www.degruyter.com/view/j/remav

ECONOMETRIC MODELS OF REAL ESTATE PRICES WITH PRIOR INFORMATION. MIXED ESTIMATION

Mariusz Doszyń
Institute of Economics and Finance
University of Szczecin, Poland
e-mail: mariusz.doszyn@usz.edu.pl

Abstract

The purpose of this paper is to estimate econometric models with sample and prior information. Prices of land property for residential development in Szczecin are modeled (the price level was determined for 2018). Modeling property prices only based on sample data generates numerous problems. Transaction databases from local real estate markets often contain a small number of observations. Properties are frequently similar, which results in low variability of property characteristics, and thus – low efficiency of parameter estimators. In such a situation, the impact of some features cannot be estimated from the sample data. As a solution to this problem, the paper proposes econometric models that consider prior information. This information can be, for example, in the form of property feature weights proposed by experts. The prior information will be expressed in the form of stochastic restrictions imposed on the model parameters. In the simulation experiment, the predictive power of mixed estimation models is compared with two kind of models: OLS models and model with only prior information. It turned out that mixed estimation results are superior with regard to formal criteria and predictive abilities.

Key words: *econometric models of real estate prices, mixed estimation, Theil – Goldberger estimator, prior information, land prices prediction*.

JEL Classification: *C15, C18, C50, C51, C52.*

Citation: Doszyń, M. (2022). Econometric models of real estate prices with prior information. Mixed estimation. *Real Estate Management and Valuation, 30*(3), 61-72.

DOI: https://doi.org/10.2478/remav-2022-0021.

1. Introduction

Transaction databases from local real estate markets are often of low quality (from an econometric modeling perspective). The number of observations is low, the variability of property characteristics is low, etc. Sometimes, property characteristics show no variability at all, especially in the primary market. For example, all land properties sold may be characterized only by favorable location, favorable transportation accessibility, etc. In such a case, it is impossible to estimate the effect of these features on price based on sample information alone.

In addition, the features of a property are often qualitative in nature, so they are not objectively measurable. Only the states of these characteristics can be determined, which involves subjectivity. Even real estate experts may perceive the same feature states differently. This is important if databases are created by combining information from different sources. It is then not certain that, in each case, the states of qualitative features were defined in the same way.

This paper proposes an econometric way of solving the problems caused by low quality of statistical information. The quality of information can be improved by using prior information from

experts who know the market well. Prior knowledge can be in the form of weights (or weight intervals) for property characteristics, based on which the ranges of parameters of the econometric model can be determined. Based on the intervals for the model parameters, the prior expected values and variances can be determined, which are used in mixed estimation. Mixed estimation combines sample information with prior knowledge about model parameters that depend on the characteristics of the modeled real estate market.

2. Literature review

There are many applications of the quantitative method in modeling real estate prices, especially in the context of mass appraisal (Jahanshiri et al., 2011; Pagourtzi et al., 2003; McCluskey et al., 2013). Econometric models of prices were considered, e.g., in Isakson (1998), Dell (2017), Kokot and Gnat (2019). Parametric, semiparametric, and nonparametric estimators of that kind of models were discussed in Pace (1995). Different analytical forms of models used in predicting real estate values were applied in (Barańska & Łuczak, 2007). Econometric models of property prices often contain spatial effects (Fik et al., 2003; Cellmer et al., 2020). For massive datasets, machine learning methods are sometimes recommended (Zurada et al., 2011; Sing et al., 2021).

The extensive research on modeling undeveloped land prices in Szczecin, from the point of view of mass valuation, is contained in the monograph (Doszyń, 2020a). This monograph presents a system for determining the impact of property characteristics on their value in the so – called Szczecin algorithm of real estate mass appraisal. A detailed description of the Szczecin algorithm can be found, e.g., in (Hozer et al., 1999).

So far, there are not many applications of mixed estimation in modeling real estate prices in the literature. Such attempts are contained in the articles of Doszyń(2022, 2021).

In Doszyń (2022), the effect of using prior information in estimation of real estate mass appraisal models was studied. Six econometric models were compared: ordinary least squares (OLS) model, mixed estimation, Bayesian model, Inequality Restricted Least Squares (IRLS) model, ridge regression, and LASSO (with regularization). Models with prior information were found to be superior.

Mixed estimation was used in Doszyń (2021) to model land property prices. The results were compared with those obtained by OLS. The mixed estimation model provided estimates consistent with theoretical expectations. Property price predictions were also better (than with the OLS model).

The possibility of using prior information in estimating parameters of econometric models was first proposed in Durbin (1953). This involves a non–Bayesian approach based on the frequency probability theory.

A systematic and detailed description of mixed estimation was given in Theil and Goldberger (1961) and Theil (1963). In these papers, a generalized least squares method was proposed for mixed estimation. The problems associated with the unknown variance of random disturbances were analyzed in detail. A special class of mixed estimators (f – class mixed estimators) was proposed. Verification of consistency of prior and sample information was also discussed. A formula for the disturbance variance estimator considering prior information was proposed. Measures of the shares of prior and sample information in posteriori distribution were also suggested. Some differences between the mixed estimator and the Bayesian approach were explained. Properties of the mixed estimator were analyzed in Swamy and Mehta (1969), Mehta and Swamy (1970) and Nagar and Kakwani (1964).

The article Mittelhammer and Conway (1988) contains a criticism of the Theil – Goldberger mixed estimator. The reason is the lack of strict rules for considering prior information. The authors proposed a more formalized approach to considering prior information in the form of the Prior Integrated Mixed Estimator (PIME). The authors show that PIME dominates the Theil – Goldberger estimator. The paper discusses many advantages of mixed estimation (e.g., reduction of multicollinearity). It was pointed out that mixed estimation can be treated as a special case of ridge regression. Mixed estimation also allows for the verification of the correctness of prior beliefs.

Prior information about model parameters might be also expressed as restrictions in the form of inequalities. Inequality restricted least squares (IRLS) models were discussed, e.g., in Grömping (2010). The IRLS models of real estate mass appraisal were presented in Pace and Gilley (1990) and Doszyń (2020b).

3. Data and Methods

3.1. Sample information

The database contains information about 40 transactions of land properties for residential development in Szczecin (2018 price level). Information on the following features is available: land area, utilities, neighborhood, transport availability, physical plot properties (geometric features) and location. All features, except area, are qualitative, expressed on an ordinal scale with three states. The states are coded as follows: 0 – unfavorable, 1 – average, 2 – favorable. For utilities, the states are defined as follows: 0 – none, 1 – incomplete, 2 – complete. In general, feature states are ordered from worst to best.

Tables 1 and 2 present basic descriptive statistics for price and property characteristics. In Table 1, statistics for quantitative variables (price, land area) are presented. In Table 2 statistics for qualitative real estate features are shown.

Table 1

Basic statistics for unit price and area (quantitative features)

Statistics	Price (zł/m2)	Land area (m2)
Min	133.70	599.00
First quartile	232.70	721.00
Median	271.75	903.50
Mean	255.96	1085.15
Third quartile	285.77	1330.25
Max	361.11	1977.00

Source: own study.

The unit price of land ranged from 134 – 361 PLN/m², with a median of 272 PLN/m². The land area ranged from about 600 – 2000 m², with a median of 903.5 m², so the plots were rather small.

Table 2

Shares of feature states (qualitative features)

Feature state	Utilities	Neighborhood	Transport availability	Physical properties	Location
0	0.075	0.150	0.650	0.125	0.000
1	0.100	0.850	0.350	0.775	0.150
2	0.825	0.000	0.000	0.100	0.850
Sum	1.000	1.000	1.000	1.000	1.000

Source: own study.

Regarding the qualitative features (Table 2), there was one dominating state for each feature. So, 82.5% of the properties had favorable utilities, 85% of the properties had an average neighborhood, 65% had unfavorable transport availability, 77.5% had average physical plot properties and 85% of properties had a favorable location.

It is worth noting that there were no properties with a favorable state for two of the features (neighborhood, transport availability). On the other hand, in the case of location, no unfavorable state appeared. It is not possible to estimate the impact of these feature states from the sample information. Zero–one variables for the feature states that do not appear in the database do not show variability, as their value is zero for each property. The sample information is not enough to estimate the impact of these variables on the unit price of the property. Hence, for three property features (neighborhood, transport availability, location), there are two states (not three) in the database. The impact of absent feature states might be estimated by using prior information, by means of a mixed estimation procedure.

3.2. Prior information

The following econometric model of property unit price will be estimated

$$lnp_i = \alpha_0 + \alpha_1 a_i + \alpha_{21} u_{1i} + \alpha_{22} u_{2i} + \alpha_{31} n_{1i} + \alpha_{32} n_{2i} + \alpha_{41} ta_{1i} + \alpha_{42} ta_{2i} + \alpha_{51} pp_{1i} + \alpha_{52} pp_{2i} + \alpha_{61} l_{1i} + \alpha_{62} l_{2i} + u_i \quad (1)$$

where:

p_i – unit price of i – th property, $i = 1, 2, \ldots, n$, n – number of properties,
a_i – land area of i – th property,
– utilities: u_{1i} – incomplete, u_{2i} – complete,
– neighborhood: n_{1i} – average, n_{2i} – favorable,
– transport availability: ta_{1i} – average, ta_{2i} – favorable,
– physical properties: pp_{1i} – average, pp_{2i} – favorable,
– location: l_{1i} – average, l_{2i} – favorable,
$\alpha_0, \alpha_1, \ldots\ldots, \alpha_{62}$ – parameters,
u_i – error term.

Model (1) is log – linear. Qualitative feature states were entered into the econometric model as zero – one variables ("1" indicates the presence of a given feature state). We only have information that a given state is present, without its intensity. Intensity is present for the "whole" qualitative feature, by means of given states. E.g. we know that a favorable location is better (has higher "intensity") than an average location. Zero – one variables were omitted for worst feature states because there is a constant term in the model. The impact of the worst states is present in the constant term. For qualitative features, zero – one variables for average and best states were added. On the other hand, land area was introduced into the model as a single variable, because it is a quantitative feature.

The prior information takes the form of weights assigned to each feature:
– land area – 15%,
– utilities – 10%,
– neighborhood – 20%,
– transport availability – 15%,
– physical properties – 10%,
– location – 30%.

There is always the question of the weight validity. These were proposed by real estate appraisers who "knew" the local market.

Model (1) contains parameters, not weights, hence some kind of transformation is needed. A feature weight might be transformed into model parameter by means of the following formula (the author's simple proposal):

$$\alpha_{kp} = w_k \frac{p}{p_k} ln\left(\frac{p_{max}}{p_{min}}\right) \quad (2)$$

where:

α_{kp} – impact of state p of feature k, $p = 0, 1, 2,$
w_k – weight of feature k,
p_k – number of states of given feature k,
p_{max} – the unit price of a property with the most favorable feature states,
p_{min} – the unit price of property with the least favorable feature states.

Formula (2) is related to model (1) only with regard to the analytical (exponential) form. All qualitative features have three states coded as "0, 1, 2", hence, for each feature $p_k = 2$. Generally, there are three states for each feature, but the worst states are omitted, thus there are two zero – one variables for each feature in the econometric model.

There are no properties in the database with the worst feature states. There is also no property in the case of which all features have the most favorable states. The values of these properties can be determined based on expert opinion. It has been assumed that the range of unit prices is between 100 – 400 PLN/m^2. Prior information about the weights and parameters has been presented in Table 3.

Weights of feature states were determined assuming that "transitions" between successive states are linear. Let us assume that the weight for transport availability is equal to 0.2. Then, the weight for average transport availability is 0.1, and for favorable transport availability – 0.2. The case is similar for other qualitative features. The worst feature states are omitted because there is intercept in the

model, hence weights for the worst feature states are always equal to zero. Impact of the worst states is contained in the intercept.

Table 3

Prior information about feature weights and parameters

Variables	Weights	Weight ranges		Expected value of a parameter	Parameter ranges		Standard deviation	Variance
const	-	-	-	4.927	4.905	4.949	0.0109	0.0001
a_i	0.150	0.100	0.200	-0.00015	-0.0001	-0.0002	0.0000	0.0000
u_{1i}	0.050	0.000	0.100	0.069	0.000	0.139	0.0347	0.0012
u_{2i}	0.100	0.050	0.150	0.139	0.069	0.208	0.0347	0.0012
n_{1i}	0.100	0.050	0.150	0.139	0.069	0.208	0.0347	0.0012
n_{2i}	0.200	0.150	0.250	0.277	0.208	0.347	0.0347	0.0012
ta_{1i}	0.075	0.025	0.125	0.104	0.035	0.173	0.0347	0.0012
ta_{2i}	0.150	0.100	0.200	0.208	0.139	0.277	0.0347	0.0012
pp_{1i}	0.050	0.000	0.100	0.069	0.000	0.139	0.0347	0.0012
pp_{2i}	0.100	0.050	0.150	0.139	0.069	0.208	0.0347	0.0012
l_{1i}	0.150	0.100	0.200	0.208	0.139	0.277	0.0347	0.0012
l_{2i}	0.300	0.250	0.350	0.416	0.347	0.485	0.0347	0.0012

- Not applicable

Source: own study.

In the mixed estimation, stochastic restrictions are imposed on the parameters, which means variability of the parameters. Therefore, the weights are also expressed in the form of intervals (Columns 3 and 4 in Table 3). The weight intervals have a length of 0.1 in each case. The weights (in Column 2) differ from the ends of the weight intervals by +/- 0.05. Relying on the weight intervals allows us to account for the uncertainty in their values.

The expected values of the parameters are given in column 5. The expected value of the intercept is the natural log of unit minimum price: $ln(p_{min}) = ln(138)$. This price was given by an expert. The interval for the intercept is $\langle 135; 141 \rangle$, which after logarithmization gives the ends of the interval for the intercept: $\langle 4.905; 4.949 \rangle$.

For land area, the expected value and the ends of the interval were determined in a different way than for qualitative variables. Land area is a quantitative variable with an interval of about 1400 m^2 (Table 1). Negative value of the parameter was adopted, because the influence of land area on unit price is negative. The expected value of the parameter for land area was defined as $-0.15\, ln(p_{max}/p_{min})/1400$, where 0.15 is the weight for the land area and 1400 m^2 is the area spread in the database. The ends of the interval for land area were determined in a similar manner, taking the ends of the interval for weight equal to 0.1 and 0.2, respectively.

For the remaining parameters (parameters with qualitative variables), the expected values and ends of the parameter intervals were calculated based on (2). In determining the expected values of the parameters, the weights from column 2 were taken (Table 3). In turn, the ends of the weight intervals (columns 3 and 4) were substituted when calculating the ends of the parameter intervals.

All parameters were assumed to have normal distribution with intervals containing the parameters with a confidence of 0.95. It follows that the ends of the interval deviate from the expected value of the parameter by two standard deviations. Let us denote the parameter interval as $\langle a; b \rangle$. The standard deviation is then equal to $(b - a)/4$. The standard deviations are in the penultimate column of Table 3. The parameter variances are in the last column. The standard deviations (and variances) for the qualitative characteristics are equal, because the lengths of the prior parameter intervals are equal. The values for the intercept and the parameter for the land area are different.

3.3. Mixed estimation of the econometric models

Model (1) in a matrix notation might be expressed as:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{u} \tag{3}$$

where:
y – vector of natural logs of unitary land prices [n x 1],
X – matrix of explanatory variables [n x k],
β – vector of parameters [k x 1],
u – error term [n x 1].

In model (3), k is a number of all explanatory variables, including intercept ($k = 12$). Model (3) is based on sample information. Standard assumptions are made according to (3). Matrix **X** is non – random and has full rank k. Error term has a zero mean and constant variance: $E(\mathbf{u}) = 0$, $E(\mathbf{uu}') = \sigma^2\mathbf{I}$. In analyzed database, explanatory variables are random, and rank of **X** is lower than k, because not all feature states are present in a database. For that reason, it is not possible to obtain $(\mathbf{X}'\mathbf{X})^{-1}$, because matrix $\mathbf{X}'\mathbf{X}$ is singular (its determinant is equal to zero). Mixed estimation resolves these problems.

Prior information has the form of stochastic restrictions imposed on parameters:

$$\mathbf{r} = \mathbf{R}\boldsymbol{\beta} + \mathbf{v} \tag{4}$$

where:
r – vector of expected values of parameters [j x 1],
R – matrix including restrictions [j x k],
β – vector of parameters [k x 1],
v – vector containing prior information about parameter errors [j x 1].

Prior expected values of parameters are presented in Table 3 (5th column). Vector $\mathbf{R}\boldsymbol{\beta}$ contains j linear combinations of parameters in **β**. Restrictions are imposed on all parameters, hence $j = k$ and matrix $\mathbf{R} = \mathbf{I}$, where **I** is a unitary matrix. It was assumed that vectors **v** and **u** are independent and $E(\mathbf{vv}') = \boldsymbol{\Psi}$. It is a covariance matrix of prior parameters. In analyzed models, matrix $E(\mathbf{vv}') = \boldsymbol{\Psi}$ is always diagonal with prior parameter variances on the main diagonal. Variances of prior parameters were presented in the last column of Table 3.

After combining sample (3) and prior information (4) we have:

$$\begin{bmatrix}\mathbf{y}\\ \mathbf{r}\end{bmatrix} = \begin{bmatrix}\mathbf{X}\\ \mathbf{R}\end{bmatrix}\boldsymbol{\beta} + \begin{bmatrix}\mathbf{u}\\ \mathbf{v}\end{bmatrix} \tag{5}$$

Matrix of error variances for (5) takes the form:

$$E\left(\begin{bmatrix}\mathbf{u}\\ \mathbf{v}\end{bmatrix}\begin{bmatrix}\mathbf{u}' & \mathbf{v}'\end{bmatrix}\right) = \begin{bmatrix}\sigma^2\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Psi}\end{bmatrix} \tag{6}$$

Off – diagonal blocks of (6) are zeros. It was assumed that vectors **u** and **v** are independent.

Mixed estimators might be expressed as a GLS (Generalized Least Squares) estimator:

$$\mathbf{b}_m = \left(\begin{bmatrix}\mathbf{X}' & \mathbf{R}'\end{bmatrix}\begin{bmatrix}\sigma^2\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Psi}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X}\\ \mathbf{R}\end{bmatrix}\right)^{-1}\begin{bmatrix}\mathbf{X}' & \mathbf{R}'\end{bmatrix}\begin{bmatrix}\sigma^2\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Psi}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{y}\\ \mathbf{r}\end{bmatrix} \tag{7}$$

$$\mathbf{b}_m = (\sigma^{-2}\mathbf{X}'\mathbf{X} + \mathbf{R}'\boldsymbol{\Psi}^{-1}\mathbf{R})^{-1}(\sigma^{-2}\mathbf{X}'\mathbf{y} + \mathbf{R}'\boldsymbol{\Psi}^{-1}\mathbf{r}) \tag{8}$$

Covariance matrix of $\boldsymbol{b}_m$ estimator has the form:

$$\mathbf{V}(\mathbf{b}_m) = (\sigma^{-2}\mathbf{X}'\mathbf{X} + \mathbf{R}'\boldsymbol{\Psi}^{-1}\mathbf{R})^{-1} \tag{9}$$

Variance of error term σ^2 is unknown, so the OLS estimator of σ^2 will be applied:

$$s^2 = (\mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y})/(n-k) \tag{10}$$

In Theil (1963), a statistic based on the χ^2 distribution is proposed to verify the hypothesis stating the consistency of prior and sample information. Measures of the contribution of prior and sample information are also introduced.

4. Empirical results – predictive accuracy of econometric land price models

Predictive accuracy should be the primary criterion for the verification of property price models. The model that better predicts prices in the test set is usually the better model. In order to evaluate the predictive ability of the models, a simulation experiment was conducted. Three types of models are evaluated in the experiment: 1) a model with only prior parameters, 2) OLS model, 3) mixed estimation model.

The experiment consisted of 1000 draws. In each draw, a set of 40 properties was randomly divided into a training set (30 properties) and a test set (10 properties). For each training set, an OLS model and a mixed estimation model were estimated. Based on these models and the model with parameters determined a priori, theoretical prices of 10 properties from the test set were predicted. For forecasts in the test set, average prediction errors were calculated, such as MPE (Mean Percentage Error) and MAPE (Mean Absolute Percentage Error).

Parameter estimates (for OLS models and mixed estimation models) were also recorded for each draw. The above scheme was repeated 1000 times (there were 1000 draws).

The quality of the prediction was assessed by the mean percentage error, which reports the biasedness of the predictions: $MPE = (p_i - \hat{p}_i)/p_i$, where p_i – the actual unit price of the property,$\hat{p}_i$ – the theoretical price, determined from the given model. The mean absolute percentage error was also used to evaluate the effectiveness of the forecasts: $MAPE = |p_i - \hat{p}_i|/p_i$.

In the experiment, for both the mixed estimation and the OLS models, approximately 1000 estimates of each parameter were obtained. For the OLS, not all parameters can be estimated. Parameters cannot be estimated for these feature states that are not present in the database. The value of these variables is always zero. In addition, there may be a situation in individual draws, where the explanatory variable(s) does not show variability. E.g. the variable may be equal to only zero or one. In such cases, parameter estimates cannot be obtained and the variables are discarded from the model. These problems do not apply to mixed estimation. If the variable does not exhibit variability, then prior information is critical to the parameter estimation.

The means and standard deviations of parameter estimates for 1000 draws are given in Table 4. The expected values of prior parameters are also given in the last column.

Table 4

Means and standard deviations of parameter estimates (OLS, mixed estimation models)

Variables	OLS		Mixed estimation		Prior expected values of parameters
	Mean	Stand. dev.	Mean	Stand. dev.	
const	5.646	0.149	4.929	0.001	4.927
a_i	-0.0002	0.0001	-0.0001	0.0000	-0.0001
u_{1i}	-0.183	0.155	0.088	0.007	0.069
u_{2i}	-0.318	0.143	0.108	0.010	0.139
n_{1i}	0.166	0.074	0.143	0.009	0.139
n_{2i}	-	-	0.277	0.000	0.277
ta_{1i}	0.059	0.053	0.072	0.011	0.104
ta_{2i}	-	-	0.208	0.000	0.208
pp_{1i}	0.266	0.081	0.112	0.011	0.069
pp_{2i}	0.213	0.102	0.127	0.010	0.139
l_{1i}	-	-	0.217	0.012	0.208
l_{2i}	-0.351	0.113	0.425	0.008	0.416

- Not possible to estimate

Source: own study.

The results obtained from OLS models are unacceptable. The effect of many feature states is negative, and thus inconsistent with theoretical expectations. This is the case for utilities or location. The impact of feature states should be monotonic, i.e., a better state should increase the theoretical price of a property more than a worse state. This is not so in the case of physical plot properties, or in the case of utilities. In the case of utilities, the best condition (complete utilities) reduces the unit price of the property to a greater extent than incomplete utilities. Such results are unacceptable and the poor quality of the sample information.

The OLS models cannot determine the distributions of parameter estimates for such variables as favorable neighborhood (n_{2i}), favorable transport availability (ta_{2i}) and average location (l_{1i}). In the mixed estimation model, the parameters at these variables are equal to the prior expected values, with zero standard deviations. The influence of the variable l_{1i} in the OLS model appears in the intercept.

In the mixed estimation model, the parameter estimate with this variable is slightly different from the prior value, and the standard deviation is greater than zero.

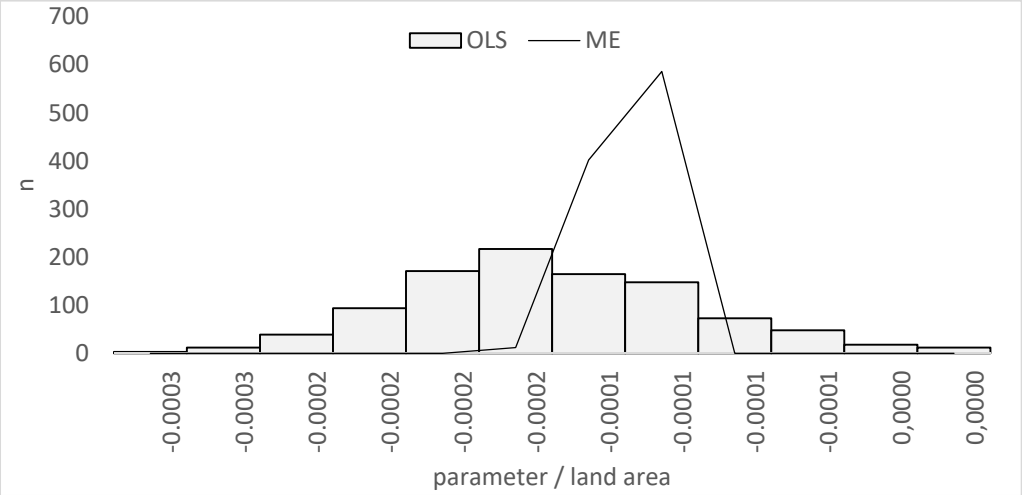


Fig. 1. Distributions of parameter estimates next to land area. *Source*: own study.

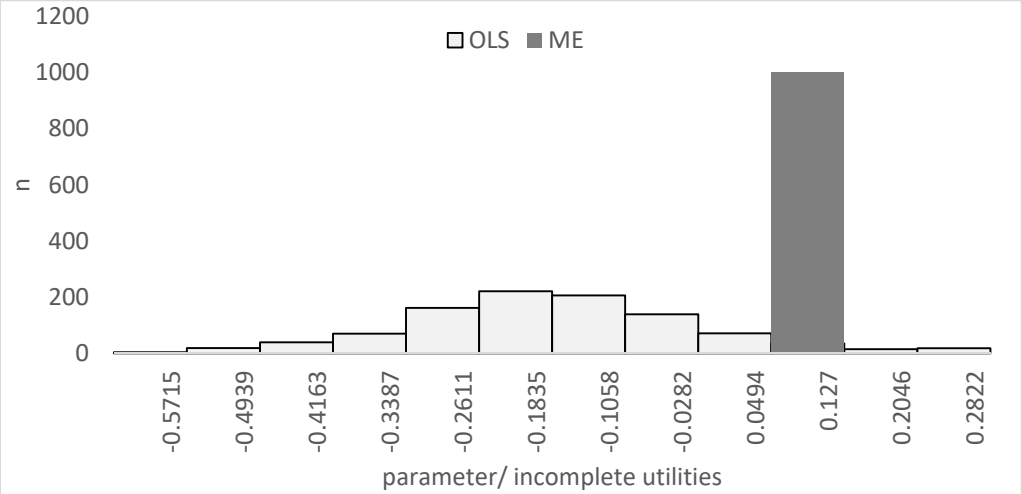


Fig. 2. Distributions of parameter estimates next to incomplete utilities. *Source*: own study.

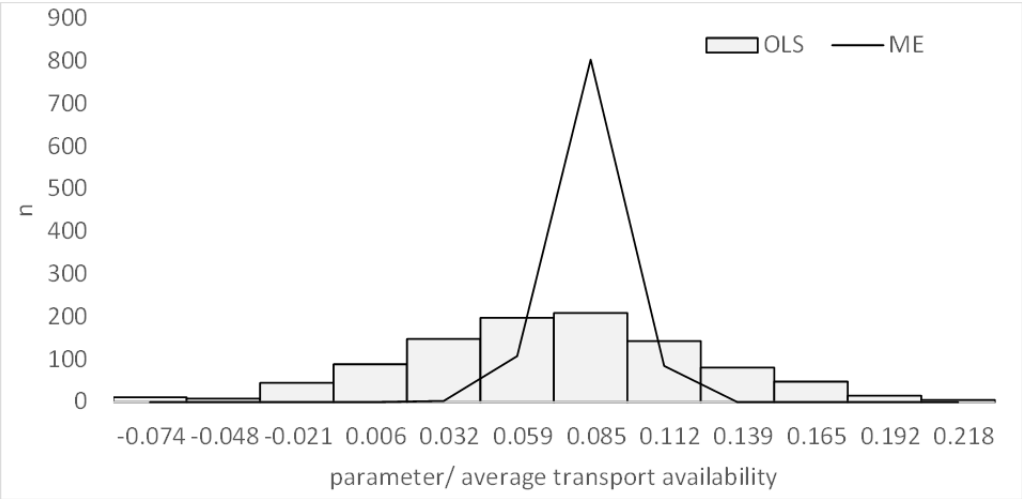


Fig. 3. Distributions of parameter estimates next to average transport availability. *Source*: own study.

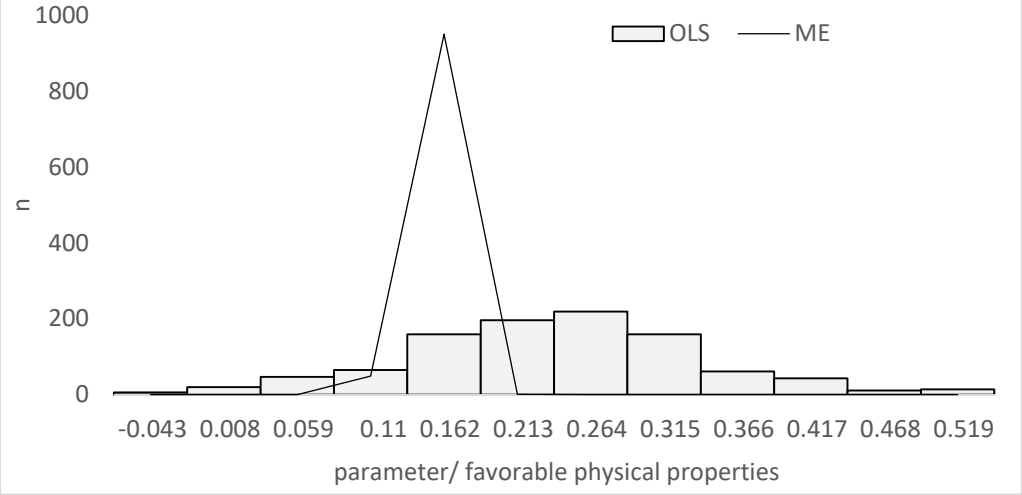


Fig. 4. Distributions of parameter estimates next to favorable physical plot properties. *Source*: own study.

The mixed estimation has a much higher efficiency than the OLS. The standard deviations of the OLS estimates are much higher than for the mixed estimation models. In the OLS models, the dispersion of the distributions of parameter estimates is much larger than for the mixed estimation models. The distributions of chosen parameter estimates are presented in Figures 1 – 4.

The MPE and MAPE characteristics for 1000 draws are shown in Table 5 and Figures 5 – 6.

Table 5

Distribution characteristics of MPE and MAPE for test sets (1000 draws)

	MPE			MAPE		
Statistic\ model	A priori	Mixed estimation	OLS	A priori	Mixed estimation	OLS
Min	-0.140	-0.195	-0.333	0.052	0.047	0.049
First quartile	-0.019	-0.041	-0.060	0.109	0.100	0.113
Median	0.015	0.000	-0.012	0.130	0.122	0.137
Mean	0.013	-0.002	-0.016	0.130	0.122	0.143
Third quartile	0.047	0.038	0.034	0.151	0.143	0.165
Max	0.144	0.152	0.186	0.222	0.227	0.384

Source: own study.

The closer the mean MPE is to zero, the more unbiased the predictions. The mean MPE was closest to zero for the mixed estimation, for which price predictions in the test sets were unbiased ($MPE = -0.002$). The median MPE was equal to zero, unlike for the other two groups of models. The MPE spread was largest for the OLS models and smallest for the model with parameters specified a priori. From the point of view of prediction unbiasedness, mixed estimation gives the best results, as can be seen in Figure 5. Price predictions based on the model with parameters specified a priori were slightly underestimated (there are more positive errors). In contrast, price predictions from OLS models were slightly overestimated (there are more negative errors).

MAPE indicates a greater efficiency of forecasts obtained from the mixed estimation models. The mean and median MAPE for the mixed estimation models was 0.12. For the model with parameters fixed a priori it was 0.13, while for the OLS model – 0.14. MAPE distribution (Fig. 6) confirms the advantage of low errors for mixed estimation. On the other hand, the OLS model shows a predominance of large errors.

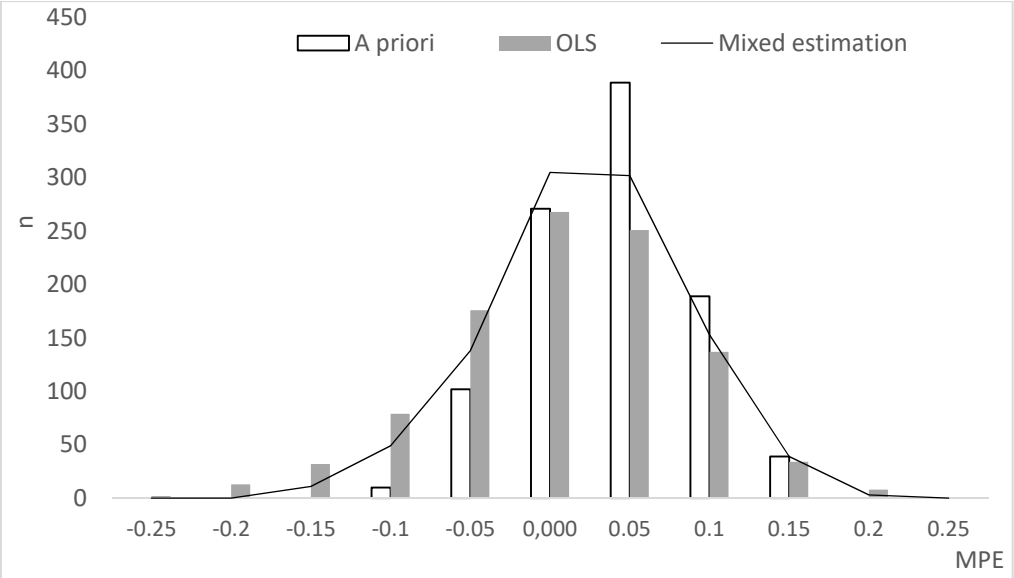


Fig. 5.MPE distribution for 1000 draws. *Source*: own study.

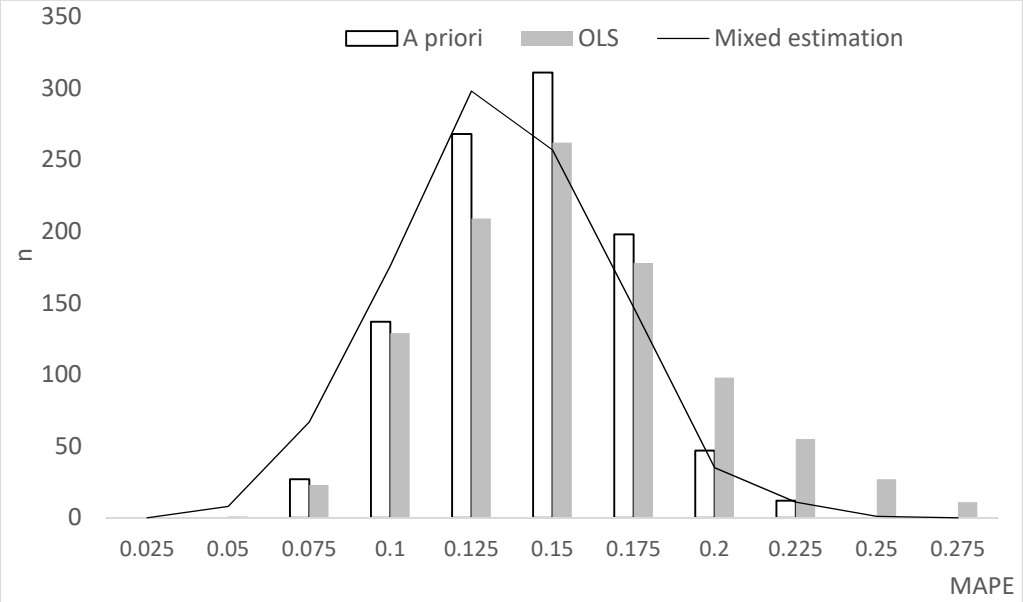


Fig. 6.MAPE distribution for 1000 draws. *Source*: own study.

5. Discussion and conclusions

The results obtained with OLS models are unacceptable. The signs of the parameter estimates are not as expected. All features (except land area) should increase the unit price of a property. In the OLS model, this effect is often negative. Moreover, in OLS models, a better feature state increases the price less than the state preceding it. The difference between states might be nonlinear, but introducing zero – one variables takes this into account. In addition, the impact of some feature states cannot be estimated because they do not appear in the database. It is worth remembering that features may affect the price in many other ways, e.g. by interacting with other features.

Mixed estimation, which incorporates prior information about parameters, can be a solution to the mentioned problems. A formula for "moving" from weights of features (or their ranges), to ranges of model parameters was proposed.

The presented analyses confirmed that mixed estimation leads to much better results than the OLS models. Mixed estimation also performs more favorably than the model with parameters specified a priori. This shows that the best solution is to combine the sample and prior information. In the mixed

estimation models, the parameter estimates are consistent with theoretical expectations. In mixed estimation, it is also possible to estimate the impact of those characteristics that do not exhibit variability (take a constant level for all properties). Then, the parameter estimate is determined by the prior expected value of the parameter.

Based on the simulation experiment conducted (1000 draws), it can be concluded that the predictive abilities in the test sets are the best for the mixed estimation models. The distributions of parameter estimates prove greater effectiveness of mixed estimators.

6. References

Barańska, A., & Łuczak, A. (2007). Comparing the results of function model estimation for the prediction of real estate market values in additive and multiplicative form. *Geomatics and Environmental Engineering, 1*(3), 19–35.

Cellmer, R., Cichulska, A., & Bełej, M. (2020). Spatial Analysis of Housing Prices and Market Activity with the Geographically Weighted Regression. *ISPRS International Journal of Geo-Information, 9*(6), 380. https://doi.org/10.3390/ijgi9060380

Dell, G. (2017). Regression, Critical Thinking, and the Valuation Problem Today. *The Appraisal Journal, 85*(3), 217–230.

Doszyń, M. (red. n.). (2020a).*System kalibracji macierzy wpływu atrybutów w szczecińskim algorytmie masowej wyceny nieruchomości*. WNUS. Szczecin [Doszyń, M. (red. n.) (2020a). *Calibration system of attributes influence matrix in Szczecin mass real estate valuation algorithm*. WNUS. Szczecin]

Doszyń, M. (2020b). Algorithm of real estate mass appraisal with inequality restricted least squares (IRLS) estimation. *Journal of European Real Estate Research, 13*(2), 161–179. https://doi.org/10.1108/JERER-11-2019-0040

Doszyń, M. (2021). Prior information in econometric real estate appraisal: A mixed estimation procedure. *Journal of European Real Estate Research, 14*(3), 349–361. https://doi.org/10.1108/JERER-11-2020-0057

Doszyń, M. (2022). Might expert knowledge improve econometric real estate mass appraisal? *The Journal of Real Estate Finance and Economics*. Advance online publication. https://doi.org/10.1007/s11146-022-09891-3

Durbin, J. (1953). A Note on Regression When There is Extraneous Information About One of the Coefficients. *Journal of the American Statistical Association, 48*(264), 799–808. https://doi.org/10.1080/01621459.1953.10501201

Fik, T. J., Ling, D. C., & Mulligan, G. F. (2003). Modelling Spatial Variation in Housing Prices: A Variable Interaction Approach. *Real Estate Economics, 31*(4), 623–646. https://doi.org/10.1046/j.1080-8620.2003.00079.x

Grömping, U. (2010). Inference with Linear Equality and Inequality Constraints Using R: The Package ic.infer. *Journal of Statistical Software, 33*(10), 1–31. https://doi.org/10.18637/jss.v033.i10

Hozer, J., Foryś, I., Zwolankowska, M., Kokot, S., & Kuźmiński, W. (1999). *Ekonometryczny algorytm masowej wyceny nieruchomości gruntowych*. [*Econometric algorithm of land property mass appraisal*]. Uniwersytet Szczeciński, Stowarzyszenie „Pomoc i Rozwój", Szczecin

Isakson, H. R. (1998). The Review of Real Estate Appraisals Using Multiple Regression Analysis. *Journal of Real Estate Research, 15*(2), 177–190. https://doi.org/10.1080/10835547.1998.12090922

Jahanshiri, E., Buyong, T., & Shariff, A. R. M. (2011). A Review of Property Mass Valuation Models. *Pertanika Journal of Science & Technology, 19*, 23–30.

Kokot, S., & Gnat, S. (2019). Simulative Verification of the Possibility of Using Multiple Regression Models for Real Estate Appraisal. *Real Estate Management and Valuation, 27*(3), 109–123. https://doi.org/10.2478/remav-2019-0029

McCluskey, W. J., McCord, M., Davis, P. T., Haran, M., & McIlhatton, D. (2013). Prediction Accuracy in Mass Appraisal: A Comparison of Modern Approaches. *Journal of Property Research, 30*(4), 239–265. https://doi.org/10.1080/09599916.2013.781204

Mehta, J. S., & Swamy, P. A. V. B. (1970). The Finite Sample Distribution of Theil's Mixed Regression Estimator and a Related Problem. *Review of the International Statistical Institute, 38*(2), 202–209. https://doi.org/10.2307/1402143

Mittelhammer, R. C., & Conway, R. K. (1988). Applying Mixed Estimation in Econometric Research. *American Journal of Agricultural Economics, 70*(4), 859–866. https://doi.org/10.2307/1241927

Nagar, A. L., & Kakwani, N. C. (1964). The Bias and Moment Matrix of a Mixed Regression Estimator. *Econometrica, 32*(1/2), 174–182. https://doi.org/10.2307/1913742

Pace, R. K., & Gilley, O. W. (1990). Estimation employing a priori information within mass appraisal and hedonic pricing models. *The Journal of Real Estate Finance and Economics, 3*(1), 55–72. https://doi.org/10.1007/BF00153706

Pace, R. K. (1995). Parametric, semiparametric, and nonparametric estimation of characteristic values within mass assessment and hedonic pricing models. *The Journal of Real Estate Finance and Economics, 11*, 195–217. https://doi.org/10.1007/BF01099108

Pagourtzi, E., Assimakopoulos, V., Hatzichristos, T., & French, N. (2003). Real Estate Appraisal: A Review of Valuation Methods. *Journal of Property Investment & Finance, 21*(4), 383–401. https://doi.org/10.1108/14635780310483656

Sing, T. F., Yang, J. J., & Yu, S. M. (2021). Boosted Tree Ensembles for Artificial Intelligence Based Automated Valuation Models (AI-AVM). *The Journal of Real Estate Finance and Economics*. Advance online publication. https://doi.org/10.1007/s11146-021-09861-1

Swamy, P. A. V. B., & Mehta, J. S. (1969). On Theil's Mixed Regression Estimator. *Journal of the American Statistical Association, 64*(325), 273–276. https://doi.org/10.1080/01621459.1969.10500969

Theil, H., & Goldberger, A. S. (1961). On Pure and Mixed Statistical Estimation in Economics. *International Economic Review, 2*(1), 65–78. https://doi.org/10.2307/2525589

Theil, H. (1963). On the Use of Incomplete Prior Information in Regression Analysis. *Journal of the American Statistical Association, 58*(302), 401–414. https://doi.org/10.1080/01621459.1963.10500854

Zurada, J., Levitan, A. S., & Guan, J. (2011). A Comparison of Regression and Artificial Intelligence Methods in a Mass Appraisal Context. *Journal of Real Estate Research, 33*, 349–388. https://doi.org/10.1080/10835547.2011.12091311