THE CLASSIFICATION OF ROMANIAN COUNTIES FROM AGRICULTURAL POINT OF VIEW

Ionela-Cătălina TUDORACHE (ZAMFIR)[1]
[1] PhD. Student, The Bucharest University of Economic Studies, Economic Cybernetics and Statistics Doctoral School, email: tudorachecata@yahoo.com

Abstract
Data mining techniques are used recently in more and more fields. Starting from pattern recognition (text, diseases, voice, images) to different prediction applications. Taking into account 33 variables (productions of different cereals, fruits and vegetables and agricultural areas that are cultivated with cereals, fruits and vegetables) are using data mining techniques such as informational synthesizing techniques (principal components analysis) and unsupervised pattern recognition (cluster analysis), the main goal of this article is to classify the Romanian counties into 3 major agricultural performance classes, and to identify hidden patterns of objects. After describing the variables used and analyzing the dataset, the principal components analysis reduced the dimension of the dataset at 6 principal components, while Ward's method confirmed the correct choice of 3 classes and K-means algorithm classified all 41 counties into 3 major clusters. The conclusions of this study reveals that counties with plains have a high performance in cereals production, while counties with hills and mountains produce more vegetables and fruits. From this point of view, each class has its own performance and characteristics.

Keywords
agricultural performance, principal component analysis, cluster analysis, Romanian counties, classification

Introduction

In the past, many studies about agricultural output performances (and not only about agriculture) were limited at graphically representations of output in terms of efficiencies, productions, areas. Due to the development of Data Mining techniques, many authors started to apply the new methodologies in different domains. In terms of studying the performance of a dataset composed by objects (companies, counties, countries, regions), one of the most adequate method is studying the performance by comparing an object with others. In this way, we can be sure that an object is better or worse than similar objects.

From this point of view, the opportunity of this research comes in relation with the development of Data Mining techniques, as well as the computing means.

By taking into account 33 variables of all 41 counties (even thou several counties were outliers at different variables, there no exclusions made due to the high interest of identifying the performance of each county) from Romania, this study aims to identify both the agricultural output efficiency and the hidden patterns of the performance classes.

The dataset (downloaded from the National Institute of Statistics site in august 2015) represents 2014's year, it is composed by 33 variables (production of different cereals, fruits and vegetables, and agricultural areas that are cultivated with cereals, fruits and vegetables), and takes into account all 41 counties (except Bucharest, that considered to have a low agricultural activity and a very high service sector performance).

At this dataset, in order to identify each county performance, several research techniques were applied. The first technique is the statistical description of dataset, followed by the description of correlations between variables. The variables correlations prove if the variables

were correctly selected and denote the correct use of principal components analysis (that is the second technique applied on dataset) in order to reduce the dataset dimensionality and eliminate informational redundancy. The last methodologies applied are the pattern recognition methods, unsupervised pattern recognition. The hierarchical Ward's method assures the right choice of three classes, while K-means algorithm gives the affiliation of each county to a performance class.
The study is divided into five major sections: section 1 presents the literature review, the main results that are relevant for this research; section 2 presents the database, the variables used and descriptive statistics for the variables taken into account; section 3 is a brief theoretically presentation of the methodologies approached; section 4 describes in detail the results obtained by applying the methodologies from the previous section, results that are both technically and economically presented; the last section is reserved for conclusions of the study and presents the main findings of the research, if the research main goal was achieved and the further research.

1. Literature review

Data Mining techniques have a wide application area, from fields like marketing, psychology, macroeconomics, finance, accounting to fields like agriculture, commerce, medicine. In agriculture there are recent studies that prove the great interest for these methods and techniques. Rotaru, Pop et.all (2012) uses principal components analysis in order to synthesize 9 variables describing the agricultural area cultivated with crops in Romania into 2 factor components that take over 80% of total information. In order to reduce the experiments costs about evaluating the performances of a plant in an environment, Meirelles and Zarate (2015) used methodologies like k-medians algorithm (that is similar to k-means algorithm) and neural computational models on a Brazil database.
On the other hand, a literature review (Behmann et.all, 2015) demonstrates the use of Data Mining techniques in agriculture: both supervised learning (neuronal networks) and unsupervised learning (k-means and self-organizing maps) are used for precision agriculture, with applications on detecting plant diseases or detecting the weed in a crop field. In the same year, Blasch et.all (2015) propose a model to generate soil information based on remote sensing data, using principal components analysis and applying the model on a field from Northeast Germany. Gorgens, Montaghi and Rodriguez (2015) use machine learning models like neuronal networks, random forests, support vector regression and a regression model to compare their accuracies in the problem of predicting the growth of forests plantations of Eucalypt. The results of the study demonstrate the applicability of these methods in agriculture.
Other significant interdisciplinary studies based on agricultural problems solved with data mining techniques are: reducing crop losses and increasing crop efficiency using the combination of spatial data, temperature and rainfall and applying k-means algorithm (Rajesh, 2011); using clustering techniques in order to determine what happened with the "agriculture land vanished in the past seven years" (Magala, Hemalatha, 2011); identifying crop pattern using clustering and classification methods of Data Mining (Fathima, Geetha, 2014); revealing the relations between soil and farmers activity using classification and clustering methods in a support system (Jeysenthil, Manikandan, Murali, 2014); the prediction of crop yield in a specific region of India, using techniques like multiple linear regression or density based clustering (Ramesh, Vardhan, 2015).

2. Dataset, variables and descriptive statistics

The dataset used was downloaded from the National Institute of Statistics (INSSE Tempo) and has 41 counties (except Bucharest) and 33 variables that indicate the agricultural output of Romania in 2014 and the area cultivated with different cereals, fruits and vegetables in 2014. This study refers the classification of Romanian counties only from economically point of view, taking into account the variables provided by INS. It does not take into account other variables that may influence the agricultural performance of a county, such as the weather, the soil humidity, natural environment or the landforms.

Table 2 Variables used in the model

Indicator's code	Explanation	Indicator's code	Explanation	Indicator's code	Explanation
p_prune	Plums	p_porumb	Corn	p_ardei	Pepper
p_mere	Apples	p_mazare	Peas	s_agrigola	Agricol
p_pere	Pears	p_fasole	Beans	s_grau	Wheat
p_piersici	Peaches	p_floare	Sunflower	s_orz	Barley
p_nectarine	Nectarines	p_rapita	Rapeseed	s_porumb	Corn
p_ciresi_visini	Cherry and sour cherry	p_cartofi	Potatoes	s_mazare	Peas
p_caise_zarzari	Apricots	p_tomate	Tomatoes	s_floare	Sunflower
p_nuci	Nuts	p_vinete	Eggplant	s_rapita	Rapeseed
p_capsuni	Strawberries	p_ceapa	Onion	s_cartofi	Potatoes
p_grau	Wheat	p_usturoi	Garlic	s_legume	Vegetables
p_orz	Barley	p_varza	Cabbage	s_livezi	Orchards

Source: Author's computation

The variables used in this research are presented in the table from above. There are variables that indicate production output (variables that have p_ in front of the variable code), and variables that show the areas cultivated (variables that have a s_ in front of the variable code). The variables that indicate production are measured in tons, while variables for areas are measured in hectares.

Table 3 Descriptives statistics for variables

Indicator	Mean	Standard Error	Standard Deviation	Kurtosis
p_prune	12079.90	2081.48	13327.99	3.98
p_mere	12516.24	2036.16	13037.79	0.72
p_pere	1494.93	171.67	1099.21	1.15
p_piersici	579.22	202.22	1294.85	28.15
p_nectarine	23.15	9.98	63.93	36.53
p_ciresi_visini	2019.29	269.41	1725.08	17.22
p_caise_zarzari	1063.05	159.62	1022.07	2.41
p_nuci	768.63	68.30	437.36	-0.29

Indicator	Mean	Standard Error	Standard Deviation	Kurtosis
p_capsuni	535.00	331.72	2124.02	38.91
p_grau	184994.00	28595.20	183098.65	-0.17
p_orz	26476.88	6444.40	41264.29	5.24
p_porumb	292402.95	29465.02	188668.18	-0.09
p_mazare	1244.32	379.38	2429.24	8.99
p_fasole	481.66	108.81	696.73	12.96
p_floare	53397.78	9318.51	59667.57	-0.31
p_rapita	25832.22	5575.21	35698.76	5.71
p_cartofi	85837.29	14756.61	94488.40	3.14
p_tomate	17223.41	2393.23	15324.12	6.98
p_vinete	3111.39	434.50	2782.17	1.76
p_ceapa	9438.76	903.18	5783.17	-0.88
p_usturoi	1531.05	166.15	1063.87	-0.01
p_varza	27393.46	4389.75	28108.08	23.33
p_ardei	5574.76	750.92	4808.23	1.87
s_agrigola	356756.59	17401.97	111426.97	1.32
s_grau	51532.93	7983.54	51119.62	0.50
s_orz	7413.76	1715.77	10986.29	3.50
s_porumb	61287.90	5738.36	36743.45	1.13
s_mazare	667.41	210.83	1349.95	13.52
s_floare	24415.12	4229.34	27081.01	0.06
s_rapita	9919.63	2046.49	13103.90	4.54
s_cartofi	4842.44	670.51	4293.39	3.98
s_legume	5839.98	519.97	3329.41	0.10
s_livezi	3434.32	619.93	3969.49	7.73

Indicator	Skewness	Range	Min	Max
p_prune	1.89	57196	315	57511
p_mere	1.37	44541	208	44749
p_pere	1.22	4587	62	4649
p_piersici	5.02	7992	0	7992
p_nectarine	5.90	412	0	412
p_ciresi_visini	3.44	10758	116	10874
p_caise_zarzari	1.54	4348	6	4354
p_nuci	0.10	1757	29	1786

Indicator	Skewness	Range	Min	Max
p_capsuni	6.17	13644	0	13644
p_grau	1.09	576626	8206	584832
p_orz	2.25	186808	149	186957
p_porumb	0.62	790446	11103	801549
p_mazare	2.96	11418	0	11418
p_fasole	3.26	3790	0	3790
p_floare	0.96	196099	0	196099
p_rapita	2.30	156822	0	156822
p_cartofi	1.78	411128	1384	412512
p_tomate	2.22	80727	501	81228
p_vinete	1.53	11003	0	11003
p_ceapa	0.45	19720	2022	21742
p_usturoi	0.78	4334	12	4346
p_varza	4.34	177337	4393	181730
p_ardei	1.40	20098	31	20129
s_agrigola	0.66	589846	101453	691299
s_grau	1.24	175896	2483	178379
s_orz	2.00	45617	56	45673
s_porumb	0.69	176648	2416	179064
s_mazare	3.48	7083	0	7083
s_floare	1.06	89884	0	89884
s_rapita	2.09	56962	0	56962
s_cartofi	1.63	21169	162	21331
s_legume	0.80	13055	717	13772
s_livezi	2.39	20542	52	20594

Source: Excel Output

Table 2 from above shows the descriptive statistics for all 33 variables. The average production of corn is about 292403 tons in 2014, with a minimum value of 11103 tons in Harghita and a maximum value of 801549 tons in Timis. On the other hand, the average area cultivated with corn is 61288 hectares (that means an average efficiency of 4.77 tons/hectare), with a maximum value of 179064 hectares in Timis and a minimum value of 2416 hectares in Harghita.

The range from table 2 is the difference between the maximum value and the minimum value and shows that, for many variables, there is a big difference between minim and maxim, that means that there is a high variability in data. A positive skewness value (all variables considered) shows that the right tail is longer for the probability distribution, while a kurtosis lower than 3 represents a platikurtic distribution, flatter than the normal distribution, with values spread on a big interval around the mean value.

3. Methodologies approached

3.1. Principal components analysis

The principal components analysis is one of the most important techniques if the variables taken into account are too numerous. The main goals of this analysis are:

- reduce variables dimensionality: if it is considered n variables (in this case 33 variables), after applying principal components analysis, there will be left k components (6 in this case), where k is smaller than n. These components are new variables that synthesize about 80% of total information.
- eliminate the informational redundancy: due to the fact that each model has variables that are correlated, the informational redundancy depends on the correlation level between variables. A high correlation (about over 60%) represents a high redundancy, while a low correlation represents low redundancy. Considering the principal components model and how the components are built, the correlation between components is zero, that gives no informational redundancy.

The principal components can be defined as a linear combination between the original variables and the eigenvectors coordinates (eigenvectors that correspond to the covariance matrix of the variables), and can be determined like (Ruxanda, 2009):

$w_i = \alpha_1^{(i)} * x_1 + \alpha_2^{(i)} * x_2 + ... + \alpha_n^{(i)} * x_n$, i=1,2,...,n where: w_1 is the first principal component. There are n principal components that take all information from the original variables, but only k of them are kept in the further model.

There are several properties of new variables that are represented by principal components, and several criterions to choose the proper number of components (k). The most important properties are:

- all n components can substitute all n variables because the components preserve the total variance from variables
- if variables are correlated one with each other, the principal components model's construction assures that the components are not correlated, and the first principal component takes the most part of total information from variables.

The most used criterions are: the coverage percent criterion sais to keep a k number of components that take about 75-80% of total information; the Kaiser criterion suggests to keep a number of k components corresponding to the number of eigenvalues higher or equal with 1 (this criterion is used only on standardized dataset).

3.2. Pattern recognition techniques

About pattern recognition techniques, there are two major categories: unsupervised pattern recognition and supervised pattern recognition. The first category includes techniques that classify a set of objects that are not classified. At this point it is unknown the number of classes, or the affiliation of each object to a specific class. There are hierarchically methods (when then number of classes is unknown), and algorithmically methods (when the number of classes is known). The supervised pattern recognition is based on unsupervised, because it uses a learning set, composed by a big number of objects, and predicts the affiliation of a new object to a specific class (with a certain error degree).

In this research, the Ward's method is used to confirm (or infirm) the number of three performance classes chosen, and K-means algorithm to identify the objects of each class, because is an algorithmically method that has a higher accuracy rate than an hierarchically method (due to the fact that it runs until a stop condition is fulfilled).

4. Research results

The results of the research is one of the most important section. Before showing the results, it is important to mention the transformations of the dataset in order to apply the principal components analysis. These transformations consists of standardizing the dataset (using z-score method). In this way, the new dataset, with standardized variables has the mean 0 and the standard deviation equal with 1 for each variable, and the correlation matrix equal with the covariance matrix. This step was important to be made, because there are different measurement units (tons versus hectares).

Table 4 A fragment of correlation matrix

_NAME	**p_prune**	**p_mere**	**p_pere**	**p_piersici**	**p_nectarine**	**p_ciresi _visini**	**p_caise _zarzari**
p_prune	**1.000**	0.472	0.697	0.059	0.029	0.250	0.204
p_mere	0.472	**1.000**	0.718	0.028	0.045	0.212	-0.143
p_pere	0.697	0.718	**1.000**	0.104	0.096	0.409	0.183
p_piersici	0.059	0.028	0.104	**1.000**	0.931	-0.046	0.503
p_nectarine	0.029	0.045	0.096	0.931	**1.000**	-0.031	0.428
p_ciresi _visini	0.250	0.212	0.409	-0.046	-0.031	**1.000**	0.533
p_caise _zarzari	0.204	-0.143	0.183	0.503	0.428	0.533	**1.000**

Source: SAS Output

Table 3 from above is a small fragment of the correlation matrix, taking into account 7 variables from 33. It is important to identify the fact that in most of the models (no matter the domain), it is impossible to have a zero correlation between any two variables. From this point of view, the information redundancy is as high as the correlation value is. In this model, it is a high correlation between the production of apples and the production of pears (about 71.8%), or between the production of sunflower and the production of wheat (89.1%). In this respect, the principal components analysis is a must.

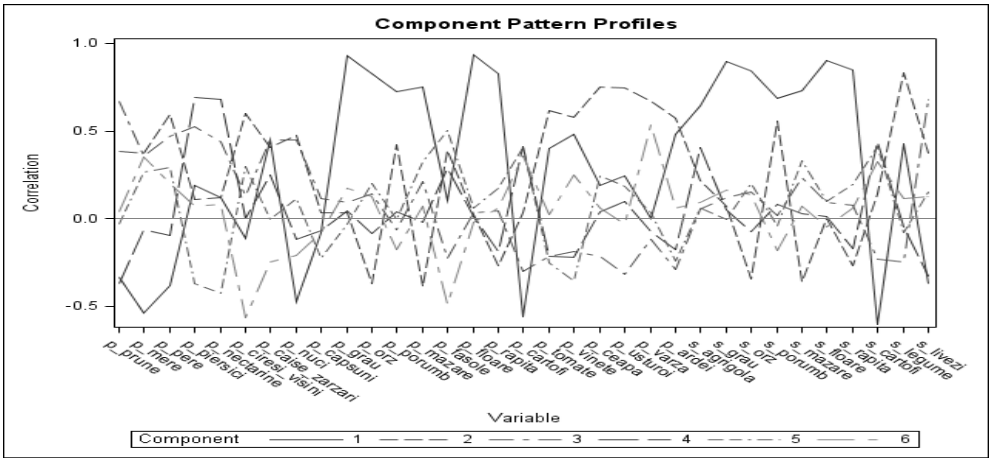
Figure 1 from above shows the first 15 eigenvalues of the covariance matrix (that is equal with the correlation matrix). According to Kaiser criterion (used because the dataset was standardized), 8 principal components can be retain in further analysis. According to coverage percentage criterion, the first 6 of the components, that synthesize almost 80% of total information may be retained. Considering that component 7 and 8 bring a less than 4% of information to the rest of 6 components, only the first 6 are considered for cluster analysis.

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	11.4344951	5.1735205	0.3465	0.3465
2	6.2609746	3.3167101	0.1897	0.5362
3	2.9442645	0.8111277	0.0892	0.6254
4	2.1331368	0.1594790	0.0646	0.6901
5	1.9736577	0.3257962	0.0598	0.7499
6	1.6478615	0.4930672	0.0499	0.7998
7	1.1547943	0.0764963	0.0350	0.8348
8	1.0782979	0.2383601	0.0327	0.8675
9	0.8399378	0.2797553	0.0255	0.8930
10	0.5601825	0.0204633	0.0170	0.9099
11	0.5397192	0.0981807	0.0164	0.9263
12	0.4415385	0.0371285	0.0134	0.9397
13	0.4044100	0.0736063	0.0123	0.9519
14	0.3308037	0.0504561	0.0100	0.9619
15	0.2803476	0.0604257	0.0085	0.9704

Source: SAS Output

Fig. 1 Eigenvalues of the covariance matrix



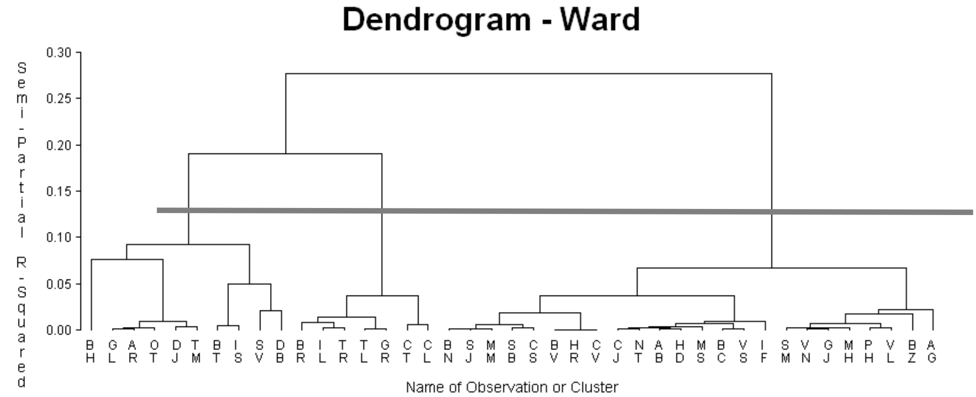
Source: SAS Output

Fig. 2 Component pattern profiles

Figure 2 from above is the graphical representation of the correlations between all 33 variables and the first 6 principal components. In this graph, the new variables (principal components) can be identified and named according the information that is contained from the original variables, like:

- W_1 takes more information (than the rest of the principal components) from: p_grau, p_orz, p_porumb, p_floare, p_rapita, p_mazare, s_grau, s_orz, s_floare, s_rapita, s_porumb, s_mazare, s_agricola and may be named ***cereals*** component;
- W_2 takes more information from: p_pere, p_ciresi_visini, p_tomate, p_vinete, p_ceapa, p_usturoi, p_varza, p_ardei, s_legume, and can be named ***most vegetables, pears and cherry*** component;
- W_3 takes more information from: p_prune, p_caise_zarzari, p_nuci, p_capsuni, s_livezi and can be named as ***most fruits*** component;
- W_4 is the ***peaches and nectarines*** component, because it takes more information from p_piersici, p_nectarine;

- W_5 is ***beans and potatoes*** component because it takes more information from s_cartofi, p_cartofi, p_fasole;
- W_6 is the ***apples*** component, because the correlation between W_6 and p_mere is 27.22% (the higher correlation between p_mere and any other component from the first 6 considered).

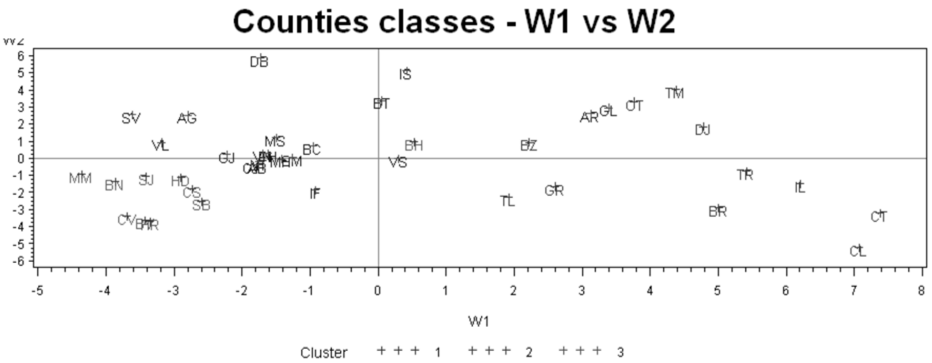


Source: SAS Output

Fig. 3 Ward's hierarchically method's dendrogram

Figure 3 is the Ward's dendrogram, or the classification tree that shows one of the many possibilities of grouping the 41 counties, using the 6 principal components from above and fulfilling the general criterion of classification (the variability within classes is low and the variability between classes is high). The red horizontal line proves that the number of 3 classes is the one that fulfills better the general criterion of classification, due to the fact that the difference from 3 classes to 4 or more classes is higher (difference measured by distances in the classification tree). Having the right number of classes, the K-means algorithm was applied on principal components.

The graphically representation of counties is in figure 4 from above. Counties with a high value for the first principal component (the class colored in red) have high value for cereals, while most counties with blue color have a high production of most vegetables, pears and cherry (component 2).



Source: SAS Output

Fig. 4 Graphically representation of classes in W_1*W_2 plan

Table 5 Classes average values for original variables

Class	Counties	p_prune	p_mere	p_pere	p_piersici	p_nectarine
1	10	10431.20	16564.90	1168.40	980.30	46.10
2	11	7458.91	3238.91	886.91	653.91	20.55
3	20	15445.80	15594.45	1992.60	337.60	13.10
Class	**Counties**	**p_ciresi_visini**	**p_caise_zarzari**	**p_nuci**	**p_capsuni**	**p_grau**
1	10	1289.10	531.20	744.80	66.00	51971.50
2	11	1348.00	1605.82	431.36	109.27	449101.55
3	20	2753.60	1030.45	966.05	1003.65	106246.10
Class	**Counties**	**p_orz**	**p_porumb**	**p_mazare**	**p_fasole**	**p_floare**
1	10	3945.10	99167.70	102.10	112.10	6756.80
2	11	79498.82	437708.00	3860.45	402.45	136269.73
3	20	8580.70	309102.80	376.55	710.00	31138.70
Class	**Counties**	**p_rapita**	**p_cartofi**	**p_tomate**	**p_vinete**	**p_ceapa**
1	10	2297.00	157736.10	4118.40	1005.70	4081.80
2	11	67213.09	20434.00	20964.73	4442.91	9285.36
3	20	14840.35	85859.70	21718.20	3431.90	12201.60
Class	**Counties**	**p_usturoi**	**p_varza**	**p_ardei**	**s_agrigola**	**s_grau**
1	10	655.90	12325.40	1590.90	317350.20	14658.40
2	11	1697.27	24341.27	7971.91	454545.00	123994.55
3	20	1877.20	36606.20	6248.25	322676.15	30116.30
Class	**Counties**	**s_orz**	**s_porumb**	**s_mazare**	**s_floare**	**s_rapita**
1	10	1138.60	23364.40	60.90	3574.40	1016.30
2	11	21879.64	87665.36	2054.82	61147.18	25790.27
3	20	2595.10	65742.05	207.60	14632.85	5642.45
Class	**Counties**	**s_cartofi**	**s_legume**	**s_livezi**		
1	10	8002.10	2445.80	3467.60		
2	11	1535.73	7027.82	1683.91		
3	20	5081.30	6883.75	4380.40		

Source: Excel computation

The table from above shows the number of counties in each performance class and the average values for the original variables for each class. At this stage, it is impossible to say that a class has a higher performance than another, in general. But, taking into account several variables, it is possible to say that:

- **class 1** (composed by counties: Bihor, Bistrita-Nasaud, Brasov, Caras-Severin, Covasna, Hunedoara, Harghita, Maramures, Sibiu, Salaj) have high average values for variables like: p_mere, p_piersici, p_nectarine, p_cartofi, s_cartofi and the lowest values for the rest of the variables and can be considered a class with high performances in producing fruits like ***apples, peaches and nectarines***, vegetables as ***potatoes***, and low performance in producing the rest of the fruits, most of the vegetables and cereals.

- **class 2** (composed by counties: Braila, Buzau, Calarasi, Constanta, Dolj, Giurgiu, Ialomita, Olt, Tulcea, Timis, Teleorman) have high average values for variables like: p_caise_zarzari, p_grau, p_orz, p_porumb, p_mazare, p_floare, p_rapita and can be considered a class with high performances in ***cereals*** production and low production performances in fruits and vegetables.
- **class 3** (the rest of the counties) have high performances in producing some of the fruits (***pears***, ***strawberries***, ***plums***) and vegetables (***tomatoes, onion, garlic, cabbage***), and low performances in producing cereals.

Source: Author's computations, the blank map was taken from: http://cmvro.cmvro.ro/cmvro/

Fig. 1 Romania divided in 3 clusters

The figure from above is the geographically representation of all classes. Class 1 is colored in green, and show that, in a mountain region it is expected to have more orchards than the rest of the regions (and therefore, the production of apples, peaches and nectarines is very high). Class 2 is colored in red and is represented by the Southern and South-East region of Romania (excepts for Timis from West). These regions are known by wide plains, so the production of cereals is very high. Class 3 is colored in blue and has counties with hills, plains (in East) and low mountains, so the orchards and vegetables are produced here, but the production of cereals is lower than the South region.

5. Conclusions

Considering 33 agricultural and economical variables, synthesized in 6 principal components, the study aim was to identify both the agricultural output efficiency and the hidden patterns of the performance classes. Three major clusters were identified and each class has its own performance. It was demonstrated that the landform "decides" what type of goods are produced in a county. Each county produces almost all types of fruits, vegetables and cereals considered, but performs only in one specific type of goods. From this point of view, it is impossible to say that a cluster has an overall higher performance than another cluster.

Finally, even thou the weather conditions or the soil humidity were not included in this analysis, the results are compatible with the reality (in my opinion, the weather and humidity variables will not change considerably the results of this study, it will just identify more cluster's patterns and will provide a full description of each class). Moreover, the study demonstrates that using Data Mining techniques in an interdisciplinary approach, the conclusions are validated with the reality, that makes that the applied techniques to be accurate.
The propose for further research is either to extend this study taking into account other variables (from a different field), or to extend the agricultural study at European countries or regions.

Acknowledgement
This paper was co-financed from the European Social Fund, through the Sectorial Operational Programme Human Resources Development 2007-2013, contract POSDRU/159/1.5/S/155463 "Supporting excellence in scientific interdisciplinary doctoral research in the economic, medical and social fields", coordinator The Bucharest University of Economic Studies.

References
1. Behmann, J., Mahlein, A-K., Rumpf, T., Romer, C., Plumer, L. 2015. A review of advanced machine learning methods for the detection of biotic stress in precision crop protection, *Precision Agriculture*, 16(3):239-260.
2. Blasch, G., Spengler, D., Hohmann, C., Neumann, C., Itzerott, S., Kaufmann, H. 2015. Multitemporal soil pattern analysis with multispectral remote sensing data at the field-scale, *Computers and Electronics in Agriculture*, 113:1-13.
3. Fathima, G.N., Geetha, R. 2014. Agriculture Crop Pattern Using Data Mining Techniques, *International Journal of Advanced Research in Computer Science and Software Engineering*, 4(5):781-786.
4. Gorgens, E.B., Montaghi, A., Rodriguez, L.C.E., 2015. A performance comparison of machine learning methods to estimate the fast-growing forest plantation yield based on laser scanning metrics, *Computers and Electronics in Agriculture*, 116:221-227.
5. Jeysenthil, KMS., Manikandan, T., Murali, E. 2014. Third Generation Agricultural Support System Development Using Data Mining, *International Journal of Innovative Research in Science, Engineering and Technology*, 3(3):9923-9930.
6. Megala, S., Hemalatha, M. 2011. A Novel Datamining Approach to Determine the Vanished Agricultural Land in Tamilnadu, *International Journal of Computer Applications*, 23(3):23-28.
7. Meirelles, W.C.L, Zarate, L.E. 2015. Data mining in the reduction of the number of places of experiments for plant cultivates, *Computers and Electronics in Agriculture*, 113, 136-147.
8. Rajesh, D. 2011. Application of Spatial Data Mining for Agriculture, *International Journal of Computer Applications*, 15(2):7-9.
9. Ramesh, D., Vardhan, B.V. 2015. Analysis of Crop Yield Prediction using Data Mining Techniques, *International Journal of Research in Engineering and Technology*, 4(1):470-473.
10. Rotaru, A.S, Pop, I.D., Vatca, A., Cioban, A. 2012. Usefulness of Principal Components Analysis in Agriculture, *Bulletin UASVM Horticulture*, 69 (2): 504:509.
11. Ruxanda, G. 2009. *Analiza multidimensională a datelor*, Academia de Studii Economice, Școala Doctorală, București.
12. http://cmvro.cmvro.ro/cmvro/ - accesed on 30.08.2015 - for Romanian counties blank map.
13. http://statistici.insse.ro/shop/?page=tempo1&lang=ro - accessed on 17.08.2015