HNU HOCHSCHULE NEU-ULM UNIVERSITY

HNU Working Paper

Nr. 17

Marina Fedorova und Michael Grabinski

Anwendung der Clusteranalyse bei der Erforschung des Landwirtschaftsmarkts in Regionen

Using of cluster analysis in regional agricultural market researching

08/2011

Marina Fedorova, Doktorandin, Marische Staatliche Universität, Fachgebiet Volks- und Agrarwirtschaft, Swerdlow-Str. 49-218, 424003 Joschkar-Ola, Republik Mari El, Russland fodorovamarina@yandex.ru

Dr. Michael Grabinski, Professor für Business and Management, Hochschule für angewandte Wissenschaften Neu-Ulm – Neu-Ulm University, Wileystr. 1, 89231 Neu-Ulm, michael.grabinski@uni-neu-ulm.de

Abstrakt

Der Landwirtschaftsmarkt entwickelt sich in verschiedenen Bundesländern ungleichmäßig. Darum entsteht die Notwendigkeit, sie nach dem Agrarentwicklungsniveau zu gruppieren, um die führenden und rückständigen Regionen zu definieren. Solche Gruppierung kann man mit Hilfe der Clusteranalyse durchführen. Dabei werden entweder hierarchische oder nichthierarchische Clusteralgorithmen benutzt. In dem Beitrag wurde die Anwendung des Ward-Verfahrens und K-means-Verfahrens dargestellt und das Ergebnis interpretiert. Die Gruppierung in Bundesländer erlaubt eine regionale Landwirtschaftspolitik. Durch stattliche Förderungen können rückständige Regionen stimuliert werden. Subventionen können dadurch (sofern nach EU-Recht erlaubt) optimiert werden.

Freie Schlagwörter: Landwirtschaftsmarkt, Gruppierung, Clusteranalyse, Ward-Verfahren, Dendrogramm, K-means-Verfahren, regionale Politik

Abstract

The agricultural markets in different federal states are developing non-uniform. Therefore it is necessary to classify them according their agricultural resulting in leading and being backward regions. Such classification can be achieved by using cluster analysis. Hierarchical and non-hierarchical cluster algorithms are used. In this article Wards-method and K-means method are used, and their results are interpreted. The classification of federal states allows the government to evaluate regional policy in order to improve the methods of state support and incentives given to leading and backward regions, respectively (as far as allowed by EU regulations).

Keywords: agricultural market, classification, cluster analysis, Wards-method, dendrogram, K-means method, regional policy

JEL-Classification: Q10, Q13, C10

1. Einleitung

Clusteranalyse hilft die Objekte nach ihrer Ähnlichkeit oder Unterschiedlichkeit zu gruppieren. Die durch eine Anzahl von Variablen beschriebenen Objekte sollen innerhalb einer Gruppe möglichst ähnlich bzgl. der Variablen sein. Objekte aus unterschiedlichen Gruppen sollen möglichst verschieden sein. Die Gruppen nennt man auch Cluster, Klassen oder Typen. Die Gruppeneinteilung wird auch als Klassifikation oder Typologie bezeichnet. Clusteranalyse wird sehr oft in verschiedenen wissenschaftlichen Bereichen angewendet: Psychologie, Soziologie, Medizin, Marktforschung. Die Hauptvorteile der Clusteranalyse sind:

- die Möglichkeit der Analyse und Verminderung der großen Datenmengen
- keine Einschränkungen der analysierbaren Daten

2. Auswahl der Daten

Es gibt verschiedene grundlegende Arten von Clusteralgorithmen: Hierarchische Clusteralgorithmen fassen sukzessiv immer mehr Personen/Objekte zu immer größeren Clustern zusammen (agglomerative Verfahren) oder teilen die Personen/Objekte sukzessiv in immer mehr, immer kleinere Cluster auf (divise Verfahren) und nichthierarchische (k-means) Verfahren.

Um die Entwicklung des Agromarktes der Bundesländer einzuschätzen, muss man die Variablen wählen, die die Lage der Landwirtschaft in den Bundesländern charakterisieren und damit die Ähnlichkeit oder Unähnlichkeit der Objekte festgestellen:

x_1 – die Zahl der landwirtschaftlichen Betriebe;

x_2 – landwirtschaftliche Arbeitskräfte;

x_3 – landwirtschaftlich genutzte Fläche, 1000 ha;

x_4 – Getreideernte, 1000 t;

x_5 – Schweinebestand, 1000;

x_6 – Rinderbestand, 1000;

x_7 – Milcherzeugung, 1000 t;

x_8 – Schlachtmenge, 1000 t;

x_9 – Kartoffelernte, 1000 t;

x_{10} – Weinbestand, 1000 hl.

Für die Clusteranalyse wurden die Daten des Jahres 2009 verwendet, die in der Tabelle 1 dargestellt sind.

Tabelle 1. Ausgangsdaten für die Clusteranalyse: Statistisches Jahrbuch 2010 für die Bundesrepublik Deutschland mit „Internationalen Übersichten“, Statistisches Bundesamt

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
Baden-Würtemberg	57 049	226,9	1 432,8	3 969	2 104	1 045	2 198	530	219	2 978
Bayern	121659	318,1	3 210,6	8 225	3 625	3 414	7 561	842,2	1 933	525
Brandenburg	6 704	38	1 327,1	3 057	772	587	1 388	175,7	341	9
Hessen	22 355	69,5	777,8	2 245	718	485	990	80,6	189	1 559
Mecklenburg-Vorpommern	5 432	28,1	1 360,0	4 240	745	568	1 433	92,3	564	14
Niedersachsen	49 917	168,2	2 605,1	7 693	8 168	2 574	5 303	1 769,7	5 507	15
Nordrhein-Westfalen	47 511	144,6	1 499,0	5 663	6 526	1 438	2 769	2 043,7	1 422	127
Rheinland-Pfalz	25 529	105,2	704,8	1 662	269	384	766	126,3	303	6 898
Saarland	1 660	4,2	77,6	146	12	53	87	2,5	6	20
Sachsen	8 313	41,4	914,9	2 805	654	509	1 608	68,3	306	124
Sachsen-Anhalt	4 842	25,9	1 171,6	4 467	1 054	353	1 066	310,6	578	436
Schleswig-Holstein	17 479	50	992,6	2 784	1 557	1 169	2 504	185,8	222	79
Thüringen	4 789	25,9	790,7	2 759	745	350	953	181	92	7
Berlin, Bremen,Hamburg	1 275	5,3	24,7	0	1	18	32	70,1	0	147

3. Methodik der Clusteranalyse

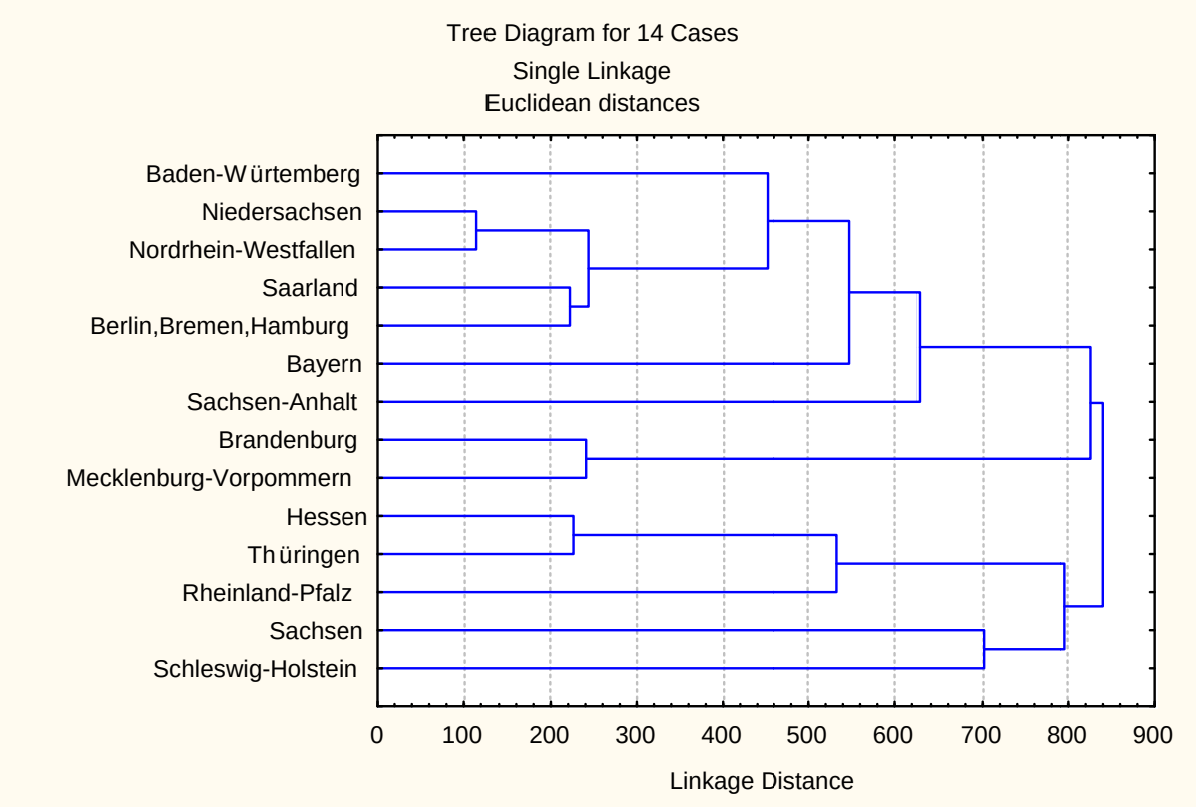
3.1 Ward-Verfahren

Als Methode der Clusteranalyse wird zuerst das Ward-Verfahren gewählt. Hierbei werden diejenigen Cluster fusioniert, deren Vereinigung die Varianz innerhalb der Cluster (Fehlerquadratsumme) möglichst wenig ansteigen lässt. Als Distanzmaß wird entsprechend die quadrierte euklidische Distanz zugrunde gelegt, das Verfahren ist insofern nur für intervallskalierte Merkmalsvariablen sinnvoll. Das Ward-Verfahren hilft statistisch homogene Cluster festzustellen. Die Analyse wurde mit Hilfe des statistischen Programms „STATISTICA 6.0“ durchgeführt.

Der Ablauf der Clusterbildung von der ersten bis zur letzten Stufe wird in einem Dendrogramm (Baumdiagramm) dargestellt (Abbildung 1). Das Dendrogramm stellt

jedoch nicht nur dar, welche Clusterbildung auf den einzelnen Stufen vorgenommen wird, sondern es zeigt zudem, wie groß die Distanz (also die Unähnlichkeit) zwischen den jeweils zusammengefassten Clustern ist. Das Dendrogramm ist von links nach rechts zu lesen und beschreibt in dieser Richtung die einzelnen Stufen der Clusterbildung. Jede Zeile des Diagramms repräsentiert ein einzelnes Objekt (eines der 16 berücksichtigten Bundesländer). Die Spalten zeigen, welche Bundesländer zu einem Cluster verbunden werden. Auf erster Stufe wurden Niedersachsen und Nordrhein-Westfalen, Saarland und Berlin, Bremen, Hamburg, Brandenburg und Mecklenburg-Vorpommern, und auch Hessen und Thüringen, Sachsen und Schleswig-Holstein zu einem Cluster zusammengefasst.

Abbildung 1. Dendrogramm der Clusterbildung für 16 Bundesländer



Auf der weiteren Stufe des Agglomerationsprozesses erfolgte die Bildung der größeren Clusters. Die Größe der Distanzen kann an der Skala „Linkage Distance“ eingeschätzt werden. Auf dritter Stufe der Clusteranalyse werden deutlich zwei Cluster festgestellt. Zum ersten gehören Baden Württemberg, Niedersachsen, Nordrhein-Westfalen, Saarland, Berlin, Bremen, Hamburg (gruppiert), Bayern, Sachsen-Anhalt, Brandenburg und Mecklenburg-Vorpommern. Die letzten zwei Bundesländer bilden getrennte

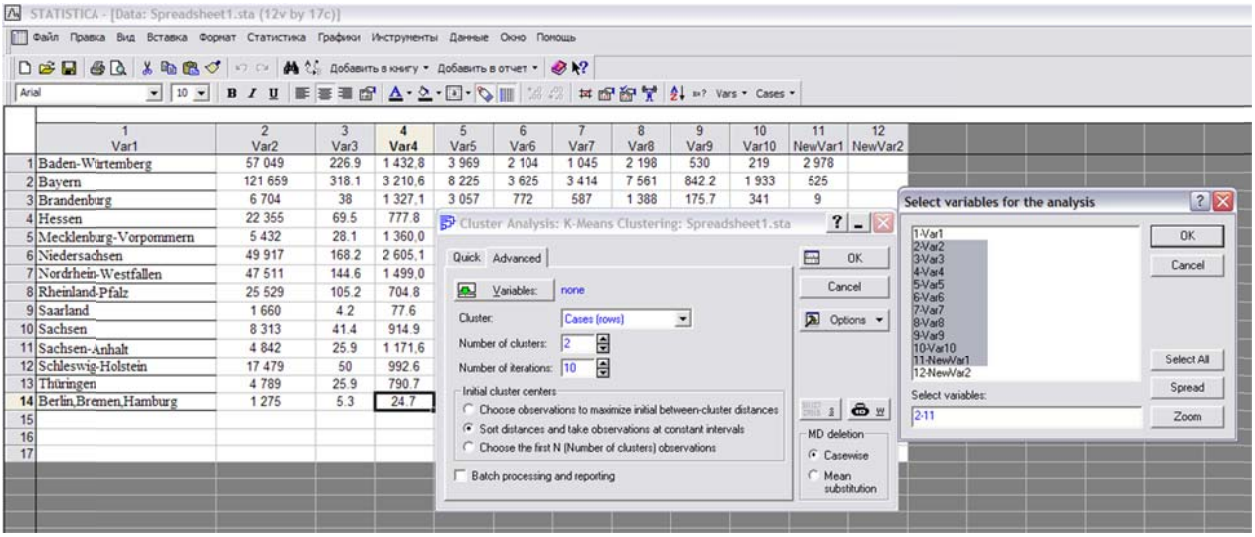
Subcluster. Diese Gruppe kann man als führende Regionen bezeichnen. Die übrigen Bundesländer bilden das zweite Cluster und können als rückständige Regionen in der landwirtschaftlichen Entwicklung genannt werden.

3.2 Anwendung des K-means Verfahren

Neben Ward-Verfahren wird bei der Clusteranalyse häufig das K-means Verfahren angewendet. Das K-means-Verfahren bildet Gruppen ohne einen Aggregationsprozess, in dessen Verlauf topologische Eigenschaften wirksam werden können. Grundlage ist wieder die Repräsentierung der Objekte in einem reellen Variablenraum mit euklidischer Metrik. Mit Hilfe dieser Metrik wird ein globales Maß für die Binnenheterogenität der Gruppen definiert. Dann wird nach einer Gruppierung gesucht, die dieses Maß minimiert, wobei die Anzahl der Gruppen vorgegeben ist.

Das Verfahren ermittelt also keine Gruppierungen, in denen einzelne Cluster besonders kompakt auf Kosten hoher Heterogenität anderer Gruppen sind, sondern Gruppen mit einer „mittleren“ Homogenität. Sein prinzipieller Nachteil liegt darin, dass die Anzahl der zu prüfenden Gruppierungen auch bei Stichproben kleineren Umfangs schon enorm groß ist, so dass Kapazitätsgrenzen auch sehr leistungsfähiger Rechner schnell erreicht werden. Daher wird das Kriterium der minimalen globalen Binnenvarianz durch ein relatives Kriterium ersetzt: Es werden Gruppierungen gesucht, bei denen jedes Objekt zum Schwerpunkt seiner Gruppe einen kleineren Abstand besitzt als zu den anderen Gruppenschwerpunkten.

Abbildung 2. Menü zu K-means in „STATISTICA 6.0“



Ein Problem des K-means-Verfahrens ist die Vorgabe der Anzahl der Cluster. Hierzu sind Vorinformationen über die Clusterstruktur notwendig. Diese können aus der Analyse anderer Daten stammen, zum Beispiel aus hierarchisch-agglomerativen Analysen oder aus der Theorie. Im unserem Fall werden die Ergebnisse des Ward-Verfahrens angewendet. In jedem Fall wird man sich in der Anwendung nicht auf eine einzige Vorgabe beschränken, sondern Lösungen zu einem Intervall von möglichen Clusteranzahlen ermitteln.

Unter dem Menüpunkt „Statistica-Multidimensional“ wird in „STATISTICA 6.0“ K-means-Clusterung angeboten, was auf der Abbildung 2 dargestellt ist. Im Hauptfenster zu diesem Menü können die Variablen und die Zahl der Cluster ausgewählt werden.

Danach werden vom Programm zwei Cluster definiert (Abbildung 3). Die Ergebnisse sind ähnlich dem Ward-Verfahren, aber Brandenburg und Mecklenburg-Vorpommern wurden wegen der großen Distanz zur zweiten Gruppe zugerechnet. Und es ist mehr zweckmäßig.

Abbildung 3. K-means Verfahren in der Clusterbildung

	Members of Cluster Number 1 (Spreadsheet1.sta) and Distances from Respective Cluster Center Cluster contains 7 cases						
	Baden-Würtemberg	Bayern	**Niedersachsen**	Nordrhein-Westfallen	Saarland	Sachsen-Anhalt	Berlin,Bremen,Hamburg
Distance	92.96197	215.7749	84.29597	64.31773	125.4275	173.5998	109.1455

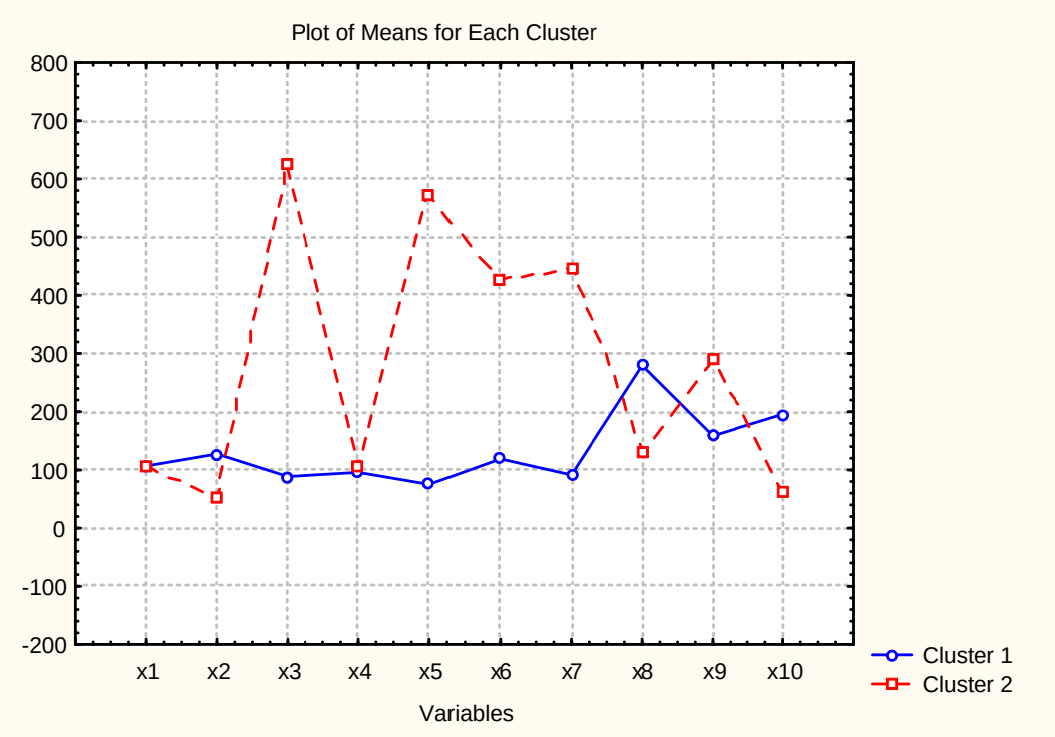
	Members of Cluster Number 2 (Spreadsheet1.sta) and Distances from Respective Cluster Center Cluster contains 7 cases						
	Brandenburg	Hessen	Mecklenburg-Vorpommern	Rheinland-Pfalz	Sachsen	Schleswig-Holstein	**Thüringen**
Distance	215.9529	188.6404	228.3243	143.6789	148.4025	240.3178	190.4973

4. Interpretation der Ergebnisse

Um die Ergebnisse besser zu interpretieren, kann man die Mittelwerte jedes Clusters analysieren und ihre Grafik ableiten, die die Unterschiede zwischen beiden Gruppen anschaulich in Abbildung 4 darstellt. Aus der Grafik kann man schlussfolgern, dass

- die mittlere Zahl der landwirtschaftlichen Betriebe und Getreideernte in beiden Cluster ähnlich sind
- die Mittelwerte der landwirtschaftlich genutzten Fläche, des Schweinebestandes, Rinderbestandes und der Milcherzeugung im zweiten Cluster dem ersten Cluster weit übersteigen
- die Kartoffelernte im zweiten Cluster auch höher als im ersten Cluster ist.

Abbildung 4. die Zentralwerte für jedes Cluster



In diesem Zusammenhang könnten das zweite Cluster als führende Bundesländer und das erste als rückständige Bundesländer gezeichnet werden. Das K-means Verfahren kann nicht für alle, sondern für einige stichprobenartige Variablen durchgeführt werden, aber letztendlich werden dieselben Cluster definiert.

Die Clusteranalyse spielt eine große Rolle für die regionale Politik verschiedenartiger Bundesländer. Sie hilft, den Beitrag der Regionen im Marktsystem einzuschätzen und die Prozesse effektiv regional zu differenzieren. So folgen Modelle für Subventionen, die auf führende und rückständige Regionen eingehen. Clustertechnologien können auch als Projekte programmatisch angewandt werden. Daran kann auch der Staat intensiv teilnehmen und solche Programme finanzieren.

5. Literatur

Bacher, J., Pöge, A., Wenzig, K. (2010): Clusteranalyse, anwendungsorientierte Einführung in Klassifikationsverfahren. München, Oldenburg.

Eckstein, P. (2010): Statistik für Wirtschaftswissenschaftler, eine realdatenbasierte Einführung mit SPSS. Wiesbaden, Gabler.

Janssen, J., Laatz, W. (2010): Statistische Datenanalyse mit SPSS, eine anwendungsorientierte Einführung in das Basissystem und das Modul Exakte Tests. Berlin [u.a.], Springer.

Schermelleh-Engel, K., Werner, C. (2007): Computerunterstützte Einführung in multivariate statistische Analyseverfahren. Frankfurt-am-Main, Universität Frankfurt.

Ein Autor (MF) bedankt sich für die großzügige finanzielle Unterstützung durch den Deutschen Akademischen Austauschdienst (DAAD).