DE GRUYTER OPEN
DOI: 10.1515/agriceng-2016-0043

ISNN 2083-1587; e-ISNN 2449-5999
2016,Vol. 20,No.3, pp.43-51
Agricultural Engineering
www.wir.ptir.org

VERIFICATION OF ECONOMIC AND AGRICULTURAL INDICATORS WITH THE USE OF STATISTICAL METHODS ON THE EXAMPLE OF INDIVIDUAL FARMS

Katarzyna Grotkiewicz*, Agnieszka Peszek, Zbigniew Kowalczyk

Institute of Agricultural Engineering and Informatics, University of Agriculture in Krakow

* *Corresponding author: email: katarzyna.grotkiewicz@ur.krakow.pl*

ARTICLE INFO

Article history:
Received: March 2016
Received in the revised form:
April 2016
Accepted: June 2016

Key words:
economical and agricultural indicators,
individual farms,
data exploration,
two step cluster analysis

ABSTRACT

The paper describes basic methodology assumptions related to construction of Bayesian networks. The paper aims at preparation of data for modeling to obtain fresh knowledge on economic and agricultural database and will constitute the first stage of research. Variables (economic and agricultural indicators) with discreet values were used for analysis with the use of two step grouping technique and previous non-typical data were explored. The research was carried out on the group of three hundred individual farms from Małopolskie and Świętokrzyskie Voiovedship. The knowledge obtained from analyses will be used in practice in agricultural engineering in order to support agricultural activity.

Introduction

Research on the scientific and technical progress and its relation to work performance and land efficiency in agriculture constitute one of the leading research trends in agricultural engineering in the country and they are popular in the world literature (Michałek et al., 1998; Tabor, 2006; Kołodziejczak, 2008; Neal, 2014; Prisecaru, 2015). This trend originates not only in the need of searching for complex measures of assessment of the level of intensity and modernity of agriculture but also in observation and direct cooperation with agricultural practice. Based on the extensive studies of the national and foreign literature and in particular those presented in the monography by Grotkiewicz et al., (2013) the attention was paid not only to positive aspects of suggested and tested methods of assessment of the intensity level of agriculture but also to methodological weaknesses of these methods, which finally influence the final research results, were emphasised (Grotkiewicz and Kowalczyk, 2015). Basic statistical methods were used for comparison of significance of differences between the analysed indicators i.e. work performance and land efficiency and scientific and technical progress and at the same time relations between them were searched for with the use of fundamental statistical methods using a single – factor analysis of variance and analysis of correspondence (Michałek et al., 2010; Grotkiewicz et al., 2013).

Another approach to present a relation between the analysed economic and agricultural indicators is investigation of relations between variables and more strictly between distribution of discreet variables basing on the probability theory with the use of Bayesian networks. A bayesian network is an acyclic graph which includes the quality part which constitutes a set of variables - graph nodes along with relations between them and the quantity part which represents probability distribution for these variables (Campos and Javier, 2007; Kusz and Marciniak, 2006; Bartnik et al., 2006).

A Bayesian network is a method of representing data which gives possibilities for conclusions (Kusz and Marciniak, 2006). Based on the quality description (namely a graphic model structure) identification of conditional relations between variables (quantity and quality) is possible. Moreover, a farm model which meets the indispensable conditions to achieve the increase of productivity and factors which shape them can be construed. Concluding in the Bayesian network comes down to indication of probability distribution a posteriori if the model variable values are reported (Aczel, 2005). Distribution of this probability may be directly used in supporting investment decisions (Kusz and Marciniak, 2006). A possibility of using the algorithm of the Bayesian network for obtaining knowledge on economic and agricultural data base requires their previous preparation for modeling. The need to carry out analysis is justified not only because of the economic data variability to reflect the present market situation in the best possible way but also there is a need to exclude the values, which may be recognized as non-typical, possibly burdened with error which consequently will influence the final concluding.

The objective of the research

The basic objective and thus the first stage of this research is verification and preparation of data for Bayesian modeling in order to obtain fresh knowledge on economic and agricultural data bases.

Based on the data obtained from the group of 300 individual farms the exploration review was carried out and Two Step Cluster Analysis was carried out. The effect of research will be obtaining information on non-typical values in the set and on grouping values, which will be analysed.

Material and research methods

Material collected for research includes small farms from 10 municipalities located in the southern and central and southern Poland i.e. from Małopolskie and Świętokrzyskie Voivodeship. 30 individual farms were selected from each municipality. Both municipalities and specific farms within municipalities were selected purposefully taking into consideration specific criteria (Grotkiewicz et al., 2013). The scope of questions was vast and included issues concerning: the land use structure, sowing structure, livestock, size and value of the machinery park, material inputs and their size, working force inputs, level of provided mechanical services, global production, clear production. Based on the collected number characteristics and calculated economic and agricultural indicators, the following quantity indicators were subjected to analysis, i.e. scientific and technical progress, work performance, land efficiency, global production, clear production. Methodology of calculation indicators of work performance and land efficiency, scientific and technical progress and the remaining economic and agricultural indicators were presented in previous papers (Michałek et al., 2008; Grotkiewicz and Michałek, 2009; Grotkiewicz et al., 2013).

In order to analyse data from economic and agricultural base, firstly the input data which constitute 300 individual farms were explored. This analysis aimed at searching for and excluding non-typical data (Chen et al., 1996). Necessity of carrying out the analysis of this type results from the process of preparation for modeling. It is also justified by assumptions which refer to the values introduced to the algorithm of the Bayesian network (Morzy, 2007). For this purpose it was decided to use the analytical technique which searches for non-typical observations in data records on account of the values of all investigated variables and not only values of a single variable. The analytical procedure searches for non-typical observations in the group of data which is based on deviations from standards calculated for the formed concentrations (groups of records with similar values).

Each separate property was subjected to analysis (TwoStep Cluster Analysis) which aims at grouping values. Euclidean distance, relevant for the quantity nature of properties was selected as a probability measure of clusters and Schwarz's Bayesian information criteria which assesses adjustment of the model to the data subjected to analysis, was used for assessment of the number of clusters.

IBM SPSS Statistics 23 is the program which was used for exploration analysis and TwoStep Clustering.

Research results

The task on the exploration analysis was to search for and exclude incomplete, incorrect or insignificant data from the set of the explored data. Such data may be created at the initial stage of data collection, e.g. through providing by respondents untrue values or lacking values in some variables in the process of imputation, when data are incomplete. At the beginning of analysis, routine investigation of distribution of features representing economic and agricultural indicators were carried out and their relations were researched with reference to the area variable where non-uniformity of the value distribution for the analysed properties and border and outstanding values were reported.

Basic descriptive statistics for the investigated variables were listed in the following table (tab.1).

When analysing values and distribution of variables it was decided that for the possessed data set an analysis should be carried out which identifies non-typical observations basing on standard deviations. Excluding only border and outstanding values could remove real values, which despite the fact that they diverge from the standards indicated by measures of central tendency for the investigated properties they can describe the whole spectrum of values accepted by particular properties.

In the input parameters of the algorithm it was assumed that not more than 5% of non-typical observations will be looked for on account of all investigated economic and agricultural indicators. In total, the algorithm found out 15 such observations recognized as non-typical after the analysis of values of all variables was carried out. The table below (tab. 2) presents the number of records of the data set, which were selected on account of the value of the specific variable which determined the observation as non-typical.

Table 1.
Descriptive statistics for the investigated economic and agricultural indicators.

Economic and agricultural indicators	N	Gap	Minimum	Maximum	Average	Standard deviation	Skewness		Kurtosis	
	Statistics	Statistics	Statistics	Statistics	Statistics	Statistics	Statistics	Standard error	Statistics	Standard error
PG, (PLN thous.)	300	269.150	0.670	269.820	46.253	43.786	2.323	0.141	6.475	0.281
PC, (PLN thous.)	300	491.530	-290.190	201.340	24.071	40.552	-1.881	0.141	23.912	0.281
WZ, (PLN thous.·ha^{-1})	300	136.330	-87.940	48.390	3.597	7.929	-3.101	0.141	68.160	0.281
WP, (PLN thous.·mhr^{-1})	300	0.413	-0.184	0.229	0.015	0.024	1.134	0.141	36.136	0.281
PT, (PLN thous.·mhr^{-1})	300	0.923	-0.686	0.237	0.004	0.068	-5.923	0.141	60.677	0.281

Table 2.
Reasons for recognition of observations as non-typical for economic and agricultural indicators

Economic and agricultural indicators	Incident as a reason		Statistics of the impact variable			
	Frequency	(%)	Minimum	Maximum	Average	Standard deviation
PG, (PLN thous.)	5	33.3	0.468	0.662	0.571	0.093
PC, (PLN thous.)	3	20.0	0.499	0.561	0.536	0.033
WZ, (PLN thous.·ha^{-1})	1	6.7	0.387	0.387	0.387	0.000
WP, (PLN thous.·mhr^{-1})	1	6.7	0.813	0.813	0.813	0.000
PT, (PLN thous.·mhr^{-1})	5	33.3	0.524	0.939	0.737	0.151

As a result of the above analysis, a new data set was obtained. It was reduced by 15 records in comparison to the original set which includes 300 observations. This new set including 285 records is an input set for preparing data for modeling with the use of the Bayesian network which will be carried out in the second stage of research.

Grouping values in economic and agricultural indicators

Bayesian networks require in the analysis properties with discreet values. However, the analysed indicators have constant values. These are variables measured in the quantity scale. At discretization in the first part it was decided to divide the values of each variable with regard to natural clusters of values in particular properties. For this purpose, a classification method was applied known under the name of Two Step Cluster Analysis (IBM, 2016). The advantage of this method in comparison to other grouping methods is, inter alia, an assumption of independence of variables which enables the analysis of variables with combined multinormal distributions (Şchiopu, 2010; Park and Bartnik, 2006). The Bayesian Information Criterion (BIC) also known as Schwarz's Bayesian information criterion is used for assessment of the number of clusters (Piłatowska, 2011). Each separate property was subjected to analysis which aims at grouping values. Euclidean distance relevant for the quantity nature of properties was selected as a probability measure.

Below the table was presented (tab. 3) with the results of two step grouping with the use of Schwarz 's Bayesian information criterion (BIC) for global production.

Table 3.
Evaluation of the number of clusters with the use of Schwarz's Bayesian information criterion for global production

Number of clusters	Schwarz information criterion (BIC)	BIC changes	BIC change quotient	Distance measure quotient
1	208.351			
2	93.121	-115.231	1.000	1.714
3	89.384	-3.737	0.032	3.193
4	98.719	9.335	-0.081	1.100
5	83.475	-15.244	0.132	1.840
6	92.724	9.249	-0.080	2.115
7	103.821	11.097	-0.096	1.121
8	99.648	-4.173	0.036	1.192
9	109.856	10.208	-0.089	1.449
10	120.430	10.573	-0.092	1.355
11	129.376	8.946	-0.078	1.011
12	140.590	11.214	-0.097	1.617
13	151.546	10.956	-0.095	1.079
14	160.732	9.186	-0.080	1.067
15	171.892	11.160	-0.097	1.054

Values of Schwarz's Bayesian criterion in case of the PG variable indicate that the best model which divides values of this variable into 5 clusters (groups) was selected. Since this criterion evaluates adjustment of the model to data. The lower are the BIC values the better is adjustment (Kapłon, 2009).

As a result of two step grouping with the suggested number of clusters for the PG cluster 5 groups of varied quantities and distributions of values, which are presented in the following table, were obtained (tab. 4).

Table 4.
Characteristic of clusters of discreted variable of global production PG

PG, (PLN thousand)							
Cluster	Number	% of N Total in table	Minimum	Maximum	Average	Median	Standard deviation
1	218	76.5	0.67	52.75	26.59	25.30	12.80
2	38	13.3	53.85	97.12	71.65	70.48	13.62
3	21	7,4	108.38	162.91	131.24	130.00	16.00
4	5	1,8	169.82	215.60	197.14	198.98	17.16
5	3	1,1	239.38	269.82	256.22	259.46	15.48
Total	285	100,0	0.67	269.82	45.72	32.41	44.23

A similar scheme of operation was assumed in case of the analysed variable.

Thus, assuming the same statistics for assessment of the model in case of analysis of clean production property its division into 5 clusters was recognized. The following table no. 5 presents its characteristics.

Table 5.
Characteristics of clusters of digitized variable of clean production PC

PC, (PLN thous.)							
Cluster	Number	% of N Total in table	Minimum	Maximum	Average	Median	Standard deviation
1	1	0,4	-281.300	-281.300	-281.300	-281.300	0.000
2	1	0.4	-62.090	-62.090	-62.090	-62.090	0.000
3	240	84,2	-7.760	43.670	15.086	13.595	10.872
4	27	9,5	44.220	89.030	67.071	68.360	15.057
5	14	4,9	102.000	148.000	118.113	115.020	15.427
6	2	0,7	164.300	201.340	182.820	182.820	26.191
Total	285	100,0	-281.300	201.340	24.938	16.000	36.670

For land efficiency WZ two step grouping algorithm divided this variable into 6 clusters. Table 6 contains their characteristics

Table 6.
Characteristic of clusters of digitized variable of land efficiency WZ

WZ, (PLN thous.·ha^{-1})							
Cluster	Number	% of N Total in table	Minimum	Maximum	Average	Median	Standard deviation
1	3	1.1	-5.310	-1.390	-3.613	-4.140	2.012
2	223	78,2	-1.230	4.920	2.145	2.150	1.354
3	49	17,2	4.940	11.800	7.346	6.910	1.805
4	5	1.8	18.170	23.710	20.762	20.340	2.020
5	2	0.7	26.370	30.070	28.220	28.220	2.616
6	3	1.1	45.570	48.390	47.283	47.890	1.505
Total	285	100.0	-5.310	48.390	3.964	2.660	6.018

Work performance WP variable has 5 separate values determined based on the number of five clusters distinguished in the two step grouping analysis of the value of this variable (tab. 7).

Table 7.
Characteristic of clusters of digitized variable of work performance WP

WP, (PLN thous.·mhr^{-1})							
Cluster	Number	% of N Total in table	Minimum	Maximum	Average	Median	Standard deviation
1	1	0.4	-0.077	-0.077	-0.077	-0.077	0.000
2	247	86,7	-0.022	0.027	0.010	0.009	0.007
3	30	10.5	0.028	0.064	0.041	0.039	0.011
4	6	2.1	0.078	0.112	0.093	0.094	0.013
5	1	0,4	0.229	0.229	0.229	0.229	0.000
Total	285	100.0	-0.077	0.229	0.015	0.010	0.022

The PT variable was digitized with TwoStep Clustering algorithm to the 5th value. Table no. 8 presents characteristics and distribution of clusters of constant variable PT which defines the value of scientific and technical progress.

Table 8.
Characteristic of clusters of digitized variable of scientific and technical variable PT

PT, (PLN thous.·mhr^{-1})							
Cluster	Number	% of N Total in table	Minimum	Maximum	Average	Median	Standard deviation
1	3	1.1	-0.686	-0.188	-0.503	-0.635	0.274
2	2	0.7	-0.150	-0.086	-0.118	-0.118	0.045
3	252	88.4	-0.082	0.042	0.001	0.001	0.016
4	24	8.4	0.042	0.115	0.078	0.076	0.021
5	4	1.4	0.185	0.237	0.206	0.201	0.025
Total	285	100.0	-0.686	0.237	0.004	0.001	0.068

As a part of cluster analysis additional verification of correctness of their separation through the use of another analytical technique was carried out. This verification aimed at ensuring as to the nature and dimension of clusters. Analysis was carried out on the same set of data and for the same constant variables using the algorithm of hierarchical analysis of clusters (method of single binding). As a result of the analysis very similar clusters of variable records as those obtained for TwoStep Clustering were obtained. There are differences in the size of clusters but for the further course of analysis they are not considerable.

Conclusion

Based on the results of analysis which were carried out in order to prepare data for Bayesian modeling an exploration data review was performed and on its basis a new set reduced by 15 records in comparison to the original one was separated. Then, a two - step grouping algorithm was used, which unfortunately did not give the best effects from the point of view of division of values into groups with similar and at the same time considerable number. It results from the nature and distribution of the value of the investigated variables which are very consolidated. Basically in the natural distribution of values there is usually only one (sometimes two) big clusters of values and several small number clusters (and even single points) for the outstanding values. Probably one should carry out further reduction of the data set which aims at elimination of the outstanding values. However, on the other hand, such treatment could remove from the data set real values which describe the situation of the investigated farms with economic and agricultural indicators in a non-uniform manner. Very similar clusters of variable records as those obtained for TwoStep Clustering using the algorithm of hierarchical analysis of clusters which aim at verification of their separation were obtained. There are differences in the size of clusters but for the further course of analysis they are not considerable. Thus, in the further part of the analysis, it was decided to use results from grouping carried out with the two step cluster analysis.

The obtained knowledge from the analyses which were carried out may be used in direct practice with the use of Bayesian networks for developing an optimal model of agricultural farm located in a well - organized technical, economical and information infrastructure. Moreover, this model should achieve high indicators which present the level of intensity and competitiveness of agriculture and thus work performance and land efficiency which in turn will be presented in the second stage of research.

References

Aczel, A.D. (2005). *Statystyka w zarządzaniu*. PWN. ISBN 83-01-14548-X.

Bartnik, G., Kusz, A., Marciniak, A. W. (2006). Modelowanie procesu eksploatacji obiektów technicznych za pomocą dynamicznych sieci bayesowskich. *Agricultural Engineering, 12*(87), 9-16.

Campos, L.M., Javier, G.C. (2007). Bayesian network learning algorithms using structural restrictions. *International Journal of Approximate Reasoning, 45*(2), 233–254.

Chen, M.S., Han, J., Yu, P.S. (1996). Data mining:an overview from a database perspective. IEEE Trans. *Knowledge and Data Enginieering, 8,* 866-883.

Grotkiewicz, K., Kowalczyk, Z. (2015). Methodological notes concerning determination of the scientific and technical progress rate and its efficiency. *Agricultural Engineering, 4*(156), 149-156.

Grotkiewicz, K., Kuboń, M., Michałek, R., Peszek, A. (2013). *Postęp naukowo-techniczny w procesie modernizacji polskiego rolnictwa i obszarów wiejskich.* Wydawnictwo Inżynieria Rolnicza. ISBN 978-83-935020-5-9.

Grotkiewicz, K., Michałek, R. (2009). Ocena poziomu produkcyjności i wydajności w rolnictwie na przykładzie wybranych regionów Polski. *Agricultural Engineering, 6*(115), 103-108.

IBM, Knowledge Center. (on-line). Obtained from: https://www.ibm.com/support/ knowledgecenter/SSLVMB_21.0.0/com.ibm.spss.statistics.help/idh_twostep_main.htm.

Kapłon, R. (2009). Rozkład a priori w czynniku bayesowskim a wybór modelu klas ukrytych. *Badania Operacyjne i Decyzyjne, 3,* 87-94.

Kołodziejczak, M. (2008). Efektywność wykorzystania zasobów pracy i ziemi w rolnictwie Unii Europejskiej. *Roczniki Naukowe SERiA, Tom X, Zeszyt 1*, 176-181.

Kusz, A., Marciniak, A. W. (2006). Dynamiczne sieci probabilistyczne jako system reprezentacji wiedzy. *Agricultural Engineering, 12*(87), 285-294.

Michałek, R., Grotkiewicz, K., Kuboń, M., Sporysz, M. (2010). Metodyczne aspekty określania postępu naukowo-technicznego w badaniach makro- i mikroekonomicznych. *Agricultural Engineering, 5*(123), 197-205.

Michałek, R., Kowalski, J., Tabor, S., Cupiał, M., Kowalski, S., Rutkowski, K. (1998). *Uwarunkowania technicznej rekonstrukcji rolnictwa.* Wydawnictwo PTIR, ISBN 83-905219-1-1.

Michałek, R., Peszek, A., Grotkiewicz, K. **(**2008). Wydajność pracy i ziemi w wybranych gminach województwa małopolskiego. *Agricultural Engineering, 10*(108), 185-191.

Morzy, T. (2007). Eksploracja danych. *Nauka, 3*, 83-104.

Neal, K. van Alfen (2014). Policy Frameworks for International Agricultural and Rural Development. *Encyclopedia of Agriculture and Food Systems, Volume 5,* 489-50.

Park, H.S., Baik, D.K. (2006). A study for control of client value using cluster analysis. *Journal of Network and Computer Applications, Vol. 29,* No. 4, 262-276.

Piłatowska, M. (2011). *Porównanie kryteriów informacyjnych i predykcyjnych w wyborze modelu.* Prace i materiały Wydziału Zarządzania UG (8).

Prisecaru, P. (2015). EU Reindustrialization on the Coordinates of Scientific and Technical Progress. *Procedia Economics and Finance, 22*, 485-494.

Rocznik statystyczny województw. 2014. GUS. ISSN 1230-5820

Şchiopu, D. (2010). Applying TwoStep Cluster Analysis for Identifying Bank Customers'Profile. *Seria Ştiinţe Economice, LXII, 3*, 66-75.

Tabor, S. (2006). Postęp techniczny a efektywność substytucji pracy żywej pracą uprzedmiotowioną w rolnictwie. *Agricultural Engineering, 10*(85), ISSN 1429-7264.

WERYFIKACJA WSKAŹNIKÓW EKONOMICZNO-ROLNICZYCH Z WYKORZYSTANIEM METOD STATYSTYCZNYCH NA PRZYKŁADZIE GOSPODARSTW INDYWIDUALNYCH

Streszczenie. W pracy omówiono podstawowe założenia metodyczne związane z budową sieci bayesowskich. Zadaniem opracowania jest przygotowanie danych do modelowania w celu pozyskiwania nowej wiedzy z ekonomiczno-rolniczej bazy danych i będzie ono stanowiło jednocześnie pierwszy etap badań. Do analiz wykorzystano zmienne (wskaźniki ekonomiczno-rolnicze) o wartościach dyskretnych posługując się techniką dwustopniowego grupowania oraz dokonano wcześniejszej eksploracji danych nietypowych. Badania przeprowadzono na grupie trzystu gospodarstw indywidualnych z województwa małopolskiego i świętokrzyskiego. Uzyskaną wiedzę z przeprowadzonych analiz będzie można wykorzystać w bezpośredniej praktyce w inżynierii rolniczej mając na celu wspomaganie działalności rolniczej.

Słowa kluczowe: wskaźniki ekonomiczno-rolnicze, gospodarstwa indywidualne, eksploracja danych, dwustopniowa analiza skupień