DE GRUYTER
OPEN

DOI: 10.1515/agriceng-2017-0017

ISNN 2083-1587; e-ISNN 2449-5999
2017,Vol. 21,No.2, pp.69-79
Agricultural Engineering
www.wir.ptir.org

APPLICATION OF BAYESIAN NETWORKS FOR FORECASTING FUTURE MODEL OF FARM

Katarzyna Grotkiewicz

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

* *Corresponding author: e-mail: katarzyna.grotkiewicz@ur.krakow.pl*

ARTICLE INFO

Article history:
Received: November 2016
Received in the revised form:
January 2017
Accepted: January 2017

Key words:
Bayesian Networks,
forecasting,
economic and agricultural indicators,
farm

ABSTRACT

Comparative analyses in the national scale were carried out in 300 individual farms from Małopolskie and Świętokrzyskie Voivodeship in order to search for relations between the production intensity level, work performance and land efficiency and factors which shape them. The analyses concerned the use of Bayesian modelling algorithms for forecasting development of various economic and agricultural indicators which decide on the intensity and competitiveness of agriculture. The paper constitutes the second stage of research, which was preceded with previous preparation of data for modelling with the use of an exploratory overview of available data and TwoStep Cluster Analysis (Grotkiewicz et al., 2016). Based on the analyses, which were carried out, networks were built which present the relations between the analyzed variables, and conditional similarities were verified.

Introduction

After Poland's accession to the European Union, the role of agriculture has gained a greater significance particularly for individual farms which have a small area structure, which may be especially visible in Małopolskie Voivodeship, where the average size of farms is 4.02 ha (ARiMR, 2016) and thus it is the lowest average in relation to the remaining regions of Poland. Common Agricultural Policy has become a chance for not only economic but also social development, which thus affected intensification of farms which produce only for their own needs and have no perspectives for future. Farms, which developed their potential by means of scientific and technical progress using the EU aid, improved living conditions and thus affected the increase of economic and agricultural indicators ensuring farmers with an appropriate level of incomes and living conditions (Grzegorek, 2012; Grotkiewicz et al., 2013). There are many various indicators of assessment of the level of intensity and modernity of agriculture in literature. However, from the point of view of competitiveness both on the national and international arena, indicators of work and land productivity in agriculture are on the top. In previous research in the design stage, numerous statistical analyses were carried out, with the use of e.g. Duncan's test and analysis of correspondence, for statistical assessment of significance of differences of particular indicators and those, which relate to the impact of scientific and technical progress, on the indicators of work performance and land efficiency and factors which shape them.

Another method, which may be used for direct practice concerning economic and agricultural indicators is the Bayesian Network, which based on the quality description (namely a graphical model structure) enables identification of conditional relations between variables (quantity and quality) i.e. economic and agricultural indicators and gives an opportunity to build a future model of a farm which meets the conditions indispensable for achieving an economic growth including present indicators of the agriculture level. The Bayesian Network, also known as a probabilistic one, is a method of data representation, which gives an opportunity to make conclusions (Bartnik et al., 2005) which as a result means determination of the a posteriori probability distribution under the condition that variable values of the model (Aczel, 2005) are reported. Moreover, it may be a useful tool, inter alia, for modelling reliability and supporting decisions in the conditions of uncertainty concerning e.g. economic processes (Bartnik et al., 2006; Kusz et al., 2015) as well as in the processes of the environment quality management (Sujak et al., 2016).

Objective of the research

Based on the previous analyses, concerning data preparation for modelling (Grotkiewicz et al., 2016), this paper constitutes the second stage of research, whose fundamental objective consists in an assessment of usefulness of the Bayesian Networks for forecasting development of economic and agricultural indicators with the use of various algorithms of Bayesian modelling and for obtaining new knowledge from the economic and agricultural data base. The effect of the research will consist in obtaining information on the probability distribution. There, a conditional probability distribution that a given component (feature) is in a given state (group or the value class) is related with each node of a network, preconditioned with the components' state (factors) represented by components related thereto (Bartnik et al., 2006).

Methodology of research

The second stage of research and thus verification of the obtained models obtained by means of the Bayesian Networks will be carried out based on the data collected from ten municipalities of Małopolskie and Świętokrzyskie Voivodeship. The collected data include the quantity variables i.e. scientific and technical progress PT, work performance WP, land efficiency WZ, global production PG, clean production PC and quality variables (production trend, degree of simplification of plants cultivation, work expenditures). Finally, tests were carried out on the group of three hundred individual farms from the southern Poland. Popular methodology used in many scientific papers was applied for calculation of economic and agricultural indicators in particular of work performance and land efficiency (Grotkiewicz et al., 2013; Tabor, 2006; Gębka, Filipiak, 2006).

However, before the construction of the network was initiated, in the first stage of research, exploration of the input data including 300 individual farms from Małopolskie and Świętokrzyskie Voivodeship was carried out (Grotkiewicz et al., 2016). Exploration analysis aimed at possible searching for and exclusion of non-typical data. Necessity of carrying out the analysis of this type results from the process of preparation for modelling.

(Michałek et al., 2010). It is also justified by assumptions, which refer to the values introduced to the algorithm of the Bayesian Network (Morzy, 2007).

In the first approach to build a model based on the Bayesian Networks all variables with discrete values were used after non-typical data were explored. Discrete values for variables (economic and agricultural indicators: PT, WP, WZ, PG, PC) were created as a result of analysis carried out in the first stage of research which uses a Two Step Cluster Analysis (Grotkiewicz et al., 2016).

The results of the research concerning data exploration and grouping of values with Two Step Cluster Analysis were presented in the paper by Grotkiewicz et al., (2016) which constitute at the same time the first stage of research.

According to the literature (Aczel, 2005) Bayes' statement is strictly related to the total probability and it may be presented in a different form with the use of the conditional probability concept with the following formula:

$$P(\mathrm{B}|\mathrm{A}) = \frac{P(\mathrm{A}|\mathrm{B})P(\mathrm{B})}{P(\mathrm{A}|\mathrm{B})P(\mathrm{B})+P(\mathrm{A}|\overline{\mathrm{B}})\mathrm{P}(\overline{\mathrm{B}})} \quad (1)$$

According to the formula, *B* probability under the condition *A* is calculated based on the knowledge of *B* and $\overline{B}$ probability and conditional probabilities *A* at the assumption *B* and *A* at the assumption $\overline{B}$. Probabilities *P(B)* and *P($\overline{B}$)* are called a priori probabilities of events *B* and $\overline{B}$. *P(B|A)* is called a posteriori probability of event *B* (Aczel, 2005).

In such a meaning, the Bayesian concept enables the use of both quality data (production trend, degree of simplification of plants cultivation, work expenditures) and quantity ones (inter alia: scientific and technical progress, work performance and land efficiency) in order to present trends in the changes of scientific and technical progress and efficiency of the progress in the national scale in comparison to the changes in the work performance and land efficiency and factors which shape them, which decide on the level of productivity and social efficiency in the Polish agriculture.

GeNie 2.0 program will be used for Bayesian analyses. It offers a scientific environment for construction and testing of prediction models which are based on various algorithms of Bayesian networks (Jongsawat et al., 2010). This program was prepared in the Decision Systems Laboratory, University of Pittsburgh.

Building of Bayesian networks

The Bayesian Networks serve for presenting relations between events (such as e.g. occurrence of a specific value of a given feature) based on the probability calculus. They are graphical models, which serve for presenting the total distribution of probability (with the use of which we may also predict occurrence of values for the indicated economic and agricultural indicators). Nodes of the network are variables (properties with discrete values) and the connections of nodes (arcs or vectors) reflect relations between properties and their direction. Thickness of links between nodes on the graphs, which present the Bayesian networks, symbolizes strength between variables (Oniśko et al., 2001; Bartnik, Kusz, 2005). If a pair of nodes in the network is not connected then variables corresponding thereto are (conditionally) independent.

In the first approach to build a model based on the Bayesian networks all variables with discrete values present in the investigated data set were applied. Discrete values for varia-

bles (economic and agricultural indicators) were formed as a result of the previous analysis which used Two Step Clustering technique (Grotkiewicz et al., 2016), where numbers and percentage values of discretized variables (numbers of clusters) referring to ranges of continuous variables values were presented in tables.

The below diagram presents a network built on GeNie program based on information from previous analyses (Grotkiewicz et al., 2016) and using discretized variables of economic and agricultural variables i.e. scientific and technical progress (PT_discrete), work peformance (WP_discrete), land efficiency (WZ_discrete), global production (PG_discrete), clear production (PC_discrete) and nominal variables i.e. production trend (Production_trend), degree of simplification of plants cultivation (Cultivation), work expenditures (Work_expenditures). A network diagram presents also values of algorithm parameters (Bayesian Search) which builds a network.

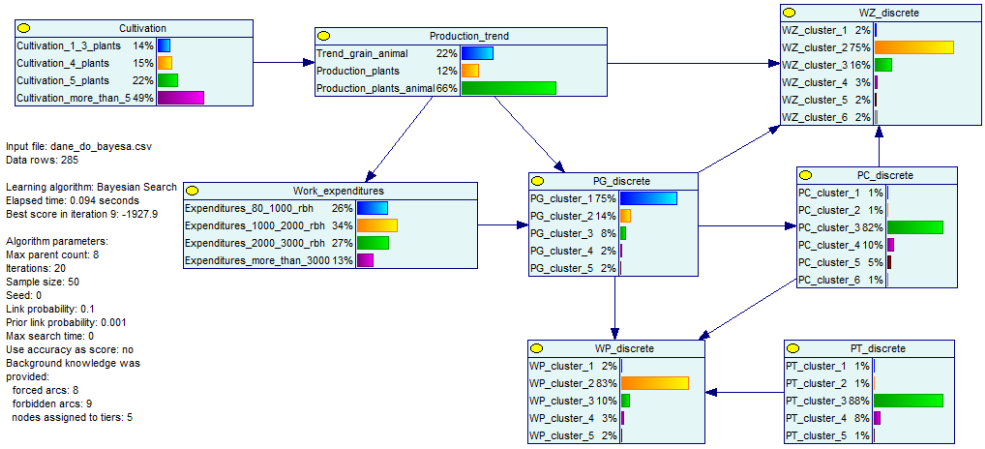

Figure 1. Distribution of conditional probabilities for agri-economic indicators

Probability of occurrence of values in a particular cluster of variable values (continuous) and its percentage value were presented in the knots of the network in the form of a bar chart.

Based on the possibility of checking conditional probabilities which take place if a specific event occurs (e.g. a value belonging to a specific cluster of a preceding node in a network), numerous various analyses were carried out in order to obtain a table of conditional probabilities for the values of particular network nodes.

An example of such calculation was presented below, where on a diagram values of conditional probabilities for the nodes PC_discrete, WZ_discrete, WP_discrete, PT_discrete are visible, which were calculated if an event occurs, that the value of PG_discrete value belongs to cluster 2 (probability of this event equals 1).

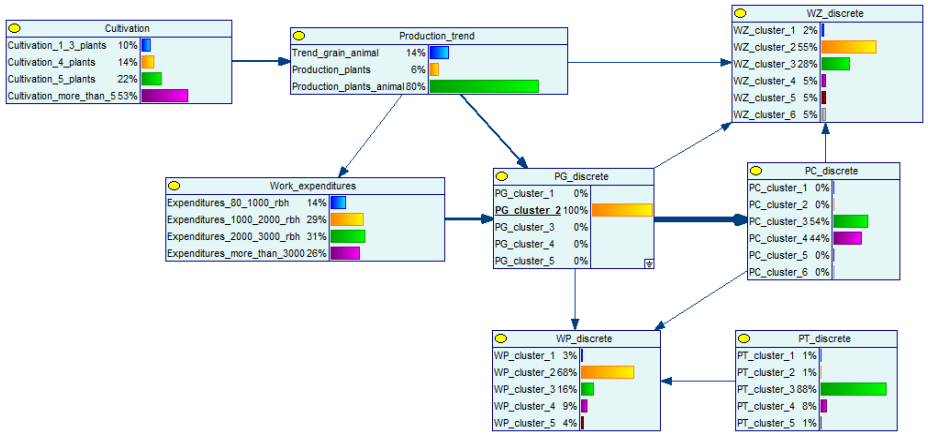

Figure 2. Distribution of conditional probabilities for economic and agricultural indicators if the event PG, belonging to the second cluster takes place.

Occurrence of such event (e.g. PG_discrete belongs to the range of values <53.85; 97.12> causes a change of probabilities in some nodes of the network which depend on PG_discrete node. For instance, a probability of occurrence of the event that WZ_discrete value will belong to cluster 2, drops (from 75% to 55%) (i.e. WZ will belong to the value range <-1.23; 4.95>).

On the other hand, the increase of conditional probability (to 87%) for the same dependable event (i.e. that value WZ_discrete belongs to cluster 2) is reported in a situation when PG_discrete value belongs to more numerous cluster 1 referring to the range of values <0.67; 52.75> variable PG. In such case, probability that the value of PC variable belongs to cluster 3 (PC with values from the range of <-0.10; 4.20>) raises to 99% and chances that PC value will be found in another range is practically levelled to zero.

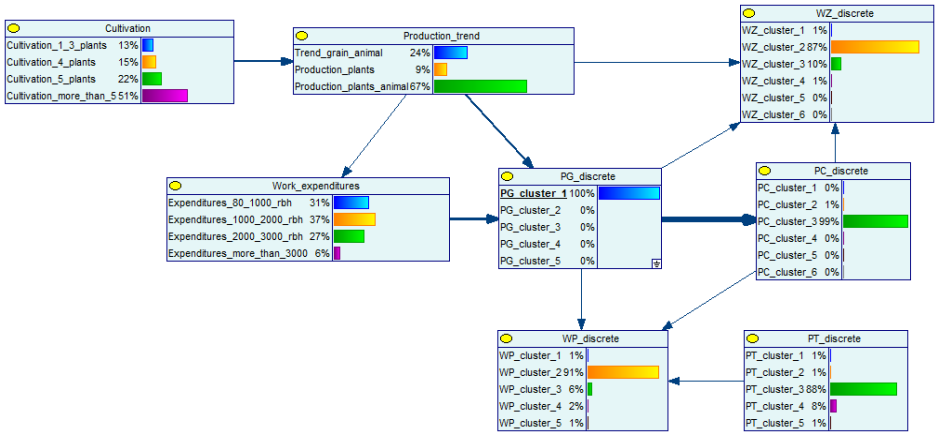

Figure 3. Distribution of conditional probabilities for economic and agricultural indicators if the event PG, belonging to the cluster 1 takes place

Also, for the same conditional event (PG_discrete = 1) chances of the event that WP value will belong to cluster 2 increases (to 91%). Change of the event in PG_discrete node (i.e. belonging of PG value to any of its natural clusters) is related to the events in the preceding nodes (Cultivation, Production_trend, Work_expenditures). It results from the system of relations of these nodes. To make an event happen in the node PG_discrete with the unit probability (100%) then the system of probability in prior nodes must be different than the calculated initial probabilities. On the other hand, for the change of event in PT_discrete node the system of probabilities in PT_discrete node does not change because in the modelled Bayesian Network this node is not related to PG_discrete node.

Alternative Bayesian networks

In order to verify the force of observed relations for natural clusters, the values of indicators used in Bayesian Networks it was decided to build similar networks but based on discrete values from another variable range. This division was formed by the use of arbitrarily determined criteria which consist in determination of ranges of the value for each variable (division of the variable value range into several parts) and attributing a discrete value to each range. Characteristics of clustered variables were presented in the following tables (table 1-5) and in figures (networks 4-5). Tables present basic statistics of the central trend for values from the arbitrarily determined ranges of constant variables i.e. global production (PG_divided), clean production PC (PC_divided), work performance WP (WP_ divided), land efficiency WZ (WZ_ divided) and scientific and technical progress PT (PT divided).

Table 1.
Characteristics of clusters of discretized variable of global production PG

	PG, (PLN thous.)						
Cluster	Number	% from N Total in the table	Minimum	Maximum	Average	Median	Standard deviation
<0;110)	22	7.7 %	0.67	9.53	7.16	8.36	2.66
<10;25)	85	29.8 %	10.13	24.97	17.92	17.77	4.64
<25;50)	104	36.5 %	25.01	49.91	36.09	35.74	7.01
<50;100)	45	15.8 %	50.95	97.12	68.56	65.72	14.46
<100;∞)	29	10.2 %	108.38	269.82	155.53	136.40	45.61
Total	285	100.0 %	0.67	269.82	45.72	32.41	44.23

Table 2.
Characteristics of clusters of discretized variable of clean production PC

PC (PLN thous.)							
Cluster	Number	% from N Total	Minimum	Maximum	Average	Median	Standard deviation
(-∞;1)	18	6.3 %	-281.300	0.940	-20.079	-0.950	66.782
<1;10)	77	27.0 %	1.020	9.770	5.863	6.020	2.535
<10;25)	102	35.8 %	10.010	24.840	16.984	16.540	4.352
<25;50)	50	17.5 %	25.130	47.780	33.716	32.065	6.624
<50;100)	22	7.7 %	51.700	89.030	71.815	75.215	12.365
<100;∞)	16	5.6 %	102.000	201.340	126.201	119.310	27.212
Total	285	100.0 %	-281.300	201.340	24.938	16.000	36.670

Table 3.
Characteristics of clusters of discretized variable of land efficiency WZ

WZ, (PLN thous.·ha^{-1})							
Cluster	Number	% from N Total in table	Minimum	Maximum	Average	Median	Standard deviation
(-) (0)	14	4.9 %	-5.310	-0.010	-0.999	-0.300	1.649
<0;1)	37	13.0 %	0.010	0.970	0.522	0.530	0.257
<1;5)	176	61.8 %	1.000	4.940	2.654	2.615	1.077
<5;10)	42	14.7 %	5.010	9.700	6.908	6.785	1.295
<10;∞)	16	5.8 %	10.150	48.390	22.935	20.265	13.587
Total	285	100.0 %	-5.310	48.390	3.964	2.660	6.018

Table 4.
Characteristics of clusters of discretized variable of work performance WP

WP, (PLN thous.·mhr^{-1})							
Cluster	Number	% from N Total	Minimum	Maximum	Average	Median	Standard deviation
(-∞;0)	12	4.2 %	-0.077	-0.000	-0.010	-0.002	0.022
<0;0.005)	48	16.8 %	0.000	0.004	,0026	0.002	0.001
<0.005;0.01)	75	20.3 %	0.005	0.009	0.007	0.007	0.001
<0.01;0.05)	137	48.1 %	0.010	0.049	0.020	0.016	0.009
<0.05;∞)	13	4.0 %	0.053	0.229	0.087	0.078	0.047
Total	285	100.0%	-0.077	0.229	0.015	0.010	0.022

Table 5.
Characteristics of clusters of digitized variable of scientific and technical progress variable PT

PT, (PLN thous.·mhr^{-1})							
Cluster	Number	% z N Total	Minimum	Maximum	Average	Median	Standard deviation
(-∞;0)	84	29.5 %	-0.686	-0.000	-0.034	-0.007	0.103
<0;0.001)	56	19.6 %	0.000	0.001	0.000	0.000	0.000
<0.001;0.005)	41	14.4 %	0.001	0.005	0.003	0.003	0.001
<0.005;0.01)	29	10.2 %	0.005	0.010	0.007	0.007	0.001
<0.01;0.05)	50	17.5 %	0.010	0.047	0.024	0.022	0.010
<0.05;∞)	25	8.8%	0.057	0.237	0.102	0.091	0.049
Total	285	100.0%	-0.686	0.237	0.004	0.001	0.068

Ranges of values of variables which were formed as a result of their arbitrary division which describes a cluster with its borders (clusters) are more varied on account of number in comparison to clusters separated analytically (with a method of two step clustering or hierarchical method) which also influences another distribution of probabilities in nodes.

Based on the knowledge and experience from previous economic analyses, the network was corrected with other connections between nodes than in the two-step cluster analysis. Connections which result from the relations of economic and agricultural indicators were introduced to the network. They are expressed with formulas, which include these relations (Grotkiewicz et al., 2013). Finally, for further analyses the network takes the form, which was presented in the following diagram (figure 4) and concerns constant variables, i.e. PG (PG_divided), PC (PC_divided), WP (WP_divided), WZ (WZ_divided) and PT (PT_divided) and quality variables.

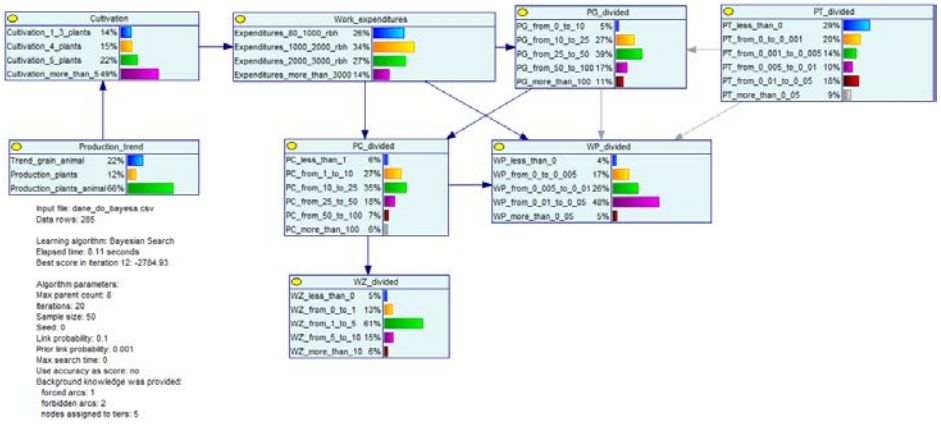

Figure 4. Distribution of conditional probabilities for agri-economic indicators

For the above network, analysis of conditional probabilities was carried out in the secondary nodes for the obtained network. Some of the results were presented in the further part.

For example, when PG value belongs to the second value range (from 10 to 2 thousand PLN) i.e. when the value of PG_divided variable belongs to the cluster <10;25) (probability of this event equals 1), then distribution of conditional probabilities of occurrence of particular values in the remaining nodes is as on the following diagram (network 5).

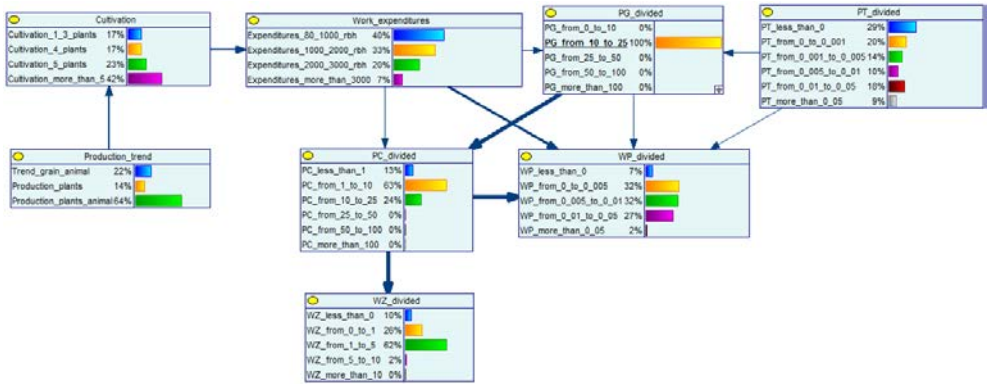

Figure 5. Distribution of conditional probabilities for economic and agricultural indicators if the event PG, belonging to the cluster (of the value range) <10;25) takes place

Occurrence of the above described event results in the probability change in some node networks which depend on the node PG_divided. For example, probability of the event occurrence that PC_divided value will belong to the value range of <1; 10), increases (from 27% to 63%). Change of the event in PG_divided node (i.e. belonging of PG value to any of separated value ranges) is related to the events in the preceding nodes (Cultivation, Production_trend, Work_expenditures). It results from the system of relations of these nodes.

Conclusion

Based on the research results, networks, which reflect the links between the analyzed variables, were built. Network nodes are variables (features with discrete values) while node links (arcs or vectors) present the relations between features and their direction). Thickness of links between nodes on the graphs, which present the Bayesian networks, symbolizes strength between variables. Using the possibility of checking conditional probabilities which take place if a specific event occurs (e.g. a value belonging to a specific cluster of a preceding node in a network), numerous various analyses were carried out in order o obtain a table of conditional probabilities for the values of particular network nodes. Based on the previous analyses (Grotkiewicz et al., 2016) networks which reflect the model of an individual farm were built including economic and agricultural indicators which are important from the point of view of intensity and competitiveness. Although they refer to the entire national economy they particularly concern an agricultural sector, where indica-

tors, investigated in the microscale, indicate the distance between southern Poland and highly developed farms from e.g. Opolskie or Zachodniopomorskie region.

The above research concerned only individual farms and included majority of small farms. Thus, the obtained results cannot be transferred to bigger facilities including agricultural farms. Therefore, the following stage of presenting methods with the use of Bayesian networks in considerably bigger facilities including agricultural enterprises is still a plan and results of analyses will serve for determination of models of future agricultural enterprises.

References

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

ARiMR. (2016) (on-line). Dostępny w internecie: http://www.arimr.gov.pl/dla-beneficjenta/srednia-powierzchnia-gospodarstwa.html

Bartnik, G., Kusz, A. (2005). Sieci probalistyczne jako system reprezentacji wiedzy diagnostycznej. Inżynieria Systemów Bioagrotechnicznych, Politechnika Warszawska. *Zeszyt 5*(14), 5-12.

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

Gębka M., Filipiak T. (2006). *Podstawy ekonomii i organizacji gospodarstw rolnych*. SGGW, Warszawa, ISBN 83-7244-756-X.

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

Grotkiewicz, K., Peszek, A., Kowalczyk, Z. (2016). Verification of economic and agricultural indicators with the use of statistical methods on example of individual farms. *Agricultural Engineering, 3*(159), 149-156.

Grzegorek, J. (2012). Miejsce Polski w Europie i Świecie według wybranych danych statystycznych. *Polska Akademia Nauk, Tom III*, 286-297.

Jongsawat, N., Tungkasthan, A., Premchaiswadi, W. (2010). Dynamic Data Feed to Bayesian *Net*work Model and SMILE Web Application. *Bayesian Network*, ISBN 978-953-307-124-4.

Kusz, A., Marciniak, A., Skwarcz, J.(2015). Implementation of computation process in a bayesian network on the example of unit operating costs determination. *Eksploatacja i Niezawodność – Maintenance and Reliability, 17*(2), 266-272.

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

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

Oniśko, A., Druzdzel, M.J., Wasyluk, H. (2001). Learning Bayesian network parameters from small data sets: application of Noisy-OR gates. *Interniatonal Journal of Approximate Reasoning, 27*, 165-182.

Sujak, A., Kusz, A., Rymarz, M., Kitowski, I. (2016). Environmental Bioindication Studies by Bayesian Network with Use of Grey Heron as Model Species. *Environmental Modeling & Assessment, Open Access.* DOI 10.1007/s10666-016-9524-4

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

WYKORZYSTANIE SIECI BAYESOWSKICH DO PROGNOZOWANIA PRZYSZŁOŚCIOWEGO MODELU GOSPODARSTWA ROLNEGO

Streszczenie. Poszukując zależności między poziomem intensywności produkcji a wydajnością pracy i ziemi oraz czynnikami je kształtującymi, przeprowadzono analizy porównawcze w skali krajowej na tle 300 gospodarstw indywidualnych z województwa małopolskiego i świętokrzyskiego. Analiza dotyczyła zastosowania algorytmów modelowania bayesowskiego do przewidywania rozwoju różnych wskaźników ekonomiczno-rolniczych decydujących o intensywności i konkurencyjności rolnictwa. Praca stanowi drugi etap badań, który poprzedzony był wcześniejszym przygotowaniem danych do modelowania wykorzystując do tego eksploracyjny przegląd dostępnych danych, oraz technikę TwoStep Cluster Analysis (Grotkiewicz i in., 2016). W oparciu o przeprowadzone analizy zbudowano sieci obrazujące związki pomiędzy analizowanymi zmiennymi oraz sprawdzono prawdopodobieństwa warunkowe.

Słowa kluczowe: sieci bayesowskie, prognozowanie, wskaźniki ekonomiczno-rolnicze, gospodarstwo rolne