Development of a Robust Bayesian Approach for Real Estate Valuation in Areas with Few Transactions

Alexander DORNDORF, Matthias SOOT, Alexandra WEITKAMP, Hamza ALKHATIB, Germany

Key words: Real Estate Valuation, Areas with Few Transactions, Weighting of Data, Variance Component Estimation

SUMMARY

The real estate valuation is realized by three different valuation methods in Germany. However, the sales comparison approach is the method, which has the nearest affinity to the real estate market. Often, regression analysis is used in this case. The regression model needs normally 15 purchases per independent variable for an accurate estimate in real estate valuation. For this reason, in areas with few transactions (if only 10 to 30 purchases exist) the solution of regression is not satisfactory. Furthermore, the detection of outliers is a challenging task, because the number of purchases is small and each detected outlier reduce the sample size. Actually, in areas with few transactions, the real estate expert estimates the value by his own experience under considering the available market data, e.g. purchases or offer prices. The purpose of this study is to demonstrate a mathematical-statistical approach for the combination of all kind of data. For this three different data sets are used, which consist of purchase prices, knowledge of real estate experts and offer prices. The focus of this paper lies on the development of an optimal weighting approach between this data, which base on the variance component estimation. First, we use a closed loop simulation to validate and optimize the algorithm. In the second step the optimized algorithm is validated on the real market. Results show the advantage of this approach for the fusion of different data sets in area with few transactions.

ZUSAMMENFASSUNG

Für die Immobilienwertermittlung in Deutschland werden drei normierte Verfahren genutzt. Das Vergleichswertverfahren ist als marktnächstes Verfahren anzusehen. Beim Vergleichswertverfahren kommt regelmäßig die Regressionsanalyse zum Einsatz. Zur Lösung der Regression ist es allerdings notwendig, dass pro modellierter Einflussgröße 15 Kaufpreise vorliegen. Dies führt insbesondere in kaufpreisarmen Lagen zu Problemen, in denen nur 10 bis 30 Kauffälle pro Auswertung zur Verfügung stehen. Des Weiteren ist die Detektion von Ausreißern in kleinen Datensätzen eine Herausforderung. Zudem wird durch ausschließen der Ausreißer die Stichprobe weiter verkleinert. Das Ziel dieser Arbeit ist die Entwicklung und die Validierung eines mathematisch-statistischen Modells zur Kombination von verschiedenen Marktdaten. Der Fokus liegt hierbei auf der optimalen Gewichtung zwischen den unterschiedlichen Datensätzen (Kauffälle, Expertenbefragung und Angebotsdaten). Als Gewichtungsansatz wird die Varianzkomponentenschätzung verwendet. Im Rahmen dieser Arbeit wird in einem ersten Schritt eine Closed-Loop-Simulation durchgeführt um den verwendeten Ansatz zu untersuchen. Im zweiten Schritt wird der Ansatz an realen Daten getestet. Die Ergebnisse zeigen das Potential der Varianzkomponentenschätzung bei der Kombination der verschiedenen Marktdaten in kaufpreisarmen Lagen.

1. MOTIVATION

The reliable determination of property values is stringently necessary for different purposes e.g. for mortgage lending valuation. Wrong assessment of real estate markets, caused by missing market transparency can lead to serious consequences as could be seen in subprime crisis in 2008. High quality market information are required to determine certain results in real estate valuation. However, this determination requires an appropriate number of market data; that can be a challenging task in regions with few transactions.

The aim is to use a comparison approach with a multiple linear regression. For an accurate estimate of the regression coefficients, the regression model normally needs 15 purchases per independent variable (Ziegenbein 2010, Kleiber et al. 2014), but in areas with few transactions only few prices are available. Hence, it is challenging to provide an accurate estimation. In addition the detecting outliers in areas with few transactions leads again to reduce the sample size, and that tends to result in statistically unsolvable estimation task

Recently, property appraisers use their market expertise to determine market values in regions with few transactions. The few purchases are often not used methodically. Reuter (2006) adapted a Delphi method for his intersubjective price comparison by using the knowledge of appraisers to derive the market value. First approaches, which combine this experts' knowledge and transactions by means of Bayesian multiple regression approach, are derived by Weitkamp & Alkhatib (2012a) and Alkhatib & Weitkamp (2012). Alkhatib & Weitkamp (2013) and Weitkamp & Alkhatib (2014) suggested a robust estimate of the Bayesian regression model to deals with the problems caused by outliers. They replaced the well known normal distribution of the likelihood data by a student-distribution which allows to keep outliers in the estimation but to down weight their influence on the estimates. They only used in their approach one prior information in form of experts' survey.

Additional data and prior information are necessary to derive an accurate estimate of the real estate market. Usable market data and the acquisition of such data are discussed in Soot et al. (2016). This study uses three different data sets of one and two-family houses, which consists of purchase prices, knowledge of real estate experts and offer prices.

In this paper, we present a new approach to combine these different heterogeneous data in the context of the regression analysis. In addition to the usual estimates, the regression coefficients and their estimated uncertainties, we also estimate optimal weight factors of the different used data by means of variance component estimate (VCE).

In Section 2 the mathematical basics of the used approach is presented. First investigations on robust estimate of the regression analysis in real estate valuation context are discussed in Section 2.3. In Section 3, the developed method is shown. The used data sets and their quality are presented in Section 4. After this, the approach is validated with a closed loop simulation (Section 5.1) as well as real data sets (Section 5.2). Finally, the results are discussed in Section 6.

2. MATHEMATICAL BASICS

2.1 Classical Multiple Linear Regression in Real Estate Valuation

Since decades, the multiple linear regression is used in the sales comparison approach (Ziegenbein 1977, Pelzer 1978). In this model, the input quantities of a real estate (e.g. area of lot or standard land value) explain the purchase price. The functional model follows:

$$\boldsymbol{y} = \boldsymbol{\beta_0} + \boldsymbol{x_1\beta_1} + \cdots + \boldsymbol{x_i\beta_i} + \boldsymbol{\varepsilon}, \quad \boldsymbol{i} = \boldsymbol{1}, \ldots, \boldsymbol{u}, \quad \boldsymbol{\varepsilon} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\sigma^2}) \qquad \text{Eq. 1}$$

The dependent variable y (in our case: standardized purchase prices) is explained by a linear combinations of the independent variables $\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{u}$ and the unknown regression coefficients $\beta_0, \ldots, \beta_\mathrm{u}$. These unknown regression coefficients are usually estimated by means of the method of least squares (Fahrmeir et al. 2009, Koch 1997). Due to the remaining disturbances between model and reality, the residuals $\boldsymbol{\varepsilon}$ arise as measure of the not explainable spread. They have to obey the normal distribution with the mean value 0 and the variance σ^2. Further discussion of regression analysis can be found in, e.g., Fahrmeir et al. 2009, Urban & Mayerl 2011.

2.2 Bayesian Approach for Regression Analysis

In areas with few transactions only an insufficient number of purchases for regression analysis is available. For this reason, a statistical estimation requires additional data, e.g. knowledge of real estate experts. A mathematical approach for the combination of purchases with prior information is the Bayes' theorem.

$$\boldsymbol{P}(\overline{\boldsymbol{\beta}}|\boldsymbol{y}) \propto \boldsymbol{P}\left(\underline{\boldsymbol{\beta}}\right) \cdot \boldsymbol{P}(\boldsymbol{y}|\boldsymbol{\beta}) \qquad \text{Eq. 2}$$

In this equation $P(\overline{\boldsymbol{\beta}}|\boldsymbol{y})$ is the posterior density. $P\left(\underline{\boldsymbol{\beta}}\right)$ is called prior density and the term $P(\boldsymbol{y}|\boldsymbol{\beta})$ is denoted as likelihood function. The likelihood function represents the information of the purchases. All additional market information are expressed and modeled in the prior density. The posterior density contains the result of regression coefficient by given data $\boldsymbol{y}$. Detailed information about the Bayesian inference can be found in, e.g., Koch 2007, Kacker & Jones 2003.
In the case of conjugate prior, the Bayes' theorem is analytically solvable. The difference between classical regression analysis and Bayesian estimation is shown in Table 1.

Table 1: Comparison between Classical and Bayesian Estimation of regression coefficients and cofactor matrix.

Classical Regression Analysis		Bayesian Parameter Estimation	
$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1}\boldsymbol{X^T y}$	Eq. 3	$\overline{\boldsymbol{\beta}} = \left(\boldsymbol{X^T X} + \underline{\boldsymbol{V}}^{-1}\right)^{-1}\left(\boldsymbol{X^T y} + \underline{\boldsymbol{V}}^{-1}\underline{\boldsymbol{\beta}}\right)$	Eq. 4
$\widehat{\boldsymbol{V}} = \boldsymbol{Q_{\beta\beta}} = (\boldsymbol{X^T X})^{-1}$	Eq. 5	$\overline{\boldsymbol{V}} = \left(\boldsymbol{X^T X} + \underline{\boldsymbol{V}}^{-1}\right)^{-1}$	Eq. 6

The classical regression analysis and likelihood function consist of the purchases with the variables $\boldsymbol{y}$ and $\boldsymbol{X}$. In Eq. 3, the parameter $\hat{\beta}$ is the result of the classical regression analysis. The difference between the classical regression and Bayesian estimation is the prior information, which is represented by the variables $\underline{\boldsymbol{V}}$ and $\underline{\boldsymbol{\beta}}$. Here, the underlined variables represent the prior information and the over lined variables stand for the posterior estimates. Experts' knowledge and offer prices are used as prior information in this study. A detailed description of the model and information about the data can be found in Section 4. The variable $\underline{\boldsymbol{\beta}}$ is the prior knowledge of the regression coefficient and $\underline{\boldsymbol{V}}$ is the appendant cofactor matrix. In Eq. 4 the variable $\overline{\boldsymbol{\beta}}$ is the posterior regression coefficient of the Bayesian estimation. A detailed discussion about conjugate prior and analytical solution of Bayes' theorem can be found in Koch (2007). Practical applications of the Bayesian regression for real estate data are demonstrated in Alkhatib & Weitkamp (2012), Weitkamp & Alkhatib (2012a), Weitkamp & Alkhatib (2012b).

2.3 Robust Estimation

The purchases often contain outliers. This is caused by the imperfect market and missing information about the origin of the purchase price. For the regression analysis, the outliers should be eliminated, otherwise the estimation is biased. In areas with few transactions the detection of outliers is a challenging task, because the number of purchases is small and each detected outlier reduces the sample size. Methods of robust parameter estimation are insensitive to outliers in the dependent variable. A variety of robust methods for the classical regression, as, e.g., M-estimators, L-estimators or R-estimator (Hartung et al. 2009), have been developed in the recent years. All methods introduce a weight matrix $\boldsymbol{\Omega}$ in the stochastic model. This matrix considers outliers with smaller weights as correct observations. In Eq. 3 and Eq. 4 an identity matrix $\boldsymbol{\Omega} = \mathbf{I}$, which indicate equal weights for all purchases, is used, therefore the weight matrix can be neglected. For the Bayesian estimation with weight matrix follows:

$$\overline{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{\Omega}\boldsymbol{X} + \underline{\boldsymbol{V}}^{-1}\right)^{-1}\left(\boldsymbol{X}^T\boldsymbol{\Omega}\boldsymbol{y} + \underline{\boldsymbol{V}}^{-1}\underline{\boldsymbol{\beta}}\right) \quad \text{Eq. 7}$$

In case of the Bayesian estimation, the robust methods for regression estimation cannot be adopted easily without modifying the proposed likelihood density. For this reason, the robustness has to be implement by means of assuming other distributions for the dependent variable and possibly for the prior distributions. Alkhatib & Weitkamp (2013) present a robust Bayesian approach for areas with few transactions, which uses an independent Student-distribution instead of the normal distribution to calculate the weight matrix $\boldsymbol{\Omega}$. The selection of Student density function (as non-conjugate prior density) leads to the problem, that the analytical solution cannot be derived easily. As a numerical solution for the resulting posterior density, the Gibbs sampler, a Markov-Chain-Monte-Carlo approach, has been used. A detailed discussion and overview of a robust Bayesian approach with Student-distribution can be found in Geweke (1993). Practical applications of robust parameter estimation with regression analysis and Bayes estimation for areas with few transactions is presented in Alkhatib & Weitkamp (2013) and Weitkamp & Alkhatib (2014).

2.4 Variance Component Estimation

In areas with few transactions, the existing purchases must be supported in the estimation by additional market data. The different data groups are characterized by heteroscedasticity and non-normality generally, which is more than challenging in terms of optimal weighting estimation. A more detailed description of the characteristics of the used data can be found in Section 4.1. For the combination of heterogeneous data sets, a statistical approach is the VCE. Eq. 4 is expanded as follow:

$$\overline{\boldsymbol{\beta}} = \left(\frac{1}{\sigma_L^2}\boldsymbol{X}^T\boldsymbol{X} + \frac{1}{\sigma_P^2}\underline{\boldsymbol{V}}^{-1}\right)^{-1}\left(\frac{1}{\sigma_L^2}\boldsymbol{X}^T\boldsymbol{y} + \frac{1}{\sigma_P^2}\underline{\boldsymbol{V}}^{-1}\underline{\boldsymbol{\beta}}\right) \quad \text{Eq. 8}$$

The terms σ_L^2 and σ_P^2 are introduced to consider the different unknown variance unit of Likelihood function and prior-information. In case of more than one type of prior information, each data set has an own variance component. Substituting Eq. 8 into Eq. 9 leads to Eq. 10, which is a general description for arbitrary number of data sets.

$$(\boldsymbol{V})^{-1}\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{y} \quad \text{Eq. 9}$$

$$\boldsymbol{\beta} = \left(\frac{1}{\sigma_1^2}\boldsymbol{X}_1^T\boldsymbol{X}_1 + \cdots + \frac{1}{\sigma_i^2}\boldsymbol{X}_i^T\boldsymbol{X}_i\right)^{-1}\left(\frac{1}{\sigma_1^2}\boldsymbol{X}_1^T\boldsymbol{y}_1 + \cdots + \frac{1}{\sigma_i^2}\boldsymbol{X}_i^T\boldsymbol{y}_i\right), \quad i = 1, \ldots, k \quad \text{Eq. 10}$$

Here, the original observations $\boldsymbol{y}$ of the prior data are used directly, e.g. offer prices. The variance component σ_i^2 of every data set i is depend on the magnitude of their variance. A data set with a large variance component decreases the influence of the corresponding data set in the whole estimation process while a small variance component increase its influence. There are different approaches in the literature to optimally estimate the variance components, like Helmert method, maximum likelihood estimation and minimum norm quadratic unbiased estimators. Further information about VCE can be found in Amiri-Simkooei (2007) and Koch & Kusche (2002). The method proposed in this paper is given in Koch & Kusche (2002).

3. Optimal VCE Approach in the multiple Regression Analysis

3.1 Developed Approach

In this paper, the focus lies on an optimal weighting approach for the different market data with the VCE. The results of this investigation will be used for the development of a robust Bayesian approach in future works. The VCE is introduced successfully in other geodetic applications, e.g. geopotential determination from satellite data or adjustment of global positioning system network (Koch & Kusche 2002, Yang et al. 2005). A practical example for real estate valuation is presented in Uhde (1982). Here, the variance components depending on the dispersion of purchase prices are estimated in one data set. For the combination of purchase prices, experts' knowledge and offer prices, VCE is not introduced before. Hence, the main aim is to investigate the suitability of VCE for these different data sets. Therefore, in areas with few transactions the challenges lies on the estimation of optimal weights. For this purpose, we assume that data sets are free from any outliers. Hence, Eq. 10 can be used directly. The unknown variance components σ_i^2 are iteratively estimated by the approach described in Koch & Kusche (2002):

$$\hat{\sigma}_i^2 = \frac{\hat{\boldsymbol{e}}_i^T\,\hat{\boldsymbol{e}}_i}{r_i}, \quad i = 1, \ldots, k \quad \text{Eq. 11}$$

where $\hat{\boldsymbol{e}}_i$ are the estimated residuals of every data set and k denotes the number of data sets. The residual $\hat{\boldsymbol{e}}_i$ are estimated by means of the estimated posterior coefficients in the corresponding

iteration step. Thus the estimated posterior $\hat{\boldsymbol{\beta}}$ are considered in the variance component estimation. The term r_i denotes the partial redundancy and considers the contribution of one data set to the total estimation with all data sets. Therefore, a data set with a small residual sum of squares and large redundancy significantly affects the result of $\hat{\boldsymbol{\beta}}$. On the contrary, a data set with large residual and small redundancy have a small influence on the estimates of coefficients. Detailed derivations can be found in Koch & Kusche (2002).
The investigation of this VCE approach follows in two steps. First, a closed loop simulation is developed to validate the algorithm on a scenario for areas with few transactions. In the second step, the algorithm is validated on the real market. Afterwards, these results are compared to the results of the simulation.

3.2 Closed Loop Simulation

The calculation process of the closed loop simulation is depicted in Figure 1 schematically. As input parameter the independent variables of purchase cases and experts' survey are used. Both, they build the data set $\boldsymbol{X}$. Estimated regression coefficient of the considered submarket are used as expected values $\mathrm{E}(\boldsymbol{\beta})$. The error model consists of variance, offset and scale factor for every simulated data set. In this study, three data sets are simulated. They include the characteristics of purchases in areas of few transactions, knowledge of experts and offer prices (Section 4 presents a detailed discussion of the used data sets and their errors). Then the expected value of the observation $\mathrm{E}(\boldsymbol{y})$ is calculated with $\boldsymbol{X}$ and $\mathrm{E}(\boldsymbol{\beta})$.

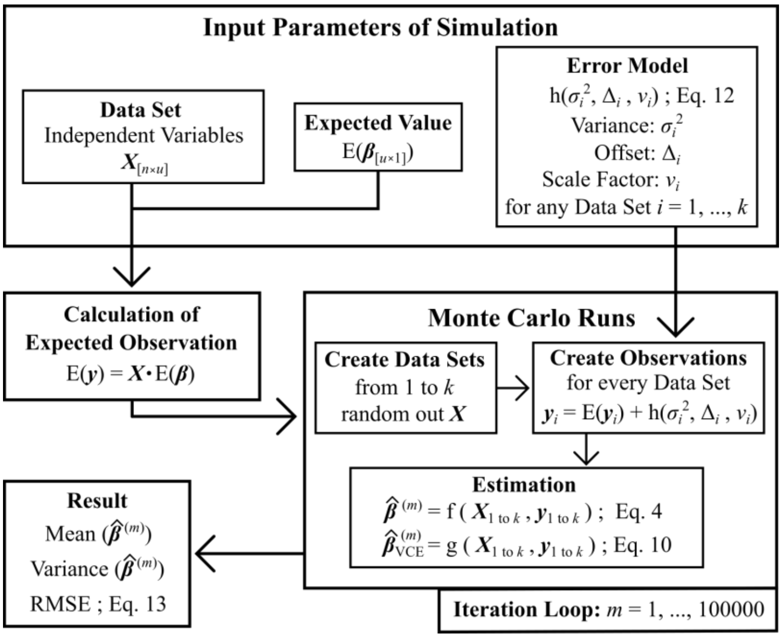

Figure 1: Schematically sequence of the closed loop simulation

All these parameters are initial values for the closed loop simulation. At first, every loop generates k random data sets from $\boldsymbol{X}$. In the second step, the errors of every data set is generated with Eq. 12

and it is added to the expected observations $\mathrm{E}(\boldsymbol{y}_{1\text{ to }k})$. The variance $\sigma^2_{1\text{ to }k}$ is used to generate a normal distributed noise $\boldsymbol{\kappa}$. This should characterize the heterogeneous dispersion of the data sets. The scale factor $\upsilon_{1\text{ to }k}$ leads to a percentage increase of the expected observation, thus the observations of a data set are skewed systematically to the other data sets. The offset is a systematic error in a data set and has a direct influence on the coefficient β_0.

$$h(\sigma_i^2, \Delta_i, \upsilon_i) = \mathrm{E}(\boldsymbol{y}_i)\,\frac{\upsilon_i}{100} + \Delta_i + \boldsymbol{\kappa}\ ;\ \ \boldsymbol{\kappa} \sim N(0, \sigma_i^2) \qquad i = 1, \ldots, k \qquad \text{Eq. 12}$$

In the third step, the three simulated data sets with the noise observation are used to estimate the coefficients $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_{VCE}$. The coefficients $\widehat{\boldsymbol{\beta}}$ are determined without VCE, thus, all data sets have the same weights in the estimation. In the case of $\widehat{\boldsymbol{\beta}}_{VCE}$, the VCE is used in the calculation. These three steps are repeated 100’000 times in the loop. The number of iteration loops determines the precision of the simulation result. Repeats of this simulation show that the different results approximately equal and therefore we fixed the number of iterations to 100’000 iteration loops, which are sufficient for this study. The results of the closed loop simulation include the means and variances of the 2 estimations and their 100’000 repeats. Furthermore, the root-mean-square error (RMSE) is calculated with the predicted observations and the expected observations.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum\left(\widehat{\boldsymbol{y}} - \mathrm{E}(\boldsymbol{y})\right)^2} \qquad \text{Eq. 13}$$

The presented closed loop simulation shall be extended in future researches, e.g. integration of outliers in the error model. Basics and further information about closed loop simulation and Monte Carlo are presented in Saltelli et al. (2008) and Kroese et al. (2011). Alkhatib (2007) demonstrates a practice example for satellite gravity missions.

4. USED DATA SETS

The data set which is used to validate our approach had been collected in the city of Nienburg (Weser). Nienburg (Weser) is located in the south of Lower Saxony. It is a small city with approximately 50’000 inhabitants. As the functional submarket, we investigated the market of one and two-family houses. We use three different types of data:

- Purchase prices
- Experts’ knowledge
- Offer prices

The spatial submarket of Nienburg (Weser) is located in a region with a regular supply and demand situation and an average number of transactions in real estate market of individual housing (single and two-family houses). For this region, good market information exists. For this approach, 242 purchase prices from the official purchase price database are used. This dataset is reduced to simulate a region with few transactions. Random sample of 30 purchase prices out of the 242 prices are generated. A detailed information on different reduction procedures to simulate regions with few purchases can be found in Weitkamp & Alkhatib (2014). The second dataset contains an experts’ survey from September 2015. Experts estimate market values for different objects.

Therefrom, 180 pseudo prices were obtained. In this approach, 130 of these pseudo prices are used for fitting the features of the purchase price data set. As third data set, 109 offer prices from a real estate offer portal (immobilienscout24.de) are used from the years between 2011 until 2015.
Detailed information on derivation procedure and first investigation on all three data sets can be found in Soot et al. 2016.
All three data sets were adjusted to the same economic situation (on date: September 2015). For all data sets the information “area of lot” [sq. m], “standard land value” [EUR per sq. m], “construction year” [age – 1946], “living space” [sq. m] and “equipping standard” [without unit] are available. The dependent variable $\boldsymbol{y}$ is the “purchase price per square meter living space” in EUR per sq. m.
The multiple linear regression is done after investigation of statistical outliers. We use a Baarda data-snooping algorithm (Baarda 1986) with a limit of 2.5 σ for the normalized error. The results of the multiple linear regression are shown in Table 2.
One can see that the result between purchase prices and experts’ survey differs significantly in parameters for area of lot, construction year and equipping standard. The estimated parameters for intercept and living space differ widely for offer prices. This can partly be explained with the correlation between parameters. Model depended the equipping standard (β_5) is correlated with the intercept (β_0) (correlation coefficients: offer prices = – 0.8, purchase prices = – 0.7, experts’ survey = – 0.5). As expected for this model, the area of lot and the living space are correlated (≈ 0.5).

Table 2: Results of the multiple linear regression for the three data sets.

	β_0 Intercept	β_1 Living space	β_2 Area of lot	β_3 Construction year	β_4 Standard land value	β_5 Equipping standard
Purchase Prices	506.65	– 4.89	0.31	15.40	3.41	110.46
Experts’ Survey	507.28	– 4.71	0.17	9.89	3.09	221.45
Offer Prices	346.91	– 2.68	0.31	11.88	3.38	135.45

The parameter for the equipping standard differs according to the particular data set. This is caused by the poor quality of this variable in the data sets of offer prices and purchase prices. If detailed information about the sold real estate is missing, the equipping standard is set to an average value, even if the real estate is a very well equipped house. Future work will deal with this problem using more information to improve the classification of the real estates in the equipping standard classes.
Following the answers of the experts the influence of the equipping standard on the market value is high. In the data set of the purchase price database as well as of the offer price data set, the influence of the equipping standard is approximately half of experts’ opinion. Unlike the information in the databases, the experts can judge about the equipping standard in a very good way. And properly arisen, the equipping standard has a great influence on the market value. Due to the circumstance, that the equipping standard in the database do not vary, no great influence affects the market value.

4.1 Quality of Different Data Sets

The precession of the data is derived from the multiple linear regression. The results of the posterior standard deviation from the adjusted parameters are shown in Table 3.

Table 3: Input parameter for the error model of the closed loop simulation.

	Posterior σ	Offset	Scale Factor
Purchase Prices	198.24	0	0 %
Experts' Survey	93.85	20	0 %
Offer Prices	266.83	180	– 15 %

The information from the experts' survey has the smallest posterior standard deviation. Therefore, it can be concluded that these data fit best to each other. The data set from the purchase price database is still a typical imperfect data set. This is caused by the different influencing parameters on the origin of prices, which are not part of our model.

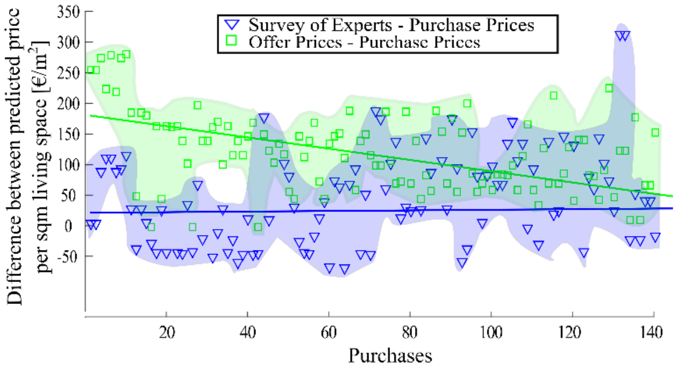

Figure 2: Difference between the predicted observations with independent variables from the experts' survey. Blue: Difference between experts' survey and purchase prices. Green: Difference between offer prices and purchase prices.

The largest posterior standard deviation arises from the offer prices, as expected. In this approach, the offset is derived from first analysis on this data set. The data from the real estate platform, which are often posted by real estate agents, are regularly scattering above the market value. A maximum offset of approximately 180 EUR per sq. m living space between offer prices and purchase prices is derived (see Figure 2). This offset decreases for high quality real estates. A systematic modeling of this offset is planned in our next investigations. In this study, the average offset of offer prices is only considered with 80 EUR per sq. m.
From this investigation follows the error model for the closed loop simulation (see Table 3). The decreasing offer prices consider with a scale factor of –15% and the offsets derive from the y–axis. The noise of the different data sets is generated with the posterior standard deviations.

5. VALIDATION OF THE APPROACH

5.1 Closed Loop Simulation

The closed loop simulation uses the presented input parameters of Section 4 for the calculation. In Figure 3, the result of the construction year is presented for 100’000 iterations. The classical regression is illustrated in red and the result of VCE is presented in blue. The solid lines represent the mean values of the histogram and the dash lines highlight the 95% confidence interval. The difference between mean values of the construction year is approximately 0.75. Hence, the mean value of regression lies marginal into the 95% confidence interval of the VCE. A significant improvements using VCE approach are almost the narrower confidence interval. The results of the other regression coefficient look very similar.

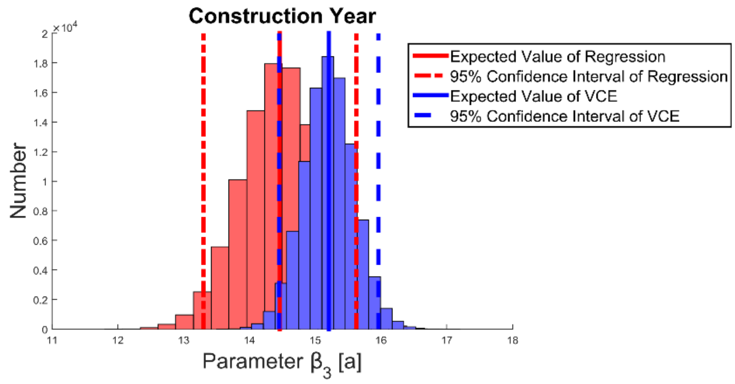

Figure 3: Distribution of the estimated construction year for the simulation with 100’000 iterations. Red: Result of the regression. Blue: Result of the VCE.

Table **4** shows an overview for all results of the simulation. The columns “Coefficient Difference to E(β_i)” represent the difference between the mean value of the estimated coefficients and the expected value E($\boldsymbol{\beta}$)) of the input parameter (from the simulation). Small differences means the estimated coefficients are determined correct. In the columns of “standard deviation σ“ the average precision of the regression coefficient is presented. The comparison between these results of classical regression and VCE shows that the results of VCE are always closer to the expected values. This could be recognised in the RMSE, too. The RMSE of VCE is approximately half as large as the RMSE of regression. Additional, the difference of intercept shows that the offset of the experts’ survey and the offer prices have an influence with approximately 26 EUR per sq. m for the VCE.

Table 4: Results of the closed loop simulation with 100’000 iterations.

Parameter	**Classical Regression**		**Regression with VCE**	
	Coefficient Difference to E(β_i)	**σ [SD]**	**Coefficient Difference to E(β_i)**	**σ [SD]**
β_0 [Intercept]	51.86	75.00	26.31	47.67
β_1 [Living Space]	0.30	0.40	0.06	0.25
β_2 [Area of Lot]	– 0.02	0.05	0.00	0.03
β_3 [Construction Year]	– 0.94	0.60	– 0.19	0.38
β_4 [Standard Land Value]	– 0.21	0.38	– 0.04	0.24

β_5 [Equipping Standard]	– 6.65	14.73	– 1.40	9.37
RMSE	41.31		21.91	

Table 5 presents the average square root of the estimated variance components. In case of classical regression, the posterior standard deviation with 200.26 is approximately equal to the mean of input simulated values σ depicted in Table 3. This result was expected, because every data set has the same influence in the estimation. In the case of VCE the variance results of purchase prices and experts' survey are approximately the same as the input variance components of the error model. Only the offer prices have a greater difference with approximately 17.41 to the input parameter. This results from the scale factor, which distort the offer prices.

Table 5: Estimated posterior square root of variance components for the closed loop simulation.

	Classical Regression	Regression with VCE		
		Purchase Prices	Survey	Offer Prices
Posterior $\hat{\sigma}$	200.26	198.26	94.53	284.24

5.2 Real Market Data

For the investigation of real data, an area with few transactions is simulated. Therefore, the experts' survey and offer prices are combined with 30 random purchase prices. This simulation is repeated 100'000 times. Figure 4 presents the result of the parameter intercept and construction year. The confidence intervals of the real market data have a comparable dispersion as in the simulation. In contrast to the close loop simulation, here the difference between the mean values of regression and VCE is larger. The results of the other regression coefficient look very similar, as shown in Table 6.

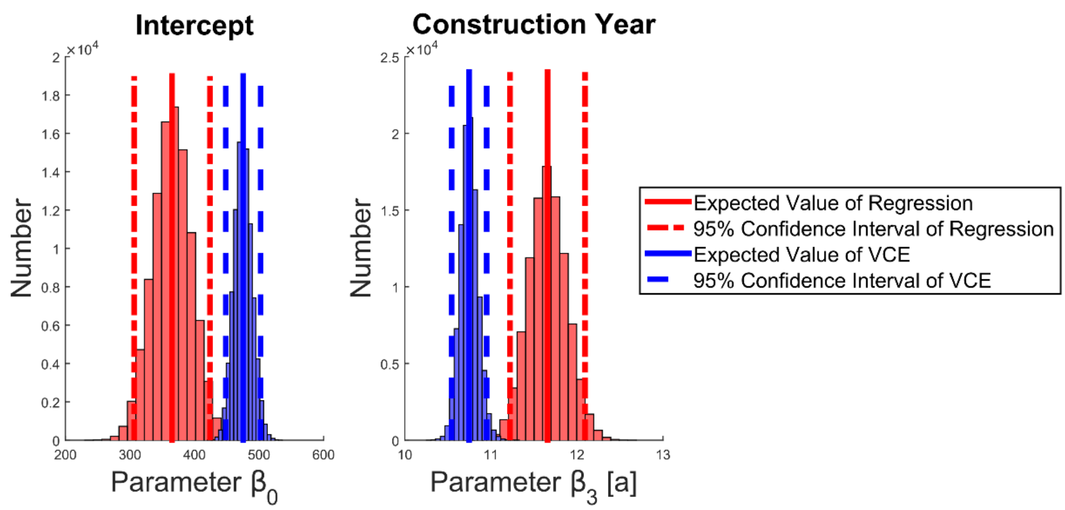

Figure 4: Distribution of the estimated intercept and construction year for 100'000 simulated areas with few transactions. Red: Result of the regression. Blue: Result of the VCE.

The comparison between regression and VCE in areas with few transactions shows, that the parameters of area of lot, construction year and equipping standard as result of the regression are closer to the coefficients of purchase price data set. Intercept, living space and standard land value behave different, however. The coefficients of VCE are closer to the purchase price data set in this case. This is caused by the difference between the coefficients of purchase price data set and experts' survey data set. An explanation could be found in the correlations between the coefficients like discussed in Section 4.

Table 6: Mean values of the simulated area with few transactions with 100'000 iterations.

Parameter	**Results of Purchase Prices**	**Results of Area with Few Transactions**			
		Classical Regression		**Regression with VCE**	
	Coefficient	**Coefficient**	$\boldsymbol{\sigma}$ **[SD]**	**Coefficient**	$\boldsymbol{\sigma}$ **[SD]**
β_0 [Intercept]	506.65	365.18	83.02	475.30	54.99
β_1 [Living Space]	– 4.89	– 3.75	0.40	– 4.48	0.30
β_2 [Area of Lot]	0.31	0.24	0.05	0.19	0.04
β_3 [Construction Year]	15.40	11.66	0.81	10.75	0.57
β_4 [Standard Land Value]	3.41	3.65	0.41	3.40	0.26
β_5 [Equipping Standard]	110.46	161.13	19.97	192.20	14.07

The estimated variance components are presented in Table 7. For the regression, the posterior $\hat{\sigma}$ with 198.72 is equivalent to the result of simulation with 200.26 (see Table 5). In the case of VCE, the $\hat{\sigma}$ of purchase data set is larger than the estimated value in the simulation and thus, it is larger than the variance in the original data sets. As an explanation for this effect, we assume that the simulated purchase data set are homogeneous as in the reality, more detailed research would done in a future work. The variance component of experts' survey with 95.89 is approximately the same as in the original data set with 93.85 (see Table 3). Furthermore, the posterior $\hat{\sigma}$ of VCE shows the consideration of each data set in the estimation. The experts' survey considers more than the other two data sets in the estimation. This result reflects the regression coefficient in Table 6.

Table 7: Estimated posterior square root of variance components for the area with few transactions.

	Classical Regression	**Regression with VCE**		
		Purchase Prices	**Survey**	**Offer Prices**
Posterior $\hat{\sigma}$	198.72	217.91	95.89	278.44

6. CONCLUSION AND OUTLOOK

The results of VCE in the close loop simulations show that the offset in the error model has a small influence on the intercept. Furthermore, the VCE considers the influence of the scale factor with a larger variance component. Hence, the simulation demonstrates that low systematic errors between the data sets lead to acceptable expected values of regression coefficients for the real estate valuation. The comparison of the simulation results with the real data results shows, that in future,

the simulation requires an improved error model. This should be allow to reproduce a larger inhomogeneity of data sets.
The investigation with the real data demonstrates a differentiated result in comparison of regression and VCE. Comparing the three coefficients, the ones of regression are closer to the purchase coefficients and the ones of VCE are closer to the ones of purchases in the other three. This depends on the regression coefficients of the experts' survey. The area of lot, the construction year and the equipping standard of survey are too different to the coefficient of purchase prices. Possible reasons for this effect are discussed in Section 4. In future, the experts' survey requires objects as base for the appraisal, which better represent the real data distribution in these three dependent variables.
Overall, this first investigation shows a high potential of VCE in contrast to an equal weighting of all data sets. The VCE correctly identifies the precision of the different data sets. A huge systematic price offset between the different data sets cannot be considered by the VCE. Therefore, the data sets have to be adjusted to each other in future investigations. The use of offer prices enables the integration of inhomogeneity of the real market in the estimation, which the experts' survey data set cannot represent. That means, in case of estimating without offer prices would be equally to the coefficient of the experts' survey. In this case, the coefficients of area of lot, construction year and equipping standard will cause larger differences to the purchase price coefficients.
The next studies should be focused on the optimisation of the VCE approach for real estate valuation. Therefore, this approach should be applied in other submarkets to analyse the general suitability for areas with few transactions. In a second step, the VCE approach should be integrated in a robust Bayesian approach.

Acknowledgement
The investigations published in this article are granted by the DFG (German Research Foundation) under the sign <60451047>. The authors cordially thank the funding Agency. Besides, we thank the surveying administration of Lower Saxony and especially the experts' committee of Nienburg for the provided data and the participation in the survey.

REFERENCES

Alkhatib, H. (2007). On Monte Carlo methods with applications to the current satellite gravity missions. PhD thesis, Hochschulschriftenserver der ULB Bonn, http://hss.ulb.uni-bonn.de/diss online, Bonn.
Alkhatib, H. & Weitkamp, A. (2012). Bayesischer Ansatz zur Integration von Expertenwissen in die Immobilienbewertung, Teil 1. ZfV, 137(02), pp. 93–102.
Alkhatib, H. & Weitkamp, A. (2013): Robust Bayesian Regression Approach for Areas with Small Numbers of Purchases. In: Royal Institution of Chartered Surveyors (Ed.): Proceedings of COBRA Conference, New Delhi 2013.
Amiri-Simkooei, A.R. (2007). Least-squares variance component estimation: theory and GPS applications. PhD thesis, Delft University of Technology, Publication on Geodesy, 64, Netherlands Geodetic Commission, Delft, The Netherlands.

Baarda W., 1968. A testing procedure for use in geodetic networks. PhD thesis, Delft University of Technology, Publications on Geodesy, 2(5), Netherlands Geodetic Commission, Delft, The Netherlands.

Fahrmeir, L., Kneib, T. & Lang, S. (2009). Regression: Modelle, Methoden und Anwendungen (2nd ed.). Heidelberg: Springer.

Geweke, J. (1993). Bayesian Treatment of the Independent Student-t Linear Model. Journal of Applied Econometrics (8): 19–40.

Hartung, J., Elpelt, B. and Klösener, K.-H. (2009). Statistik: Lehr- und Handbuch der angewandten Statistik; [mit zahlreichen durchgerechneten Beispielen], 15th edn. München: Oldenbourg.

Kleiber, W., Fischer, R., & Werling, U. (2014). Verkehrswertermittlung von Grundstücken: Kommentar und Handbuch zur Ermittlung von Marktwerten (Verkehrswerten) und Beleihungswerten sowie zur steuerlichen Bewertung unter Berücksichtigung der ImmoWertV (7., vollst. neu bearb. Aufl.). Köln: Bundesanzeiger Verlag

Kacker, R. & Jones, A. (2003). On use of Bayesian statistics to make the Guide to the Expression of Uncertainty in Measurement consistent. Metrologia, 40: 235–48.

Koch, K.-R. (1997). Parameterschätzung und Hypothesentests in linearen Modellen. 3. Auflalge, Dümmler, Bonn.

Koch, K.-R. (2007). Introduction to Bayesian Statistics. 2. Edition, Heidelberg: Springer-Verlag Berlin.

Koch, K.-R. & Kusche, J. (2002). Regularization of geopotential determination from satellite data by variance components. Journal of Geodesy (2002) 76: 259–268

Kroese, D.P., Taimre, T. & Botev. Z. (2011). Handbook of Monte Carlo Methods. Published by John Wiley & Sons, Inc., Hoboken, New Jersey.

Pelzer, H. (1978). Ein indirektes Vergleichswertverfahren unter Anwendung statistischer Methoden, ZfV 6/1978, pp. 245–254.

Reuter, F. (2006). Zur Ermittlung von Bodenwerten in kaufpreisarmen Lagen. Flächenmanagement und Bodenordnung, 68(3), pp. 97–107.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J. Gatelli, D., Saisana, M. & Tarantola, S. (2008). Global Sensitivity Analysis. The Primer. John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England.

Soot, M., Weitkamp, A., Alkhatib, H., Dorndorf, A. & Jeschke, A. (2016). Analysis on Different Market Data for Real Estate Valuation – Investigations on German Real Estate Market. FIG Working Week 2016, Christchurch, New Zealand.

Uhde, C. (1982). Mathematische Modelle zur Analyse von Grundstücksmärkten. PhD thesis, Technische Universität Hannover, Hannover.

Urban, D. & Mayerl, J. (2011). Regressionsanalyse: Theorie, Technik und Anwendung. 4. Auflage . VS Verlag für Sozialwissenschaften, Springer Fachmedien Wiesbaden GmbH

Weitkamp, A. & Alkhatib, H. (2012a). Bayesischer Ansatz zur Integration von Expertenwissen in die Immobilienbewertung, Teil 2. ZfV, 137(02), pp. 103–114.

Weitkamp, A. & Alkhatib, H. (2012b). The Bayesian approach in the valuation – a strategy to handle markets with low purchasing prices?. In: FIG Working Week 2012, Rome, Italy.

Weitkamp, A. & Alkhatib, H. (2014). Die Bewertung kaufpreisarmer Lagen mit multivariaten statistischen Verfahren: Möglichkeiten und Grenzen robuster Methoden bei der Auswertung weniger Kauffälle. AVN (Allgemeine Vermessungs-Nachrichten), 121(01), pp. 3–12.
Yang, Y., Xu, T., Song L. (2005) Robust Estimation of Variance Components with Application in Global Positioning System Network Adjustment. J. Surv. Eng. 131(4):107-112.
Ziegenbein, W. (1977). Zur Anwendung multivarianter Verfahren der mathematischen Statistik in der Grundstückswertermittlung. PhD thesis, Technische Universität Hannover, Hannover.
Ziegenbein, W. (2010). Immobilienwertermittlung. In: Kummer, K.; Frankenberger, J. (Hrsg.): Das deutsche Vermessungs- und Geoinformationswesen 2010, 421-468.

BIOGRAPHICAL NOTES

Alexander **Dorndorf** received his Master of Science in “Geodesy and Geoinformatics” at the Leibniz Universitaet Hannover in 2014. Since then he has been at the Geodetic Institute at the Leibniz Universitaet of Hannover. His main research interests are: Bayesian Statistics, Monte Carlo Simulation and modelling of measurement uncertainty.

Matthias **Soot** received his Master of Science (M.Sc.) in Geodesy at the Technical University of Dresden in 2014. For half a year, he worked as a valuation expert in free economy. Since March 2015, he is working at the Geodetic Institute of the Technical University of Dresden at Chair of Land Management. His research focus is on statistical analysis of different market information and development of purchasing price databases.

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.

Prof. Dr.-Ing. Alexandra **Weitkamp** received her diploma (Dipl.-Ing.) in Geodesy at the University of Hannover in 1999. She passed the highest level state certification as “Graduate Civil Servant for Surveying and Real Estates” in Lower Saxony in 2001. After two-year experience at Bayer AG, she returns to Leibniz Universitaet Hannover. In 2008, she received her Ph.D. in “Geodesy and Geoinformatics” at the University of Bonn. Until 2014, she has been postdoctoral fellow at the Geodetic Institute at the Leibniz Universitaet Hannover. In October 2014, she became Chair of Land Management at Technical University of Dresden. Her main research interests are: adaption of innovative evaluation methods for valuation, stakeholders behavior in rural and urban development, and decision-making methods.

CONTACTS

Alexander Dorndorf (M.Sc.) Geodetic Institute of Leibniz Universitaet Hannover – Evaluation methods Nienburger Str. 1 Hannover – 30167 GERMANY Tel. +49 (0) 511 762 2464 Fax +49 (0) 511 762 2468 Email dorndorf@gih.uni-hannover.de Web site: http://www.gih.uni-hannover.de/	Matthias Soot (M.Sc.) TU Dresden – Geodetic Institute Helmholzstraße 10 Dresden – 01069 GERMANY Tel. +49 (0) 351-46331653 Fax + 49 (0) 351-46337190 Email: matthias.soot@tu-dresden.de Web site: http://www.tu-dresden.de/gi/lm
Dr.-Ing. Hamza Alkhatib Geodetic Institute of Leibniz Universitaet Hannover – Evaluation methods Nienburger Str. 1 Hannover – 30167 GERMANY Tel. +49 (0) 511 762 2464 Fax +49 (0) 511 762 2468 Email alkhatib@gih.uni-hannover.de Web site: http://www.gih.uni-hannover.de/	Prof. Dr.-Ing. Alexandra Weitkamp TU Dresden – Geodetic Institute Helmholzstraße 10 Dresden – 01069 GERMANY Tel. +49 (0) 351-46337115 Fax + 49 (0) 351-46337190 Email: alexandra.weitkamp@tu-dresden.de Web site: http://www.tu-dresden.de/gi/lm