THE ESTIMATION OF THE ECONOMETRIC MODEL OF GRAIN YIELD: A COMPARISON OF RESULTS USING DIFFERENT METHODS OF DATA MINING

R. Põldaru, J. Roots, A.-H. Viira

Estonian University of Life Sciences

ABSTRACT. *This paper presents a comparison of estimated parameters of econometric model of grain yield in Estonia. For parameter estimation various data mining (DM) methods are used principal component regression (PCR), Bayesian statistics (BUGS), artificial neural network (ANN) models, fuzzy regression (FR) and support vector machines regression (SVMR). The data are a balanced panel of fifteen Estonian counties observed during the period from 1994 to 2001. Our analysis shows that Bayesian methods are the most acceptable and widespread. This result is primarily obtained due to assignment of informative priors for the parameters for fertilizer use.*

Keywords: *data mining, econometric models, grain yield.*

Introduction

The amount of information that is created and stored in the electronic databases is increasing rapidly. Everyday transactions, protocols, and documents are being stored in the databases and automated monitoring systems create vast information repositories.

With the advent of the Internet, these information resources have become available to individuals and companies regardless of national borders and constraints of time and space. As a consequence, information overload is becoming the new plague of the information society. It is, therefore, becoming increasingly important to provide effective tools to help users organize, manage, understand, and access large repositories of information.

Data mining is the process of discovery of useful information from large collections of data. It has common frontiers with several fields including Data Base Management (DBM), Artificial Intelligence (AI), Machine Learning (ML), Pattern Recognition (PR), and Data Visualisation (DV). New data analysis procedures provided by data mining have substantially changed the situation in the field of data processing (DP). The situation in data mining is the most challenging. Data mining, often called knowledge discovery in databases (KDD), started to depart from the statistics and machine learning ghettos and moved into the mainstream 10–15 years ago.

This paper presents an overview of different DM methods investigated by the researchers of the Institute of Economics and Social Sciences of Estonian University of Life Sciences, and as an example, makes a comparison of the results of estimated parameters of econometric model of grain yield in Estonia, using different DM methods and discusses the implementation of those methods for analysing the grain yield (results). The data used for parameter estimation are the same for all methods mentioned above.

In traditional econometrics a regression problem is handled by the ordinary least squares (OLS). As an attempt to meet future challenges, we have constructed a special econometric model to explain the relationship between the grain yield in Estonian counties and 12 explanatory variables. The possibilities of alternative methods for estimating the parameters of an econometric model of grain yield are investigated.

Materials and Methods

Data

The data used were obtained from various publications of Statistical Office of Estonia. A balanced panel of fifteen Estonian counties was observed during the period 1994 to 2001. The characteristics of the data are reported in Table 1.

The dependent variable is average grain yield (y), and independent variables are time dummies (x_1, x_2, x_3, x_4 and x_5), variables for fertilizer use (x_8, x_9 and x_{10}), variable of land quality (x_{11}) and variables of production structure (x_7 and x_{12}).

It is, however, vital to distinguish variables of primary interest from those which specify secondary structure for the model. In the econometric model of grain yield the parameters of primary interest were parameters for the nutrients (Nitrogen, Phosphorous and Potassium fertilizer). From previous economic and

agronomic research the approximate values for these parameters are known. These approximate values can be used for appreciating the results of estimating the parameters of econometric model of grain yield using different methods.

Table 1 provides the coefficients of correlation between independent variables and grain yield. In most of cases the coefficients of correlation between the output and inputs are statistically significant. Only one coefficient from 12 is not significant ("use of Potassium fertilizer" r = 0.085). It is important to note that variables for fertilizer use (Potassium) are not significant. The economic theory and practice assert that these variables must be significant. Consequently, some of the fertilizer parameters cannot be well estimated by the data. Obviously, the main reason for that situation is inadequate data for classical regression analysis. It is important to remember that statistical data are collected for administrative purposes, not for the benefit of econometric research. But it is also the case that statisticians and econometricians have developed and used a large body of techniques, aimed at making the best of inadequate data.

Table 1. Definitions of independent variables (inputs)
Table 1. *Sõltumatud muutujad (sisendid)*

Definitions of independent variables (inputs) *Sõltumatud muutujad (sisendid)*	Measure *Mõõtühik*	X_i	Coefficient of correlation *Korrelatsioonikordaja*
Dummy variable for year 1999 / *Fiktiivne muutuja 1999. a kohta*	–	x_1	–0.477
Dummy variable for year 1997 / *Fiktiivne muutuja 1997. a kohta*	–	x_2	0.152
Dummy variable for year 1996 / *Fiktiivne muutuja 1996. a kohta*	–	x_3	0.266
Dummy variable for year 2000 / *Fiktiivne muutuja 2000. a kohta*	–	x_4	0.332
Dummy variable for year 2001 / *Fiktiivne muutuja 2001. a kohta*	–	x_5	0.290
Sown area of grain / *Teravilja kasvupind*	ha	x_6	0.523
Fraction of fertilized area in total grain sown area *Väetatud kasvupinna osakaal*	%	x_7	0.323
Use of Nitrogen fertilizer / *Lämmastikväetis*	kg/ha	x_8	0.361
Use of Phosphorous fertilizer / *Fosforväetis*	kg/ha	x_9	0.279
Use of Potassium fertilizer / *Kaaliumväetis*	kg/ha	x_{10}	0.085
Quality of land / *Hindepunkt*	points	x_{11}	0.467
Fraction of grain sown area in total sown area *Teravilja osakaal kogu kasvupinnast*	%	x_{12}	0.381

Methods of investigation

Data mining is a rapidly growing field, whose development is driven by strong research interests as well as urgent practical, social, and economic needs. Some authors forecast that in the not-too-long term, DM may become as common and easy to use as E-mail. Therefore, DM should be implemented also in the area of agrarian research.

The most commonly used techniques in DM are (Friedman 1997; Goebel, Gruenwald, 1999):

- artificial neural networks;
- association rule discovery ("market basket analysis");
- Bayesian methods (belief networks, graphical models, statistical methods);
- classical statistical methods (parameter estimation, hypothesis testing, fitting models to data *etc*.);
- classification rules (studied in statistics, machine learning, neural networks, and expert systems);
- clustering analysis or data segmentation (studied in statistics, machine learning);
- decision tree induction (cart, chaid);
- genetic algorithms (optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution);
- multivariate statistical methods (principal component analysis, discriminant analysis *etc*.)
- nearest neighbor method ("case based reasoning");
- neuro-fuzzy systems or fuzzy sets (methodology for representing and processing uncertainty);
- self-organising maps;
- support vector machine (SVM) regression.

We at the Estonian University of Life Sciences (Institute of Economics and Social Sciences) have investigated the possibilities of some new DM methods and of implementation of algorithms used in DM packages (Bayesian statistical methods, neural networks, principal components method, decision trees and rules - CART (Classification and Regression Trees)). The results are published in many papers and conference theses (Põldaru *et al*., 2003 a, b, c, d; Põldaru 2005; Põldaru *et al*., 2005).

Most of these papers discuss estimation of grain model parameters, but there are also papers discussing the estimation of milk yield and milk cost models (Põldaru *et al*., 2005a; Põldaru, Roots, 2005; Põldaru *et al*., 2006). One paper discusses usage of stochastic frontier analysis (SFA) for grain yield data (Põldaru, Roots, 2004).

In current paper the following methods for grain yield model parameter estimation are compared:

- ordinary least squares;
- principal component regression;
- Bayesian regression;
- artificial neural network;
- fuzzy regression;
- support vector machine regression.

Next an overview is given about the data mining methods used for estimating model parameters.

Principal Component Analysis

Principal component regression is the most common method besides of ordinary least squares (OLS). The principal component regression method combines the principal component analysis (PCA) and ordinary least squares regression method to create a quantitative model for complex economic systems.

Principal component analysis involves a mathematical procedure that transforms a number of possibly correlated variables into a (smaller) number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. If the independent variables (inputs) are linearly related and are contaminated by errors, the first few components capture the relationship between the variables, and the remaining components are comprised only of the error. Thus, eliminating the less important components reduces the contribution of errors in the input data and represents it in a compact manner.

It is important to note that PRC is a two-step process.

PCA (the first step) is an input (independent variables) transformation method that extracts projection directions, or principal component loadings by satisfying the following criterion:

$$\max_{\alpha_m}\{\mathrm{var}(X\alpha_m)\}. \quad (1)$$

The projection directions are constrained to be orthogonal, and are eigenvectors of the input covariance matrix. The dimensionality of the input space may be decreased by selecting a subset of the latent variables (principal component scores), that capture most of the variance in the input data.

The PCR extract the scores (latent variables) from the input data (independent variables) (the first step) and then performs the OLS regression between the selected latent variables and dependent variable (grain yield) (the second step). The model determined by PCR may be expressed by equation (2):

$$y = \sum_{m=1}^{M} \beta_m \cdot z_m = \sum_{m=1}^{M} \beta_m \sum_{j=1}^{J} \alpha_{jm} \cdot x_j \text{ , where} \quad (2)$$

y – dependent variable (output);
x_j – independent variables (inputs);
z_m – latent variables (principal component scores);
J – number of independent variables;
M – number of selected latent variables (principal component scores);
α_{jm} – principal component loadings (weights);
β_m – regression coefficients.

Unlike OLS, inversion of the covariance matrix of the principal component scores to find the regression coefficients in PCR is possible even when the inputs are highly correlated, since the principal component loadings, or scores, are mutually orthogonal and uncorrelated. PCR considers only the input space in finding projection directions while ignoring the input output relationship.

An acceptable econometric model of grain yield was obtained when two principal component was used (Põldaru, Roots, 2001a).

Bayesian Statistical Methods

Advances in computers and numerical methods have made it possible to implement Bayesian analysis previously considered infeasible. The limitations of traditional statistical methods are part of the reason for the new popularity of this approach.

The general-purpose MCMC-based software, BUGS (Spiegelhalter *et al*., 1996) was used for estimating the parameters of the econometric model.

In Bayesian analysis, a comprehensive probabilistic model is employed to describe the relationships among various quantities under consideration: those that we observe (data and knowledge), those about which we learn (scientific hypotheses), and those that are needed in order to construct a proper model.

We have a probabilistic model $f(y \mid \theta)$ for the observed data y = (y_1, …, y_n) given a vector of unknown parameters θ=(θ_1, … , θ_p). While the classical statistician would assume that θ is an unknown but fixed set of parameters to be estimated from y, the Bayesian statistician places a prior distribution (probability density) $g = g(\theta)$ on θ. The Bayesian analysis uses Bayes rule (theorem) to compute the posterior distribution:

$$p(\theta \mid y) = \frac{f(y \mid \theta) \cdot g(\theta)}{\int_{-\infty}^{\infty} f(y \mid \theta) \cdot g(\theta) d\theta} \tag{3}$$

Let's consider a linear econometric model of the form

$$y = X \cdot \beta + \varepsilon \tag{4}$$

where X is n × p matrix of explanatory variables, β is p × 1 vector of parameters, and ε is an n × 1 vector of independent identically distributed random errors.

An important issue in applying the Bayes approach is the choice of a prior distribution for the unknown parameters. The prior distribution is a key part of the Bayesian inference and represents the information about the uncertain parameter β that is combined with the probability distribution of the new data to yield the posterior distribution, which in turn is used for the future inferences and decisions involving β. The key issues in setting up a prior distribution are:

- what information is included in the prior distribution, and
- the properties of the resulting posterior distribution.

The potential variants for prior distribution are:

- non-informative prior distributions;
- highly or moderately informative prior distributions.

It is vital to distinguish parameters (variables) of primary interest from those, which specify secondary structure for the model. In the econometric model of grain yield the parameters of primary interest were parameters for the nutrients (nitrogen, phosphorous and potassium fertilizer). We know the approximate values for these parameters from previous economic and agronomic research, and thus informative prior distributions can be used. A normal prior with plausible parameters was assigned for the parameters of primary interest. "Non-informative priors" – a normal prior with an extremely small precision (large variance) – were assigned for other parameters.

An acceptable econometric model of grain yield was obtained when the informative priors were assigned for the parameters for the nutrients (nitrogen, phosphorous and potassium fertilizer) (Põldaru, Roots, 2001b, c; Põldaru, Roots, 2003b).

Artificial Neural Network

Neural networks provide a new approach to the problem of parameter estimation of nonlinear econometric models (Kaashoek, van Dijk, 2000; Kuan Chung-Ming, White, 1994).

The model of artificial neural network consists of three layers: a layer of "input" units is connected to a layer of "hidden" units, which is connected to a layer of "output" units. The cells of the input layer correspond to the 'regressors' or 'explanatory variables' in the standard linear regression model. The cells in the output layer correspond to the dependent variables. The hidden layer contains cells, which transmit the signals from the input layer to the output layer. These cells may be interpreted as unobserved components built into the linear model. It is the presence of this hidden layer which permits the nonlinear mapping, since similar networks lacking a hidden layer can only affect a multivariate linear mapping. The activity of the input units represents the raw information that is fed into the network. The activity of each hidden unit is determined by the activities of the input units and the weights on the connections between the input and the hidden units. The weights between the input and hidden units determine when each hidden unit is active, and so by modifying these weights, a hidden unit can choose what it represents.

The network transmits the signals as follows. A weighted sum of the signals of the input cells are sent to the hidden layer cells. Within the cells of this layer the values of the signals received are transformed by the so called 'activation function'. Then a weighted sum of the transformed signals is sent to the cells of the output layer. The weights in the neural network correspond to unknown parameters in the linear model.

The mathematical structure of a neural net y is equal to:

$$y = \sum_{h=1}^{H} c_h \cdot g\left(\sum_{i=1}^{I} a_{ih} \cdot x_i + b_h \right) + d \text{ , where} \quad (5)$$

i index of input cells (explanatory variables), $i = 1;...; I$;
h index of hidden layer cells, $h = 1;...;H$;
$g(z)$ activation function;
x_i value of input cell (explanatory variables) i;
y value of output cell (dependent variable);
a_{ih} weight of the signal from input cell i to hidden cell h;
b_h constant input weight for hidden cell h;
c_h weight of the signal from hidden cell h to output cell y;
d constant weight for output cell.

Neural networks are flexible, but the price of increased flexibility is the danger of "overfitting". This statement may be explained as follows. In empirical econometric models one assumes that an observed model consists of a part that can be explained and a part that is labeled unexplained or "residual noise". With "overfitting" this noise is also "fitted". "Overfitting" with neural networks may occur by increasing the number of hidden cells, which increases the number of parameters, without increasing the number of explanatory variables or inputs. Because of this possibility neural nets are more sensitive to "overfitting" than other classes of econometric models.

The grain model parameters were estimated on the basis of alternative variants of models (Põldaru, Roots, 2002a, b; Põldaru, Roots, 2003a). The neural network approach was also used for estimating parameters of grain yield model (Põldaru *et al*., 2005b) and milk cost model (Põldaru *et al*., 2006) using FADN data.

For neural network model (parameter) estimation different software was used: a) the Excel Solver, b) the ANN module of Programming Environment R, c) STATISTICA (data analysis software system) version 7.

Fuzzy Regression

Fuzzy regression aims to model vague and imprecise phenomena using the fuzzy model (Tanaka, Ishibuchi, 1992). Tanaka *et al*. (1982) introduced fuzzy linear regression as a means to model casual relationships in systems when ambiguity or human judgment inhibits a crisp measure of the dependent variable. Unlike conventional regression analysis, where deviations between observed and predicted values reflect measurement error, deviations in fuzzy regression reflect the vagueness of the system structure expressed by the fuzzy parameters of the regression model.

In modeling a fuzzy system with fuzzy linear functions, the vagueness of the fuzzy output data may be caused by both the indefiniteness of model parameters and the vagueness of the input data. Fuzzy regression is a fuzzy variation of classical regression analysis.

The fuzzy parameters of the model are considered to be possibility distributions, which corresponds to the fuzziness of the system. The fuzzy parameters are determined by a linear programming procedure, which minimizes the fuzzy deviations subject to constraints of the degree of membership fit.

The fuzzy linear regression model has the following form:

$$\widetilde{Y}_i = \alpha_0 + \alpha_1 \cdot \widetilde{X}_{i1} + ... + a_k \cdot \widetilde{X}_{ik} + \widetilde{\varepsilon}_i, \quad (6)$$

where α_0 and α_i are the crisp regression coefficients, $\widetilde{Y}_i$ and $\widetilde{X}_{ij}$ are fuzzy observations with the membership functions $\mu_{\widetilde{Y}_i}$ and $\mu_{\widetilde{X}_{ij}}$, respectively, and $\widetilde{\varepsilon}_i$ is the fuzzy error associated with the regression model. The value "crisp" indicates that the variable or parameter is conventional (ordinary) number, as used in classical regression analysis. The value "fuzzy" indicates that the variable or parameter is fuzzy number and for that number the membership function is known or calculated.

A fuzzy number may be defined as $\widetilde{A} = (\alpha, \beta, \gamma)$, where α denotes the center (or mode), β and γ, are the left spread (or width) and right spread, respectively. The main algebraic and geometric characteristic of fuzzy number is the triangular membership function μ.

Different fuzzy regression models were used for estimation of the econometric model of grain yield in Estonian counties (Põldaru *et al*., 2004a).

Support Vector Machines

Support Vector Machines (SVM) regression provide a new approach to the problem of parameter estimation of linear and especially nonlinear econometric models. Model parameter estimation involves optimization of a convex cost function: there are no false local minimum to complicate the estimation process (Vapnik, 1998).

The goal in SVM regression is to find a function *f(x)* that has at most ε deviation from the actually obtained targets y_i for all the data, and at the same time, is as flat as possible. SVM regression uses the ε-insensitive loss function shown in Figure 1. If the deviation between the actual and predicted value is less than ε, then the regression function is not considered to be in error. Thus mathematically we would like

$$-\varepsilon \le a \cdot x_i + b - y_i \le \varepsilon \quad (7)$$

Geometrically it can be viewed as a band or tube of size 2ε around the hypothesis function *f(x)* and any points outside this tube can be dealt with errors. All training (data) points (x_i, y_i) for which $\left|f(x_i) - y_i\right| \ge \varepsilon$ are known as support vectors; it is only these points that determine the parameters of *f(x)*. In other words, we do not care about errors as long as they are less than ε, but will not accept any deviation larger than this.

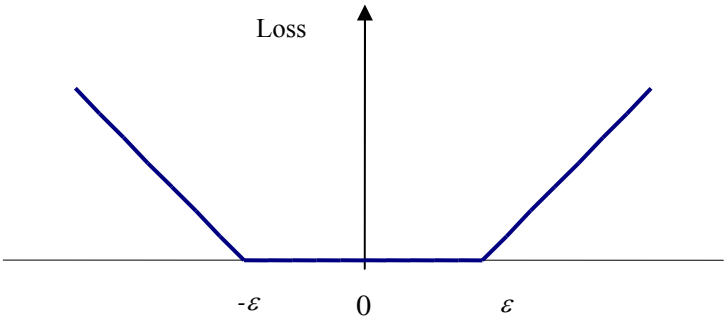

Figure 1. A piecewise linear ε-insensitive tube loss function
Joonis 1. *Osakaupa lineaarne ε-sõltuv kaofunktsioon*

The value of regression parameter α and predicted value ***f(x)*** can be calculated as follows

$$a = \sum_{i=1}^{n} \left(\alpha_i - \alpha_i^*\right) \cdot x_i \quad (8)$$

and

$$f(x) = \sum_{i=1}^{n} \left(\alpha_i - \alpha_i^*\right) \cdot x_i \cdot x + b. \quad (9)$$

This is the so-called Support Vector expansion, i.e. α can be completely described as a linear combination of the training patterns x_i.

Different SVM regression models were used for estimation of the econometric model of grain yield (Põldaru *et al*., 2004b). The SVM approach was also used for estimating parameters milk cost model (Põldaru *et al*., 2005a), and milk yield per cow model (Põldaru, Roots, 2005) using FADN data. For estimating model parameter special software was used (Meyer, 2003).

Results and discussion

Table 2 presents the estimates of parameters of the econometric model. The parameters are estimated on the basis of alternative DM methods (principal component regression – PCR, Bayesian regression – BUGS, artificial neural network – ANN, fuzzy regression – FR and support vector machines – SVM) and compared to ordinary least squares (OLS).

Table 2. Estimates for parameters of the econometric model using different methods
Table 2. *Ökonomeetriliste mudelite parameetrite hinnang erinevate meetodite kasutamisel*

Xi	OLS	PCR	BUGS	ANN	FR	SVM
x_1	–294.10*	–133.8*	–237.2*	–224.17	–245	–367
x_2	318.12*	41.6*	346.1*	334.52	314.1	299
x_3	549.33*	–11.7	597.1*	539.96	620.7	497
x_4	500.07*	–	–	518.27	428.2	438
x_5	423.90*	–	–	405.14	428.9	473
x_6	0.0071	0.0073*	0.0163*	0.0137	0.0047	0.0021
x_7	4.83*	1.14*	3.2*	4.19	6.21	7.3
x_8	1.06	3.66*	9.0*	3.77	6	2.75
x_9	8.20*	7.95*	3.6*	7.58	6.99	8.4
x_{10}	–1.64	5.14*	2.0*	1.97	4.63	–1.8
x_{11}	20.98*	9.82*	16.2*	17.25	6.44	18.7
x_{12}	0.74	4.22*	11.3	6.92	7.1	–3.1
x_8^2	–	0.0273*	–0.0612*	–	–	–
x_{12}^2	–	0.0463*	–0.0349	–	–	–
R^2	0.833*	0.611	0.885	0.840*	0.76	0.807

* indicates that coefficient is statistically significant / *koefitsient on statistiliselt usutav*

Next we discuss the results in Table 2. Most popular OLS method (Põldaru, 2000) gave a very good prediction, coefficient of determination $R^2 = 0.833$ (variant OLS) and most regression coefficients are statistically significant. From the formal point of view the model is adequate. Economic analysis evaluation demonstrates that many economically chosen variables are not significant (x_8, x_9, x_{10}). It is important to note that estimated sign of "Potassium fertilizer use" (x_{10}) is negative. The economic theory and practice assert that for the fertilizers model parameters should be positive. The independent variable for Potassium fertilizer is contaminated by errors. Consequently, from economic point of view the model was incorrect and must be improved.

An acceptable econometric model of grain yield was obtained when the two first principal components are used. In the PCR model (variant PCR in Table 2) the coefficients for fertilizer use are positive and significant. The deficiency of the model is inadequate coefficients for dummy variables and the relatively low value of coefficient of determination $R^2 = 0{,}611$.

The Bayesian regression (BR) for estimating the parameters of econometric model of grain yield gave acceptable estimates for the model parameters (variant BUGS in Table 2). An acceptable econometric model of grain yield was obtained when informative priors were assigned for the parameters for the nutrients (Nitrogen – x_8, Phosphorous – x_9 and Potassium fertilizer – x_{10}). The estimates for the fertilizer use (x_8, x_9 and x_{10}) are positive and mostly significant.

The neural network approach was also used for estimating parameters of grain yield model. The variant with one hidden cell and constraint for fertilizer parameters give acceptable results (variant ANN in Table 2). The estimates for the fertilizer use (x_8, x_9 and x_{10}) are positive.

The results in Table 2 (variant FR and variant SVM) show that both the fuzzy regression and support vector machine regression give mostly acceptable estimates for the parameters of grain yield model.

Next we compare the estimates of different variants. For the most dummy variables ($x_1 \ldots x_5$) the estimates of the parameters of the econometric model do not differ substantially for different methods of estimation except the method of principal components. For example the value of regression coefficient a_2 ranges from 299 to 346, i.e. the range is relatively moderate. This is also a case for the coefficient a_3, which ranges from 497 to 620 in different variants. It is important to note that at the same time the signs of the parameters remained the same. Therefore we can conclude that the estimates for dummy variables are robust and do not depend on the used estimation method.

When the estimates of model parameter a_6 (grain area) are compared, one can see that if BUGS and ANN methods are used, the parameter values are two or three times higher than in other variants. At the same time the sign of the parameters did not change. For the parameter a_7 (the share of fertilized area) the PCR gave unstable estimates.

The most important parameters of current model are those describing fertilizer use. As mentioned before, for an acceptable model the coefficients for fertilizer use must be positive and significant. Table 2 shows, that for variants OLS and SVM the sign for Potassium fertilizer (x_{10}) is negative. Consequently, estimates of parameter a_{10} are inadequate. For other methods all estimates of parameters for fertilizer use are positive. Consequently, these variants should be discussed in detail. If in the field experiments the fertilizers usage increases by 1 kg per hectare then the grain yield increases in average by 10–13 kg per hectare (Kärblane, 1997), then in the real

production the value of parameters should be lower. In the case of variant PCR added 1 kg of fertilizers in average increases the grain yield value by 5.6 kg per hectare, for variant BUGS 4.9, for variant ANN 4.4 and for variant FR 5.9 kg per hectare. These estimates are acceptable (appropriate) since the statistical data (collected for administrative purposes, not for the benefit of econometric research.) are used.

Next we compare the estimates of different fertilizers. The value of regression coefficient a_8 (use of Nitrogen fertilizer) ranges from 3.7 to 9.0 i.e differs approximately two times. The value of regression coefficient a_9 (use of Phosphorous fertilizer) ranges from 3.6 to 7.9 i.e differs also approximately two times. The value of regression coefficient a_{10} (use of Potassium fertilizer) ranges from 2.0 to 5.1 i.e differs also approximately two times. It should be mentioned, that for different type of fertilizers and different methods the maximum and minimum values do not coincide. For x_8 (use of Nitrogen fertilizer) parameter a_8 has maximum value in the BUGS variant and minimum value in the PCR variant. For x_9 (use of Phosphorous fertilizer) parameter a_9 has maximum value in the PCR variant and minimum value in the BUGS variant. For x_{10} (Use of Potassium fertilizer) parameter a_{10} has maximum value in the PCR variant and minimum value in the BUGS variant.

When the estimates of model parameter a_{11} (quality of land) are compared, one can see that the value ranges from 6.44 (variant FR) to 21.0 (variant OLS) in different variants, i.e. approximately three times. At the same time the sign of the parameters did not change. It should be mentioned, that parameter a_{11} is correlated with the mean value of fertilizers parameters (a_8 … a_{10}). In the variants, where the mean value of fertilizers parameters is high the value of parameter a_{11} is low. For example, for the variant FR the mean value of fertilizers parameters equals to 5.9 (the highest value) and parameter a_{11} equals to 6.44 (the lowest value). When the mean value of fertilizers is low the parameter a_{11} is high. For variant ANN the mean value of fertilizers is lowest (4.4) and parameters a_{11} value is relatively high (17.25).

It is interesting to note, that the sum of four parameters (a_8 … a_{11}) do not differ substantially for different estimation methods (lowest value is in variant FR – 24.06 and highest value is in variant BUGS – 30.8). Even in variants OLS and SVM (in both cases parameter a_{10} is negative) the sum is relatively high – correspondingly 28.6 and 28.1. Consequently, the combined effect of four independent variables x_8 … x_{11} is for all alternative variants practically the same. At the same time in different variants of models the combined effect of four variables is reallocated differently between variables of fertilizer use and land quality. Consequently, the influence of the fertilizer was included on coefficient of independent variable "land quality" when the classical model OLS and SVM models are used.

When the estimates of model parameter a_{12} (fraction of grain sown area in total sown area) are compared, one can see that the value ranges from –3.1 (variant SVM) to 11.3 (variant BUGS) in different variants, i.e. differ substantially and change sign (parameter a_{12} is in variant SVM negative).

For two variants (PCR and BUGS) in Table 2 are provided estimates for parameters a_{13} and a_{14} (parameters of quadratic variables x_8^2 and x_{12}^2). It should be mentioned, that signs of parameters a_{13} and a_{14} for variants PCR and BUGS are different. Economic theory asserts, that increasing the quantity of resource the effectiveness that resource decreases. Consequently, the parameter of quadratic variable must be negative.

Departing from that point of view and previous discussion, we should conclude, that BUGS variant is most acceptable, while all parameters are in accordance with economic theory.

In Table 2 are also provided the values of coefficients of determination of all variants. The coefficients are relatively high and ranges from 0.611 (variant PCR) to 0.885 (variant BUGS) in different variants.

Figure 2 illustrates the scatter plots of grain yield depending on predicted value for different model alternatives. Continuous line on graphs is predicted value of grain yield. The distance of points from continuous line is a residual. Figure 2 shows that for different models the graphs are different. Relatively analogous are graphs for BUGS model (C) and ANN model (D). Most different is the graph for PCR model (B). When for alternatives (C) and (D) the graphs have typical shape, the points intensity diminish in both ends of graph, then for PCR model (B) the points variability at upper end of graph is relatively high.

From Figure 2 it can be observed that in different models the same yield values are situated in different locations compared to continuous line (predicted value of grain yield). That means that the residuals are different in different models. Let us compare the situation of two concrete points on the graphs – those with values approximately 2500 kg/ha. Their accurate values are 2505 kg/ha (average yield in Järvamaa county in 1996) and 2490 kg/ha (average yield in Jõgevamaa county in 1996). From graphs A and B one can see that the actual yield in these counties exceeds the calculated yield and in both cases the yield of Järvamaa county is modelled with more precision (the difference between actual and calculated yield is smaller). At the same time the difference between actual and predicted yield is different in graphs. On the A the difference is 98 kg/ha and in case of B 385 kg/ha. From the C, D, E and F in Figure 2 we can see that in all cases the Järvamaa county yield is modelled accurately but the yield in Jõgevamaa county exceeds the calculated yield by approximately 200 kg/ha in all cases. From this analysis we can conclude that the yield is best modelled in case of Järvamaa county. There can be several reasons for that but these that must be investigated in future researches.

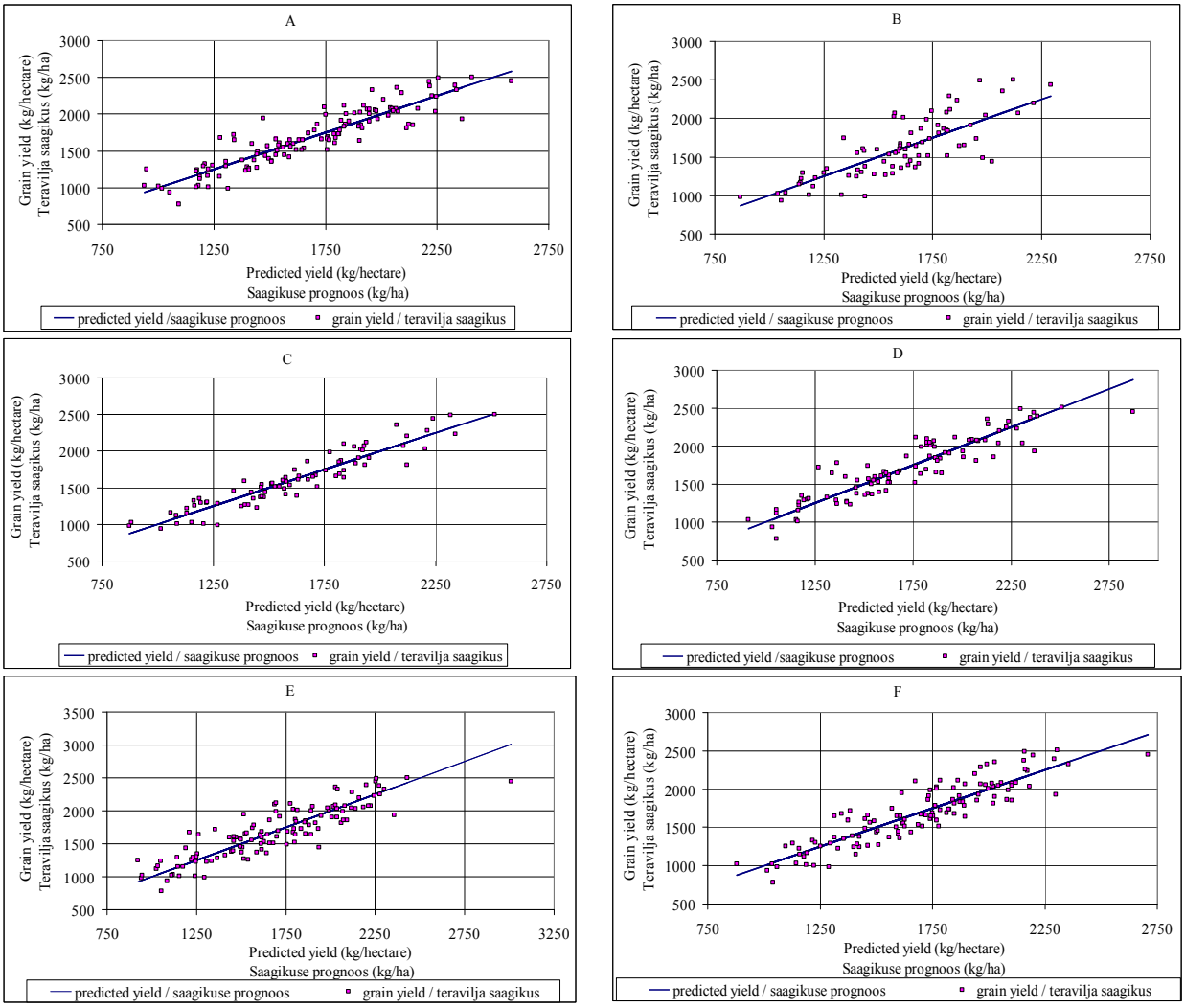

Figure 2. Relationships within grain yield and predicted grain yield for different model variants: OLS (A), PCR (B), BUGS (C), ANN (D), FR (E), SVM (F)
Joonis 2. *Tegeliku saagikuse ja prognoositava saagikuse vahelise sõltuvuse graafikud erinevate meetodite kasutamisel: OLS (A), PCR (B), BUGS (C), ANN (D), FR (E), SVM (F)*

Conclusions

Data mining is a fast expanding field with many new research results reported and new systems or prototypes developed recently. This article is an attempt to provide a reasonably comprehensive survey, from a user's point of view, on using different data mining techniques for estimating the parameters of econometric model of grain yield in Estonia. This paper can be viewed as an example of a new approach to one of such well-known problem, as variable selection for an econometric model.

The discussion may now be summarized in the following conclusions:

1. All considered methods may be used for estimating the parameters of the econometric model and may be recommended for such use.
2. Each method has its own advantages and disadvantages; there are no "silver bullets" in this case.
3. Our experience shows that Bayesian methods are most acceptable and widespread.
4. Estonian experience is still insufficient, and investigations (research) in this direction must be extended.

References

Friedman, J. H. 1997. Data Mining and Statistics: What's the Connection? http://www-stat.stanford.edu/~jhf/ftp/dm-stats.ps (Nov 1997).

Goebel, M., Gruenwald, L. 1999. A Survey of Data Mining and Knowledge Discovery Software Tools. SIGKDD Explorations, Vol 1, No 1, 1999, p. 20–33.

Kaashoek, J. F., van Dijk, H. K. (2000) Neural networks as econometric tool// Econometric Institute Rapport EI2000-31A / Econometric Institute, Erasmus University Rotterdam. – Rotterdam, 2000.

Kuan Chung-Ming, White, H. (1994) Artificial Neural Networks: An Econometric Perspective // Econometric Reviews. – 1994, No. 13, p. 1–92.

Kärblane, H. 1997. Mineraalväetiste kasutamisest. – Agraarteadus, nr 1, Tartu, lk 114–117.

Meyer, D. (2003) Support Vector Machines. The Interface to libsvm in package e1071. User Guide. December 10, 2003. http://cran.r-project.org/src/contrib/e1071_1.3-15.tar.gz.

Põldaru, R. 2000. Teravilja saagikuse ökonomeetriline mudel Eesti maakondade 1994.–1999. aastate andmete alusel. EAA väljaanne nr 13/2000. Põllumajanduse tulevik ja Euroopa Liit. Tallinn, lk 165–179.

Põldaru, R., Roots, J. 2001a. On the Implementation of the Principal Component Regression for the Estimation of the Econometric Model of Grain Yield in Estonian Counties. – International Scientific Conference Reports (Proceedings) "Problems and Solutions for Rural Development". Latvia University of Agriculture, Jelgava p. 340–345.

Põldaru, R., Roots, J. 2001b. On the Implementation of the Bayesian Statistics in Agricultural Reseach. In: Ulf Olsson and Jaak Sikk (Eds): Third Nordic – Baltic Agrometrics Conference. Jelgava, Latvia, May 24–26, 2001. Proceedings of the International Conference. Jelgava, Latvia University of Agriculture, 2001, p. 48–53.

Põldaru, R., Roots, J. 2001c. Bayesian Statistics (BUGS) in the Estimation of the Econometric Model of Grain Yield in Estonian Counties. EAA No. 15/2001, Agriculture in Globalising World, Proceedings (volume II) of International Scientific Conference, Tartu, p. 178–189.

Põldaru, R., Roots, J. 2002a. The Estimation of the Econometric Model of Grain Yield in Estonian Counties Using Neural Networks. Information Technologies in Agriculture: Research and Development. Theses of International Scientific Conference October 16–17, 2002, Kaunas, p. 14–16.

Põldaru, R., Roots, J. 2002b. Changes in Using Statistical Methods. Strategies of Rural Development: Theses of International Scientific Conference, Kaunas, Akademija p. 46–47.

Põldaru, R., Roots, J. 2003a. The Estimation of the Econometric Model of Grain Yield in Estonian Counties Using Neural Networks. "VAGOS", No. 57 Mokslo Darbai 57(10). Akademija, Kaunas, p. 124–130.

Põldaru, R., Roots, J. 2003b. Perspectives of Teaching and Training New Data Analysis Procedures. Information, Information Technology for Better Argi-Food Sector, Environment and Rural Living, Proceedings EFITA 2003, 4-th Conference of the European Federation for Information Technology in Agriculture, Food and Environment, 4–9 July, 2003. Debrecen-Budapest, Hungary, p. 525–530.

Põldaru, R., Roots, J., Ruus, R. 2003a. A Perspective of Using Data Mining in Rural Areas. Rural Development 2003. Globalization and Integration Challenges to the Rural Areas of East and Central Europe. Proceedings of International Scientific Conference. Kaunas, p. 256–257.

Põldaru, R., Roots, J., Ruus, R. 2003b. Implementation of Data Mining Methods in Agricultural Research and Education. In: Ulf Olsson and Jaak Sikk (Eds): Fourth Nordic – Baltic Agrometrics Conference, Uppsala, Sweden, June 15–17, 2003. Conference Proceedings, Uppsala, SLU, Department of Biometry and Informatics, Report 81, p. 109–118.

Põldaru, R., Roots, J., Ruus, R. 2003c. A Perspective of Using Data Mining (Association Rules) in Rural Areas. Transactions of the Estonian Agricultural University, No 218, Perspectives of the Baltic States' Agriculture under the CAP Reform, 19–20 September, 2003. Proceedings of International Scientific Conference, Tartu p. 184–199.

Põldaru, R., Roots, J., Ruus, R. 2003d. A Perspective of Using Data Mining in Rural Areas. "VAGOS", No. 61 Mokslo Darbai 61 (14). Akademija, Kaunas, 2003, p. 133–141.

Põldaru, R., Roots, J. 2004. Modeling Grain Yield in Estonian Counties: A Stochastic Frontier Analysis Approach. Data Envelopment Analysis and Performance Management. In: A. Emrouznejad, V. Podinovski (Eds.) Data Envelopment Analysis and Performance Management, Proceedings of 4th International Symposium of DEA, Aston University in Birmingham, Warwick print, UK, p. 261–267.

Põldaru, R., Roots, J., Ruus, R. 2004a. Using Fuzzy Regression in Rural Areas. Economic Science for Rural Development – Possibilities of Increasing Competitiveness, Proceedings of the International Scientific Conference No 7., Jelgava p. 43–48.

Põldaru, R., Jakobson, R., Roosmaa, T., Roots, J., Ruus, R., Viira, A.-H. 2004b. Support Vector Machine Regression in Estimating Econometric Model Parameters. Information Technologies and Telecommunication for Rural Development, Proceeding of the International Scientific Conference Jelgava, Latvia, 6–7 May, 2004, Jelgava, p. 66–77.

Põldaru, R. 2005. Implementation of Data Mining Methods in Agricultural Research. PhD paper. Estonian Agricultural University, Tartu, 2005, 149 pp.

Põldaru, R., Roots, J. 2005. The Estimation of the Econometric Model of Milk Yield per Cow: A Support Vector Machine Regression Approach, In: J. Boaventura Cunha, R. Morais (Eds.) 2005 EFITA/WCCA Joint Congres on IT in Agriculture. Proceedings of the EFITA/WCCA 2005 Joint Conference, Vila Real, Portugal, 25–28 July, 2005, p. 119–126.

Põldaru, R., Roots, J., Viira, A.-H. 2005a. Estimating Econometric Model of Average Milk Total Cost: A Support Vector Machine Regression Approach. "Economics and Rural development". Research papers, Vol 1(1), Akademija, Kaunas, 2005, p. 23–31.

Põldaru, R., Roots, J., Viira, A.-H. 2005b. Artificial neural network as an alternative to multiple regression analysis for estimating the parameters of econometric models, Agronomy Research 3(2), Tartu, Estonia 2005, p. 177–187.

Põldaru, R., Roots, J., Ruus, R. 2006. The Potential Use of Neural Network Models in Agricultural Research, In: U. Olsson and J. Sikk (Eds.): Fifth Nordic – Baltic Agrometrics Conference. Otepää, Estonia, June 15–17, 2005. Conference report. Uppsala, SLU, Department of Biometry and Informatics, Report 1 2006, p. 99–107.

Spiegelhalter, D., Thomas, A., Best, N., Gilks, W. 1996. BUGS 0.5. Bayesian inference using Gibbs sampling. Manual (version ii). MCR Biostatistics Unit, Institute of Public Helth, August 14, 1996.

Tanaka, H., Uejima, S., Asai, K. 1982. Linear Regression Analysis with Fuzzy Model. – IEEE Transactions on Systems, Man and Cybernetics, 12(6), 1982, p. 903–907.

Tanaka, H., Ishibuchi, H. 1992. Possibilistic Regression Analysis Based on Linear Programming. J. Kacprzyk, M. Fedrizzi (Eds.) Fuzzy Regression Analysis. Physica-Verlag, Heidelberg, 1992, p. 47–60.

Vapnik,V. 1998. Statistical Learning Theory, Springer, N.Y.

Teravilja saagikuse ökonomeetrilise mudeli parameetrite hindamine: andmekaeve meetoditel saadud analüüsitulemuste võrdlus

R. Põldaru, J. Roots, A.-H. Viira

Kokkuvõte

Antud artikli eesmärgiks on anda ülevaade andmekaeve meetodite rakendusvõimalustest ökonomeetrilise mudeli sõltumatute muutujate parameetrite hindamisel. Selleks on analüüsitud peamiste komponentide meetodi (PCR), Bayesi statistiliste meetodite (BUGS), tehisnärvivõrkude meetodi (ANN), hägusa regressiooni (FR) ja tugivektorite regressiooni meetodite (SVR) sobivust teravilja saagikuse ökonomeetrilise mudeli parameetrite hindamisel ja saadud tulemusi on võrreldud vähimruutude meetodil (OLS) saadud lahenditega.

Erinevate meetodite võrdlus näitab, et alternatiivseid andmetöötluse meetodeid on võimalik kasutada ökonomeetriliste mudelite parameetrite hindamisel, kusjuures mõningad meetodid konkureerivad omavahel. Ökonomeetrilise mudeli parameetrite hinnangud olid analoogilised Bayesi statistiliste meetodite, tehisnärvivõrkude ja hägusa regressiooni meetodil saadud lahendite korral ja vastavuses majandusteooria ning -praktika kogemustega. Kõige tõepärasem on Bayesi meetodil saadud lahend, mis on saavutatud eeskätt seetõttu, et väetiste regressioonikordajatele on ette antud küllaltki väikese varieeruvusega aprioorsed jaotused.

Edasist uurimist vajab see, miks teatud maakondade (Järvamaa) kohta annavad kõik võrreldud meetodid kõige täpsemaid tulemusi.