signal $\mathbf{s_p}$, the stochastic model for the collocation results in

$$\overline{\boldsymbol{\Sigma}} = \begin{bmatrix} \boldsymbol{\Sigma}_{yy} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{ss} & \boldsymbol{\Sigma}_{ss_p} \\ \mathbf{0} & \boldsymbol{\Sigma}_{s_p s} & \boldsymbol{\Sigma}_{s_p s_p} \end{bmatrix} = \sigma_0^2 \, \overline{\mathbf{Q}} \qquad [11]$$

where $\overline{\boldsymbol{\Sigma}}$ is the VCM, σ_0^2 the a priori variance factor and $\overline{\mathbf{Q}}$ is the cofactor matrix (Moritz 1980). An important task in the collocation is the allocation of $\boldsymbol{\Sigma}_{ss}$, the VCM of the signal. The signal is stochastically dependent, since it involves the systematic effects. This can be realized by assuming that neighboring observations have an equivalent or similar behavior, while observations that are farther away from each other, can accordingly be considered by an independent behavior. The matrix $\boldsymbol{\Sigma}_{ss}$ is thus given by

$$\boldsymbol{\Sigma}_{ss} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \cdots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & 1 \end{bmatrix} \qquad [12]$$

in which the correlation coefficient ρ_{ik} refers to the correlation of the signal components s_i and s_k, and in this case express the accordance concerning the value of two compared objects i and k. Of particular importance in this context is to find a function to evaluate $\rho_{ik} = f\left(s_i, s_k\right)$, which represents the stochastic context correctly. The solution to this problem is the implementation of autocovariance functions, which approximate the behavior of the signal and can be mathematically derived from the empirical data. In geodetic applications, autocovariance functions are commonly used in time series analysis to model the stochastic relationships between values of one or several time-ordered sequences of observations, the so-called time series (see Moeser et al. 2000). The determination of an appropriate autocovariance function for use in valuation purposes is described in section 4.3.

3.3 Estimation, Filtering and Prediction

Using the stochastic context between the signal components the three main tasks - estimation of trend parameters, filtering and prediction - can be estimated by means of Gauss-Helmert model (see Eq. 8). After the introduction of the first simple trend model based on multiple linear regression analysis, the trend parameters are estimated again with the additional information from the VCM of the signal and thus are improved. The adjusted coefficients are given by:

$$\hat{\mathbf{b}}_{Koll} = \left(\mathbf{X}'\mathbf{H}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{H}^{-1}\mathbf{y} \qquad [13]$$

where $\mathbf{H} = \mathbf{Q}_{yy} + \mathbf{Q}_{ss}$ and $\hat{\mathbf{b}}_{Koll}$ denotes the coefficients of the trend component. Compared to the multiple linear regression analysis, which is optimized in regard to the required valuation standards, the regression coefficients now change by re-estimating the collocation model. After estimating the Lagrange multipliers by

$$\hat{\mathbf{k}} = \mathbf{H}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}_{Koll}\right), \quad [14]$$

the residuals can be calculated by

$$\hat{\mathbf{v}} = \mathbf{Q}_{yy}\hat{\mathbf{k}}\,. \quad [15]$$

The estimated signal can be determined by

$$\hat{\mathbf{s}} = \mathbf{Q}_{ss}\hat{\mathbf{k}}\,. \quad [16]$$

Based on these results, the next step is the prediction. For this purpose, the values of the validation sample are predicted on the basis of the estimation and filtering, in which no observational data, but only the independent parameters are given. The relationship

$$\hat{\mathbf{s}}_{\mathbf{p}} = \mathbf{Q}_{s_p s}\hat{\mathbf{k}} \quad [17]$$

can identify the signal components in the prediction data. The calculation of the trend component is done by solving the equation

$$\hat{\mathbf{y}}_{\mathbf{p}}^{\,+} = \mathbf{X}_{\mathbf{p}}\hat{\mathbf{b}}_{Koll} \quad [18]$$

where $\hat{\mathbf{y}}_{\mathbf{p}}^{\,+}$ is the predicted observation and $\mathbf{X}_{\mathbf{p}}$ contains the independent variables of the validation sample. If the observations are determined in this way, they have to be improved by the predicted signal component:

$$\hat{\mathbf{y}}_{\mathbf{p}} = \hat{\mathbf{y}}_{\mathbf{p}}^{\,+} + \mathbf{s}_{\mathbf{p}}\,. \quad [19]$$

4. PRACTICAL RESEARCH APPROACH

In the following chapter, the collocation introduced in chapter 3 will be put into practice for valuation approaches. The data base used here is a sample of the Automated Purchase Records (APR) of the expert committee for real estate valuation in Hanover. According to the research approach displayed in Fig. 1, the experiment is divided in the steps of data preparation, estimation, filtering and prediction.

4.1 Description of the Data Sample

As a first spatial submarket the capital of Lower Saxony, the city of Hanover, is chosen. For the objective submarket a homogeneous sample has to be formed, for which a sufficiently large number of cases is available and where the objects are distributed evenly across the study area. Due to the available data, undeveloped land is selected which is reserved for future residential use. In addition to the value affecting characteristics, which are included in the database of the APR, additional data about each object is collected, in particular with regard to the characterization of the location. Overall, based on the available criteria, 1.199 sales from the years 2000 to 2011 with a total of 37 features for each object can be determined as the data sample used for this study. In order to be able to assess the quality of the collocation approach, the method of cross-validation is applied. The basic idea of this method is to separate a part of the original sample and to perform the estimation with the remaining data. For the subsequent validation the observations from the previously separated sample are

estimated by the collocation approach and then are compared with the original observations of the validation sample (Cressie 1993). As a preparation step (Fig. 1) a data sample with a size of 10 % of the whole data set is randomly separated as a validation sample.

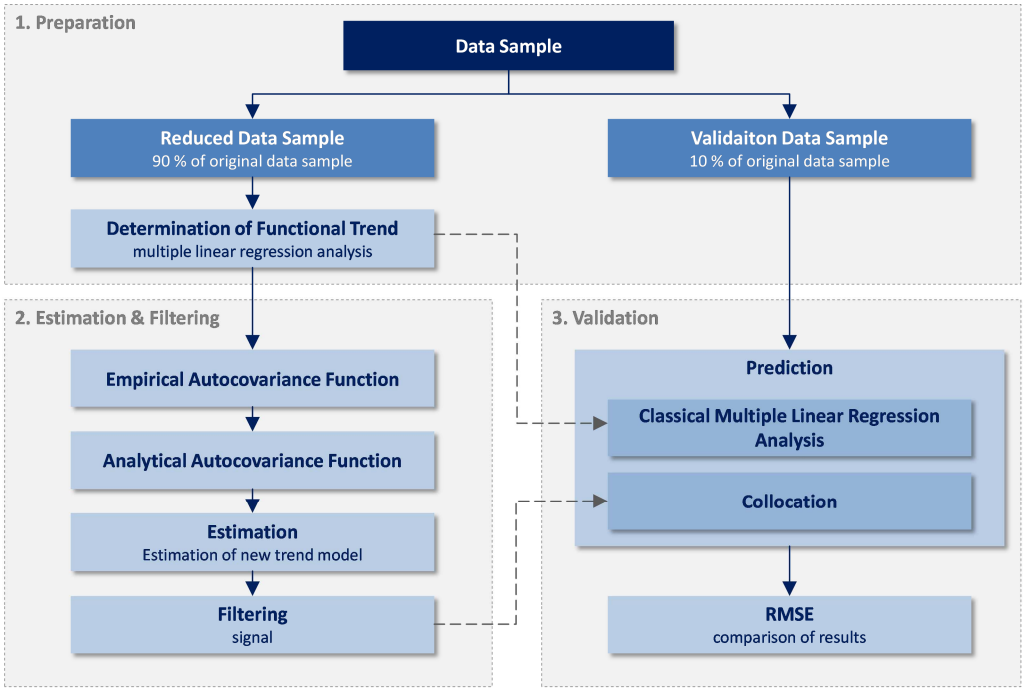

Fig. 1: Research approach.

4.2 Estimation of the Trend Function

According to Fig. 1, a further step of data preparation is to perform a regression analysis with the aim of estimating a suitable trend function. This procedure also allows to reduce the collected value influencing characteristics to a few significant factors (by the determination of significant regression coefficients). After reduction of the total sample for the validation sample, a total of 1.069 data sets is available for analysis. The dependent variable is appointed to purchase price per area (€/m²). The determination of an optimal regression function is based on the requirements that are applied in the context of the valuation standards and which are not discussed in this paper (e. g. Ziegenbein 1977, Uhde 1982). The estimated simple trend function is introduced into the collocation approach to determine the current non exhausted information components.

4.3 Determination of the Empirical Autocovariance Function

After the determination of the functional trend, an empirical autocovariance function has to be estimated in order to be able to derive an appropriate analytical autocovariance function. As described in Section 3.2, the estimation, filtering and prediction require determining an appropriate autocovariance function that reflects the stochastic behavior of the data in the form of an autocovariance function correctly. In contrast to the classical time series analysis, valuation approaches require a modified definition of this function as the observations are not defined by time, but take an undefined number of independent variables into account.

Ziegenbein (1977) and Pelzer (1978) have solved the problem by the introduction of so-called Mahalanobis distance, which can be used to express the diversity of single lots as an *m*-dimensional distance of value. The number of dimensions *m* denotes the number of features that are considered as value influencing. If the value characteristics of two compared objects *i* and *k* are summarized in a vector, the result is given by

$$\mathbf{x}_i = \left[x_{i1}\ x_{i2} \ldots x_{im} \right], \quad [20]$$

$$\mathbf{x}_k = \left[x_{k1}\ x_{k2} \ldots x_{km} \right]. \quad [21]$$

From these vectors a difference can be computed, denoted as the indicator δ_{ik} which expresses the diversity of the two objects as it can be concluded from the selected value characteristics:

$$\delta_{ik}^2 = \left(\mathbf{x}_i - \mathbf{x}_k \right) \left(\mathbf{x}_i - \mathbf{x}_k \right)' = \sum_{j=1}^{m} \left(x_{ij} - x_{kj} \right)^2. \quad [22]$$

The variable $\sqrt{\delta_{ik}^2} = \delta_{ik}$ is called the distance of value between two objects. Mathematically Eq. [22] describes a *m*-dimensional space, where δ_{ik} expresses the distance between the position vectors $\mathbf{x}_i$ and $\mathbf{x}_k$. The smaller this distance is, the lower the value differences between the objects are. Pelzer (1978) has proofed, that the correlation coefficient ρ_{ik} between the signals s_i and s_k of two objects can be regarded as a function of the distance of values:

$$\rho_{ik} = f\left(\delta_{ik} \right). \quad [23]$$

For the determination of the values of the empirical autocovariance function, the distance of values to all objects according to Eq. [22] is calculated for each object in the data sample. The so-formed value intervals are summarized in classes for which an appropriate class width has to be specified. For the current sample, 24 classes are defined in which all calculated distances of value can be sorted. In addition, the linear dependence of the successive distance classes are estimated (Niemeier 2008). The tendency that the successive values, here the classes k_i and k_{i+d} (where *d* is the distance of the classes), are not stochastically independent from each other, is expressed by the autocovariance function *C(d)* with

$$C\left(d \right) = \frac{1}{n\text{-}d} \sum_{i=1}^{n-d} u_i\, u_{i+d} \quad [24]$$

where *d* is the class, *n* the number of values in the class and u_i are the residuals calculated in the trend model. For $d = 0$ the calculation leads to the empirical variance of the observations; for $d > 0$, the empirical covariances of the values, which are classified in the distance *d*, are calculated. By normalizing the empirical autocovariance function with the empirical variance *C(0)*, the empirical autocorrelation function is calculated which is denoted by *K(d)*. A graphical illustration of the empirical autocorrelation function for the present sample is depicted in Fig. 2.

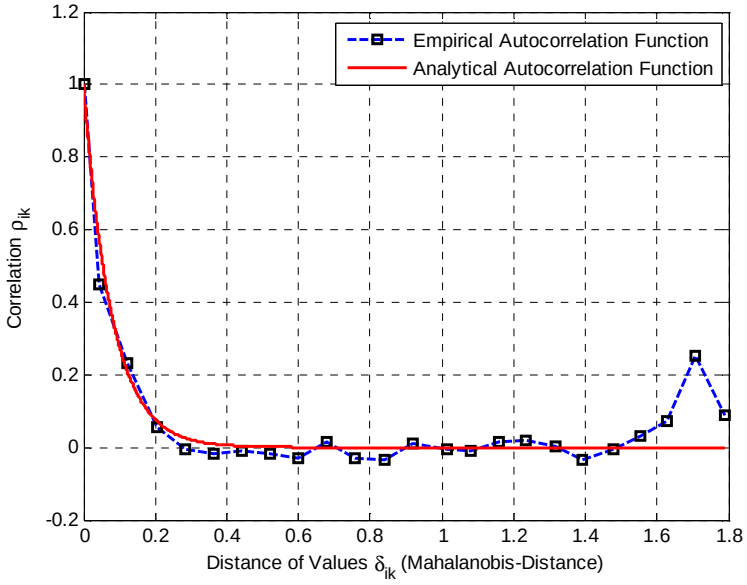

Fig. 2: Empirical and analytical autocorrelation functions.

The empirical function forms the basis for the analytical determination of autocorrelation function and thus to fill the required matrix $\mathbf{\Sigma}_{ss}$. Based on the empirical curve in Fig. 2 the assumption of an exponential function of the form

$$\rho_{ik} = e^{a(\delta_{ik})} \qquad [25]$$

can be assumed. The choice of the exponential function can be justified so that objects with small distance of value are higher correlated than those with larger distances. Therefore the autocorrelation has to decrease with increasing value-distance. For the transition of the empirical values to an analytic solution, the parameter *a* in Eq. [25] is to estimate as a functional determinant coefficient. Here, a simple regression analysis can be used to find an appropriate estimate. For the present data sample the estimation of the coefficient is -3.99. Additionally, the distance of value is weighted for a better adaptation of the analytical function to the empirical value:

$$\rho_{ik} = e^{-3{,}99\left((\delta_{ik})/max(\delta_{ik})^2\right)}. \qquad [26]$$

The resulting analytical autocorrelation function is also depicted in Fig. 2. The determined functional relationship now allows the filling of the covariance matrix $\mathbf{\Sigma}_{ss}$ by substituting Eq. [26] into Eq. [12]. The matrix $\mathbf{\Sigma}_{s_p s_p}$ according to Eq. [11] corresponds to $\mathbf{\Sigma}_{ss}$ (Eq. [12]) concerning the structure, but here the calculated distance of value δ_{ik} between all objects to be predicted is used. Accordingly, in $\mathbf{\Sigma}_{ss_p}$ the distance of values between the objects of comparison and the objects to be predicated are used.

5. DISCUSSION OF THE RESULTS

In the final step, the results of both methods have to be compared: the prediction based on the application of classical multiple linear regression analysis and the prediction based on the collocation. For this purpose, the validation sample is used (see Fig. 1), which has been separated from the original sample. Both models are used to predict the dependent variables (€/m²) based on the validation sample. In addition, the results of the original observations, the real purchase prices are compared to the prediction.

5.1 Prediction of Purchase Prices

As mentioned before, the data sample is evaluated concerning general valuation standards. By using the top-down-strategy to identify the significant coefficients the number of value influencing characteristics of the originally collected 37 features can be reduced to only 7. Tab. 1 gives an overview of the trend model, which is obtained by applying the regression analysis on the given data base. It should be noted that this is the result of a random sample, where the validation sample was separated without any systematic approach. For any other combination of randomly seperated data samples, there will be numerically different values.

Independent variable	Regression coefficient	Range	
Intercept	112,965	-/-	[-]
GFZ	10,504	0,2 – 0,4	[-]
KD	0,005	08.02.00 – 22.03.11	[-]
BRW	0,698	105 – 520	[€/m²]
KOMP	-1,131	15,1 – 82,6	[-]
DISTZENTR	-5,213	1,16 – 9,60	[km]
DISTGRUEN	-11,931	0,01 – 2,40	[km]
VERA	13,807	0 or 1	[-]
Sample size	n = 1.096		
Determination coefficient	R^2 = 0,61		

Tab. 1: Results of the regression analysis (trend model).

Significant independent variables for the description of the selected submarket are: the floor space index according to the local land use planning (GFZ), the normalized date of purchase to 01/01/2000 (KD), the land value per m² (BRW), the compactness of the property as a measure of the lot form (COMP), the distance to the city center of Hanover (DISTZENTR), the distance to the next green and recreation area (DISTGRUEN) and as an additional dummy variable, particularly the seller (VERA, 0 for a private person, 1 otherwise). The determination coefficient of the calculated regression function is 0.61 and can be rated based on experience in valuation approaches as high representation of dependent variables by means of independent variables (Ziegenbein 1977). It should be noted that the regression results according to Tab. 1 are only valid in the range of the parameters; limits are set by the minimum and maximum of each independent variable. Every data set of the validation sample has to be within these limits. In cases, where single variables are beyond these limits, the complete data set has to be removed from the validation sample. After this review, 126 data sets remain in the sample, 4 data sets do not fulfill these requirements. For the remaining data, the dependent variable is predicted by the classical regression approach. Similarly, the dependent variable is predicted in the collocation model. As a result of the different

evaluations the 126 predicted values based on the regression ($\hat{\mathbf{y}}_r$), the collocation ($\hat{\mathbf{y}}_p$) and the original observations ($\mathbf{y}$) can be compared. A representation of the results is shown in Fig. 3. In the figure, the original purchase prices of the validation sample are sorted in ascending order, the predicted values from regression and collocation are also shown (in dashed red and blue lines, respectively). It can be clearly seen that both results, based on only regression and based on collocation, strongly deviate from the original prices at the edges of the sample (< 175 €/m² and > 325 €/m²). In contrary, a relatively good fit between prediction and real price purchases can be observed in the middle part. However, this result was expected because less data are available for the interpolation of estimates at the functional borders.

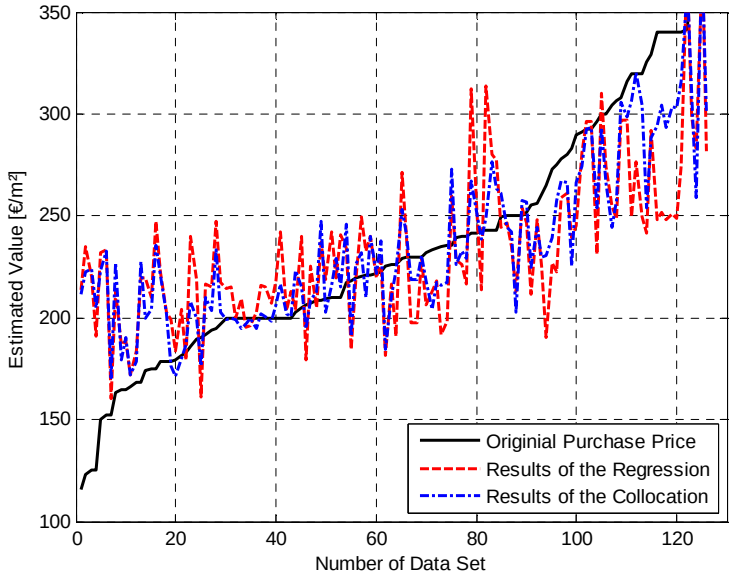

Fig. 3: Comparison of the results

The direct comparison of both model approaches shows that in most cases the residuals of the collocation are below the regression and thus, as expected, an additional information component is exploited, that provides a better fit to the original observations. To evaluate the quality of the models the mean square error (MSE) can be determined for both approaches (Willmott and Matsuura 2005). The MSE is based on the variance of the sample and is given by

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 , \qquad [27]$$

where n is the sample size of the validation set (here: n = 126) and for the predicted $\hat{y}_i$ the values from the regression or collocation function are used. For a more interpretable measure, the root of the MSE is additionally drawn (root mean square error, RMSE). The RMSE of the regression analysis is at approx. 44 €/m², the RMSE of the collocation approach at about 34 €/m². On average, the consideration of the signal component is able to improve the classical approach by about 23 %. Since the validation sample is removed as part of cross-validation randomly from the total sample, the results presented are only appropriate for these samples. In order to demonstrate the general improvement of the sales comparison approach extended

by the collocation, the methodology presented in Fig. 1 is repeated overall 500 times, i. e. the total sample in each of these evaluations is divided at random and the subsequent analysis steps are performed based on the current data set.
In this way, the validity of an individual evaluation enables a general assessment of the collocation approach in valuation applications. As a result of the 500 simulations, the RMSE for each run are presented. Fig. 4 (left) contains the RMSE for the regression analysis and evaluations of the associated collocation results of all simulation runs. In addition, the mean values over the 500 simulations are shown. The results of 500 runs confirm the first run, a general model improvement. The mean RMSE is improved by the collocation of approximately 44 €/m² to 36 €/m², which corresponds to a mean percentage increase in the information exploitation from 17.5% (Fig. 4, right). The comparative values, which have been predicted in the collocation approach, differ on average by approx. 11% of the original purchase price, which confirms the investigations in the 1970s of Ziegenbein and Hawerk (1978).

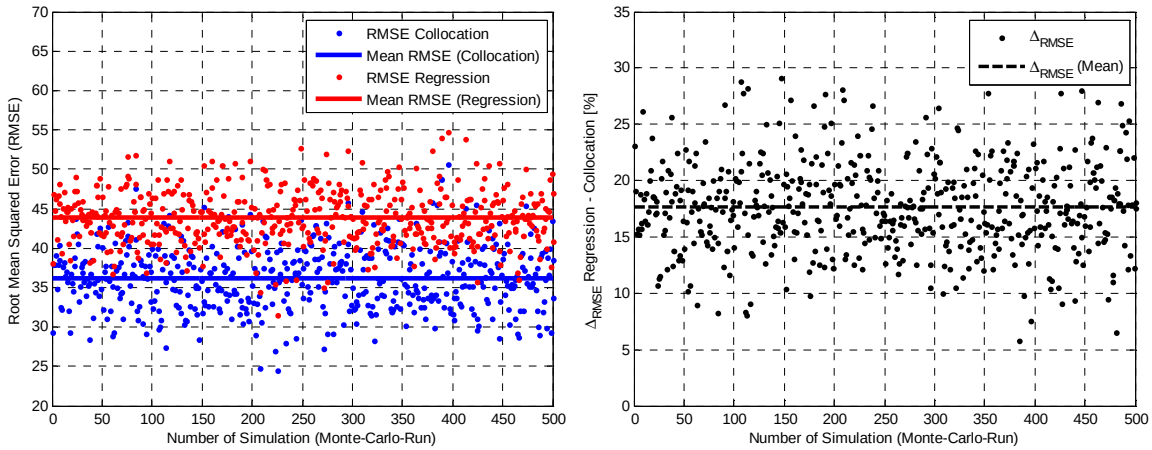

Fig. 4: Comparison of the RMSE.

6. OUTLOOK

The presented results have shown that the extension of the classical linear regression analysis for a collocation approach is a useful addition to valuation applications. The focus of the presented approach lies in the prediction of reliable comparative values, which are based on the data base by a few dominant value features that are easy to grasp and represent the original purchase prices in a good approximation. An increased computational effort to exploit the additional information can be tolerated. An advantage in the presented approach is the inclusion of an optimized regression analysis as a first trend approach to collocation, because in this way the advantages of the standard proven approach leads to a more refined evaluation approach. The goal of future research is to improve the modeling approach of collocation. These are in particular the geographical distances between the comparison and valuation objects, which have to be introduced as an additional determinant. In the expert's practice, it is common that objects that are close to the valuation object simply have a greater influence on the evaluation outcome. This situation has to be considered in determining the value-distances in the form of an additional weight. One first approach to this was already presented in Pelzer (1978) and Ziegenbein and Hawerk (1978). However, the studies refer to a

single object which is predicted. In the presented methodology reference values for 126 objects are estimated simultaneously, which makes an extension of previously existing approaches concerning the integration of geographical distance necessary. Another task is to transfer the methodology to other submarkets with the objective review of the general applicability of the approach.

REFERENCES

Cressie, N. A. C. (1993): Statistics for Spatial Data. Revised Edition Auflage. New York, Chichester, Toronto, Brisbane, Singapore: John Wiley & Sons, Inc.

Fahrmeir, L., Kneib, T., Lang, S. (2009): Regression. Modelle, Methoden und Anwendungen. 2. Auflage. Heidelberg, Dordrecht, London, New York: Springer Verlag.

Koch, K.-R. (1995): Statistische Grundlagen zur Untersuchung von Immobilienwerten. In Schmalgemeier, H. (Hrsg.): Statistische Methoden in der Grundstückswertermittlung. Band 16, Stuttgart: Verlag Konrad Wittwer, S.7–12.

Moritz, H. (1980): Advanced Physical Geodesy. Band 13, Sammlung Wichmann. Karlsruhe: Herbert Wichmann Verlag,.

Moeser, M., Mueller, G., Schlemmer, H., Werner, H. (Hrsg.) (2000): Handbuch Ingenieurgeodäsie. Auswertung geodätischer Überwachungsmessungen. Heidelberg, Herbert Wichmann Verlag, Hüthig GmbH.

Niemeier, W. (2008): Ausgleichungsrechnung: Statistische Auswertemethoden. 2. Auflage. Berlin, New York: Walter de Gruyter GmbH & Co. KG.

Pelzer, H. (1978): Ein indirektes Vergleichswertverfahren unter Anwendung statistischer Methoden. ZfV- Zeitschrift für Vermessungswesen, Jg. 103, Nr. 6, S.245–254.

Uhde, C. (1982): Mathematische Modelle zur Analyse von Grundstücksmärkten. Band Nr. 118, Wissenschaftliche Arbeiten der Lehrstühle für Geodäsie, Photogrammetrie und Kartographie an der Technischen Universität Hannover. Hannover.

Willmott, C. J. und Matsuura, K. (2005): Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing averagre model perfomance. Climate Research Clim Res, Nr. 30, S.79–82.

Ziegenbein, W. (1977): Zur Anwendung multivariater Verfahren der mathematischen Statistik in der Grundstückswertermittlung. Dissertation, Technische Universität Hannover, Hannover.

Ziegenbein, W. und Hawerk, W. (1978): Erfahrungen bei der Prädiktion von Grundstückswerten. ZfV - Zeitschrift für Vermessungswesen, Jg. 103, Nr. 6, S.254–261.

BIOGRAPHICAL NOTES

Dipl.-Ing Sebastian Zaddach graduated in Geodesy and Geoinformatics at the Leibniz Universitaet Hannover in 2007. He passed the second state exam as "Graduate Civil Servant for Surveying and Real Estates" in Lower Saxony in 2010. Since 2010 he has been working as a scientific assistant at the Geodetic Institute at the Leibniz Universitaet Hannover. His main research interests are: improvement of real estate valuation based on new statistical approaches, evaluation of uncertainty in real estate valuation.

Dr. Hamza Alkhatib received his Dipl.-Ing. in Geodesy and Geoinformatics at the University of Karlsruhe in 2001 and his Ph.D. in Geodesy and Geoinformatics at the University of Bonn in 2007. Since 2007 he has been postdoctoral fellow at the Geodetic Institute at the Leibniz Universitaet Hannover. His main research interests are: Bayesian Statistics, Monte Carlo Simulation, Modeling of Measurement Uncertainty, Filtering and Prediction in State Space Models, and Gravity Field Recovery via Satellite Geodesy.

CONTACTS

Dipl.-Ing. Sebastian Zaddach
Leibniz Universitaet Hannover
Geodetic Institute Hanover
Nienburger Straße 1
30167 Hanover
GERMANY
Tel. +49 511 762 17201
Fax +49 511 762 2468
Email: zaddach@gih.uni-hannover.de
Web site: www.gih.uni-hannover.de

Dr.-Ing. Hamza Alkhatib
Leibniz Universitaet Hannover
Geodetic Institute Hanover
Nienburger Straße 1
30167 Hanover
GERMANY
Tel. +49 511 762 2464
Fax +49 511 762 2468
Email: alkhatib@gih.uni-hannover.de
Web site: www.gih.uni-hannover.de