Клъстерният анализ – инструмент за групиране на отраслите от преработващата промишленост по области

ас. д-р Калин Крумов
Университет за национално и световно стопанство - София

Въведение

Приложение на клъстерния анализ

Огромното количество информация за отраслите от преработващата промишленост в България по области по избрани променливи не позволява на изследователя самостоятелно да извърши прецизно класифициране на отраслите в относително еднородни групи. Именно прецизната класификация на отраслите от преработващата промишленост налага използването на специализиран софтуер за обработка на информацията. Предимствата софтуера намират израз в това, че той: позволява обработката на огромни масиви от данни; позволява осъществяването на процедура на валидизация на броя на получените клъстери (bootstrap процедура); дава възможност за добра визуализация на получените резултати чрез вградена процедура на многомерно скалиране.

Клъстерния анализ в настоящото изследване е осъществен в седем стъпки[1]:

· **Първа стъпка**: Определяне на класификационни критерии (количествени характеристики на обектите);

· **Втора стъпка**: Избор на мярка за дистанция;

· **Трета стъпка:** Избор на клъстерен метод;

· **Четвърта стъпка**: Определяне на броя на клъстерите (групите);

· **Пета стъпка:** Тълкуване на съдържанието на клъстерите (групите);

· **Шеста стъпка:** Оценка на валидността на клъстерите (групите);

· **Седма стъпка:** Профилиране на клъстерите (групите).

В изследването вместо думата клъстер ще бъде използвана думата група.

Относно избора на класификационни критерии

За групиране на отраслите от преработващата промишленост в България е необходимо да бъдат приложени показатели, които характеризират тяхното състояние по области. Източник на данните за стойността за отраслите от преработващата промишленост в България през 2000 г. и 2006 г. е Националния статистически институт.

Първият етап от клъстерния анализ обхваща избора на класификационни критерии. За тази цел са подбрани 15 променливи за отраслите от преработващата промишленост по области в България за 2000 г. и 2006 г. Изборът на класификационни критерии изисква проверка на връзките и зависимостите между тези променливи за двете години. Тази процедура е извършена за отстраняване на променливи, които индикират силна връзка помежду си, като по този начин силно е ограничена възможността за формиране на несъществуващи групи (клъстери) от отрасли. Всички 15 променливи за 2000 г. и 2006 г. са представители на силните скали, поради което връзките между тях са изследвани с коефициентът на корелация на Пирсън. Изчисленията са извършени със софтурер SPSS 17.0.

Получената корелационна матрица показва, че през 2000 г. между 6 от общо 15 променливи съществува слаба връзка. Така общият брой на променливите през 2000 г. е редуциран от 15 до 6. Всички връзки между посочените променливи са статистически значими при двустранна критична област и риск за грешка $\alpha = 0{,}01$. От тях средна сила[2] на

[1] Желев, С., Маркетингови изследвания за маркетингови решения, София, И. Тракия-М, 2000

[2] Коефициентите на корелация са нормирани в границите от 0 до 1. Условно е прието, че когато стойността им е в границите от 0 до 0.3 връзката е слаба, от 0.3 до 0.7 средна или умерена и над 0.7 силна.

връзката показват 7 коефициента, докато останалите коефициенти показват слаба сила на връзката.

Корелационната матрица за 2006 г. показва, че между 6 от общо 15 променливи съществува слаба връзка. Чрез премахване на променливите, които индикират силна връзка с отаналите, са подбрани 6, които са използвани като класификационни критерии. Корелационната матрица за 2006 г. показва, че връзките между тях са статистически значими при двустранна критична област и риск за грешка α = 0,01. Изключение прави единствено само един коефициент на корелация от корелационната матрица, който е статистически значим при риск за грешка над α = 0.01 равнище. От всички корелационни коефициенти в матрицата 10 показват средна сила на връзката, а останалите индикират слаба връзка между променливите.

Резултатите от направения корелационен анализ показват, че между селектираните класификационни променливи през 2000 г. и 2006 г. не се наблюдават силни връзки, което означава, че на този етап те отговарят на изискванията и коректно могат да се използват за целите на клъстерния анализ.

След определяне на класификационните променливи възниква следният въпрос: Така избраните класификационнни критерии (променливи) съизмерими ли са една спрямо друга? По правило за коректното осъществяване на клъстерния анализ е необходимо скалите към които принадлежат отделните променливи да са идентични. В противен случай трябва да бъде осъществена процедура по стандартизиране на данните. В случаите когато всички променливи принадлежат на интервалните (количествени) скали, но са измерени с различни единици, както е в настоящото изследване, е необходимо да се приложи процедура на стандартизация на данните. За постигането на съизмеримост на данните е необходимо осъществяване на стандартизация на всяка една променлива по вариацията на единиците или приоритетно по всеки метод, който се базира на стандартната девиация на обектите, които ще бъдат клъстеризирани[3]. Изходните данни за клъстеризиране на отраслите от преработващата промишленост в България по области през 2000 и 2006 г. съответстват именно на този случай – бройни са, но са измерени в различни единици. Това прави задължително провеждането на процедура по стандартизация. Стандартизирането на данните в изследването е извършено посредством z-scores[4] с програмата Clustan Graphics. Стандартизацията по z-scores е за предпочане пред стандартизацията по рангове. При стандартизацията по z-scores получените стойности не са детерминирани по двете екстремни стойности x_{max} и x_{min}, а по дисперсията на променливата k, като измерител на нейната вариация σ_k^2 или стандартната девиация σ_k. Това е и причината, поради която стандартизирането на данните по z-scores е препоръчван метод.[5] Стандартизирането на класификационните променливи за 2000 г. и 2006 г. удовлетворява изискването за съизмеримост на променливите и позволява да се премине към следващия етап от клъстрния анализ, а именно избор на мярка за разстояние.

Относно избора на мярка за разстояние

Втората стъпка в процеса на реализация на клъстерния анализ се свързва с избор на мярка за разстояние между обектите, което да направи възможно групирането им. За сравняването на обектите в изследването е използван квадрата на евклидовото разстояние (Squared Euclidean Distance). Според Арсо Манов[6] квадрата на евклидовото разстояние „намира най-широко приложение при измерване на разстояния в многомерното пространство". Този метод позволява стойностите на обектите i и j и се сравняват с помощта на следното стандартно уравнение[7]:

$$d_{ij}^2 = \frac{\sum_k w_{ijk} \left(x_{ik} - x_{jk} \right)^2}{\sum_k w_{ijk}},$$

където x_{ik} показва стойността на променливата k на обект i, а w_{ijk} в зависимост от това дали сравнението е валидно или не заема стойност от 1 или 0. В случаите когато липсват данни за единия или за двата обекта едновременно - сравнението е невалидно, т.е.w_{ijk}=0.

[3] Brian S. Everitt, Sabine Landau, Morven Leese, Cluster Analysis, London, "Arnold", 2001, p. 51-52

[4] Стандартицазията (трансформацията) на данните посредством тази опция се извършва по начин, по който стойностите за всяка променлива придобива средна 0 и стандартна девиация от 1. Източник: www.clustan.com/general_distances.html

[5] Източник: www.clustan.com/general_distances.html

[6] Манов, А. Многомерни стастистически методи със SPSS, София, УИ. Стопанство, 2002, с. 204

[7] Източник: www.clustan.com/general_distances.html2

Изчислението на разстоянието между отраслите от преработващата промишленост по области в страната по отношение на класификационните променливи има своите особености. Така например в изследването са изпълнени следните условия:

· класификационните променливи за отраслите от преработващата промишленост по области са количествени както за 2000 г., така и за 2006 г.;

· стойностите на променливите са стандартизирани с z-scores функцията т.е. имат за средни 0, а стандартната им девиация е 1.

Когато са изпълнени тези условия изчислението на евклидовото разстояние между обектите се извършва с помощта на уравнение[8] (2).

$$x_{ik}^{*} = \frac{x_{ij} - \mu_k}{\sigma_k},$$

където μ_k е средната, а σ_k е стандартната девиация на променлива k. Уравнение (2) в изследването е използвано за трансформиране на квадрата на евклидовото разстояние до записа в уравнение[9] (3), който е използван като мярка за изследване на разстоянието между обектите.

$$d_{ijk}^{2} = \frac{(x_{ij} - \mu_k)^2}{\sigma_k^{2}},$$

Изчислението на квадратите на евклидовото разстояние за отраслите от преработващата промишленост в България по области през 2000 г. и 2006 г. е извършено с уравнение (3) посредством програмата Clustan Graphics.

Относно избора на клъстерен метод

В изследването е използван клъстерен метод, който принадлежи към групата на йерархическите агломеративни методи на клъстеризация. Основните предимства на тези методи се свързват с добрите възможности за визуализация и интерпретация на получените резултати, което е добра основа за вземане на решения в анализа. Йерархическите методи на клъстеризация имат сериозно предимство пред нейерархическите при определяне на броя на групите обекти. При методите на йерархическа клъстеризация тази процедура се осъществява в рамките на утвърдени правила, което силно ограничава полето на субективните решения в анализа, докато при нейерархическите съществува голяма вероятност от субективна намеса при определяне броя на клъстерите. Освен това йерархическите методи са прилагани често в изследвания в областта на социалните науки, биологията, медицината, астрономията, химията, изследването на космическото пространство, архитектурата и други, поради което тези методи са добре проучени от към надеждност. Недостатък на йерархическите методи е липсата на възможност за извършване на итерации, след като изследваните единици са обособени в групи.

Изборът на йерархически метод по-нататък наложи проучване на литературата в областта на клъстерния анализ и по-точно акцентира върху изводите относно надеждността на получените резултати от изследователите използвали различни клъстерни методи в своите изследвания. Редица автори[10] посочват метода на Уорд[11] и метода междугруповото свързване (Average linkage), като инструменти, които дават най-добри резултати, а получените с помощта на тези методи групи обекти могат леко да се тълкуват и интерпретират. Това е и причината, както посочва Блешвийлд[12] методът на Уорд да е един от най-широко използваните в социалните науки.

Отговорът на въпроса кой от двата препоръчвани агломеративни методи на йерархическа клъстеризация е по-подходящ за приложение в настоящото изследване, изисква анализ на получените от всеки едни от тях резултати. Изводите за приложимостта и адекватността на клъстерните методи са направени въз основа на резултатите от бутстрап (bootstrap)[13] процедура по валидизация на броя на значимите групи от отрасли за всеки един от двата метода. Резултатите от тази процедура са представени в Таблица 1.

[8] Източник: www.clustan.com/general_distances.html
[9] Източник: www.clustan.com/general_distances.html

[10] Duflou H., Maenhaut W., Application of principal component and cluster analysis to the study ot the distribution of minor and trace element in the normal human brain, Cheomometrics and Intelligent Laboratory Systems, 1990; Baxter J., Exploratory Multivariate Anaysis in Archaeology, Edinburgh, Edinburgh University Press, 1994; Milligan G., An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms, Psychometrika, 1980 и други.
[11] Ward J., Herarchical grouping to optimizean objective function, Journal of the American statistical Association, 1963
[12] Blashfield R. K., The growth of cluster analysis: Tryon, Ward and Johnson, Multivariate Bechavioral Research, 1980

Таблица № 1

№	Клъстерен метод	2000 г.	2006 г.
1	**Increase in Sum of Squares** (Метод на Уорд)	**6 significant**	**8 significant**
2	**Average linkage** (Междугрупов метод на свързване)	**0 significant**	**0 significant**

При всеки клъстерен метод е използвана бутстрап (bootstrap) процедура за валидизация на броя на клъстерите с 1000 завършени опита

Изчисленията са извършени с програмата Clstan Graphics

Резултатите от Таблицата показват, че приложението на метода на Уорд води до формирането на 6 статистически значими групи от отрасли от преработващата промишленост през 2000 г. и 8 статистически значими групи от отрасли от преработващата промишленост през 2006 г. В същото време приложението на методът на Междугруповото свързване показва, че той не се явява надежден инструмент за групиране на отраслите от преработващата промишленост по области в България, защото нито една от получените групи от отрасли през 2000 г. и 2006 г. не е статистически значима. Именно обстоятелството, че този метод отчита групи от отрасли, но не е достатъчно надежден инструмент за да ги определи като статистически значими, показва, че той е неподходящ за приложание в настоящото изследване. Доказателството за високата надежност на получените резултати от двата метода е приложената бутстрап процедура с 1000 завършени опита.

Всичко това дава основание да се направи заключението, че от двата най-често използвани (и препоръчвани) от специалистите в областта на клъстерния анализ методи за клъстеризация, методът на Уорд е по-подходящ за определяне броя на групите от отрасли на преработващата промишленост по области в България. Приложението му в настоящото изследването е свързано не само с висока надежност, но и позволява получените резултати да бъдат смислено интерпретирани.

Приложението на метода на Уорд в изследването преминава през следните етапи:

Първо: изчисляване на средните на класификационните променливи за отделните групи от отрасли;

Второ: изчисляване на квадратите на евклидовото разстояние на всеки отрасъл, включен в изследването до груповите средни;

Трето: сумиране на разстоянията между квадратите за всяка група от отрасли по уравнение[14] (4):

$$E_p = \sum_{i \in p} \sum_{j} \frac{(x_{ij} - \mu_{pj})^2}{v},$$

където: x_{ij} - отразява стойността на случай i по променливата j;

μ_{pj} - средна на група p;

v – брой на класификационните променливи.

Четвърто: обединяване на случаите в групи според критерия за най-малко нарастване на общата сума на квадратите на вътрешногруповите разстояния.

В изследването процедурата по приложение на метода на Уорд е осъществена с програмата Clustan Graphics. Изборът на подходящ клъстерен метод за определяне на групите от отрасли от преработващата промишленост позволява да се премине към следващия етап от клъстерния анализ.

Определяне броя на групите от отрасли

В настящото изследване броя на групите от отрасли от преработващата промишленост през 2000 г. и 2006 г. е определен с помощта на бутстрап (*bootstrap*) процедура по валидизация на броя на групите. Тази процедура на база на предварително определени алгоритми и поредица от предварително зададен брой опити подбира броя на групите по начин, при който техния брой съответства на поредния номер на онази стойност на коефициентите на сливане, след която стойностите на следващиите я коефициенти на сливане[15] рязко нараства.

[13] В статистиката, bootstrap е подход за получаване на статистически изводи, който попада в рамките на широк клас от методи за взимане на проби. Методът е много близък до техниките използвани при методът Монте Карло, но вместо напълно да генерира данни и на тяхна основа да прави заключения, bootstrap методът прави поредица от произволни извадки от оригиналните данни, въз основа на които прави заключения

[14] Източник: www.clustan.com

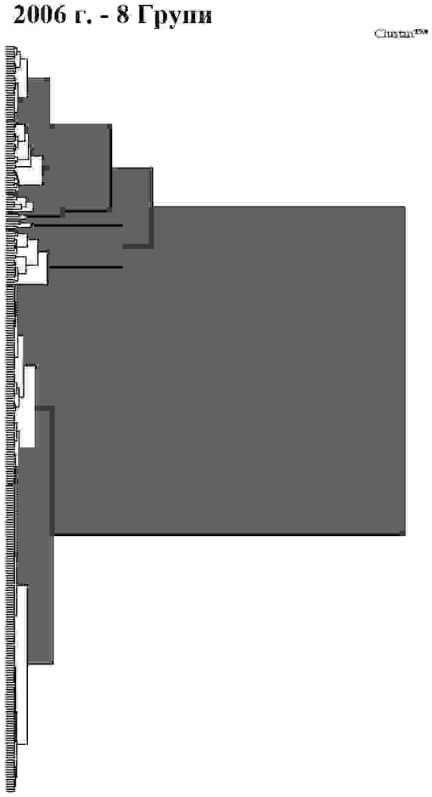
[15] Коефициентите на сливане показват разстоянието след, преминаването на което две групи от обекти формират нова по-голяма група. По принцип колкото по-висока е стойността на коефициентите на сливане, толкова по различни една спрямо друга са формираните по класификационните критерии групи от обекти.

Логическата интерпретация подсказва, че колкото по-голям е броят на зададените опити в бутстрап процудурата, толкова по-точни ще са получените от нея резултати. В настоящото изследване при 308 отрасъла по области, групирани по 6 класификационни променливи извършването на 1000 опита в рамките на бутстрап процедурата са напълно достатъчни за да предоставят достатъчно надежна информация[16] за определяне на стойностите на коефициентите на сливане, а от там и на броя на групите. Както бе посочено в предходната точка, при метода на Уорд за данните от 2000 г. резултатите от бутстрап процедурата показват 6 статистически значими групи от отрасли, а за данните от 2006 г. 8 статистически значими групи от отрасли. Това е реалният брой на съществуващите групи от отрасли през двете години. Нарастването на броя на еднородните групи от отрасли от преработващата промишленост по области в България през периода 2000-2006 г. е следствие от тяхното развитие, изразено в промяна в класификационните променливи. Прави впечатление, че със засилване на развитието на отраслите от преработващата промишленост по области в страната броят на статистически значимите групи от отрасли нараства, което е доказателство за увеличаване на различиията между тях и формирането на процеси на задълбочаваща се диференциация.

Клъстери по отрасли на преработващата промишленост в България през 2000 г. и 2006 г.

Дендограма № 1
2000 г. - 6 Групи

Дендограма № 2
2006 г. - 8 Групи



Източник: Собствени изчисления

[16] В изследването на доходите на населението в 50 американски щата плюс окръг Колумбия през периода 1983-1988 г. Jeffrey A. Mills и Sourushe Zandvakili при 52 наблюдения за всяка година от изследвания период залагат общо 500 опита (итерации) в бутстрап процудурата, а при 4266 наблюдения за всяка година през периода 1979-1989 г. залагат общо 2000 опита (итерации) в бутстрап процедурата. В настоящото изследване при 1848 наблюдения за 2000 г. и 1848 наблюдени за 2006 г., отделно за двете години са заложени по 1000 опита (итерации) в бутстрап процедурите, чиито брой е достатъчен за получаване на надежни резултати. Виж. Jeffrey A. Mills, Sourushe Zandvakili, "Statistical Inference via Botstrapping for Measures of Ineguality", University of Cincinati, Departament of Economics, 1995.

Броят на групите от отрасли от преработващата промишленост по обалсти може да се види и на Дендограмите за 2000 и 2006 г. През 2000 година Дендограма №1 визуализира още агломерационните нива и позволява да се проследи кои групи на колко подгрупи се дезагрегират. Подобен анализ може да се направи и за 2006 г. И на двете Дендограми със жълт цвят са отбелязани групите които са статистически значими, докато всички останали групи са представени със син цвят. Заключенията, които могат да се направят на база на двете Дендограми са в следните направления:

На първо място. Колкото по-ниска е стойността на коефициентите на сливане, толкова по блзки една спрямо друга са формираните по класификационните критерии групи от отрасли. Стойността на коефициентите на сливане за групи № 5 и № 6 през 2000 г. е много по-близка отколкото стойността на тези коефициенти за останалите четири групи през годината. През 2006 г. се наблюдава същата зависимост – стойността на коефициентите на сливане за групи № 7 и № 8 е много по-близка спрямо стойността на коефициентите на сливане на всички останали шест групи през годината. Нещо повече стойността на коефициентите на сливане, които водят до образуване на двете големи групи от отрасли през 2006 г. спрямо 2000 г. значително намалява (от 16.056 до 7.991), което показва, че различията между двете големи групи от отрасли през 2006 г. спрямо двете големи групи от отрасли през 2000 г. са значителни. Всичко това показва, че сходството между двете големи групи от отрасли през 2006 г. е по-голямо спрямо сходството между двете големи групи от отрасли през 2000 г. В същото време трябва да се посочи, че преобладаваща част от отраслите по области, които образуват двете големи групи през 2000 г., са индентични с отраслите по области, които образуват двете големи групи през 2006 г. За всички отрасли, които попадат в тези групи през двете години е характерна ниска степен на развитие по класификационните променливи, което е причината за високото сходство помежду им.

На второ място. През 2000 г. между останалите четири групи се наблюдават по-големи различия в стойността на коефициентите на сливане, отколокото различията в стойността на коефициентите на сливане при групи № 5 и № 6 през същата година. Това показва, че различията между отраслите в първите четири групи са по-големи отколкото различията между отраслите от пета и шеста група през 2000 г. Отраслите в първите четири групи имат по-висока степен на развитие, което е причината за формирането на по-големи различия между тях. С увеличаване степентта на развитие нарастват и различията между отраслите по класификационни променливи, като едни отрасли имат по-високи стойности по едни променливи а други отрасли по други променливи. Именно процесите на развитие на отраслите от първите четири групи през 2000 г. водят до увеличаване на различията по класификационни кретирии между тях и се явяват причината за нарастване на броя на групите през 2006 г.

Ако се обобщи всичко казано до тук изводът, който може да се направи, е че общият брой на статистически значимите групи от отрасли от преработващата промишленост по области в България нараства от 6 на 8 през периода 2000-2006 г. Причина за увеличаване броя на групите е позитивната промяна по класификационни променливи на част от отраслите от първите четири групи през 2000 г. Информацията за броя на значимите групи от отрасли през изследвания период ще послужи в следващите части на изследването при интерпретацията на т.н. външна валидизация на броя на групите, която е най-силният метод за определяне на техния брой.

Тълкуване съдържанието на групите.

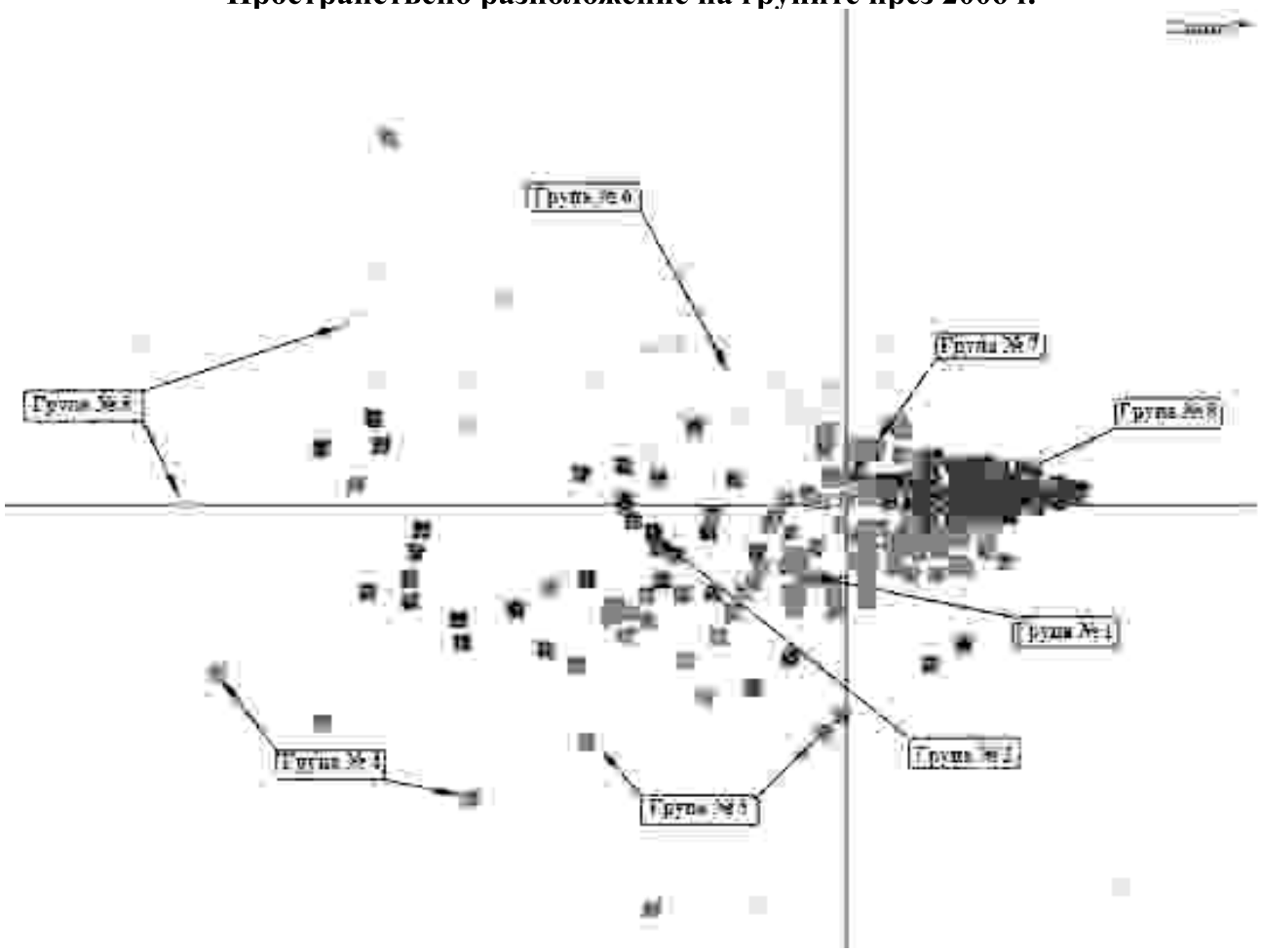
Съдържанието на групите от отрасли през 2000 г. и 2006 г. е интерпретирано въз основа на средните на класификационните променливи за всяка една от тях. Средните на променливите за всяка група от отрасли са получени с програмата Clustan Graphics в хода на приложение на клъстерния анализ. На второ място за тълкуване на съдържанието на групите е използван методът на Многомерно скалиране[17]. Изходни данни за приложението му са резултатите получени от клъстерният анализ

[17]Многомерното скалиране е метод за представяне на наблюдаваните обекти като точки в пространството, по начин който позволява разстоянията между тях да се запазят

до този етап. Този метод значително подобрява възможностите за интерпретацията на резултатите от клъстерния анализ чрез представянето им в многомерното пространство. Целта на метода на Многомерно скалиране е чрез използването на съвкупност от процедури да се намери множество от точки в пространството, които представят изследваните обекти - в настоящото изследване това са отраслите от преработващата промишленост по области в България. При този метод взаимоотношенията между изследваните обекти са представени в многомерното пространство чрез пространствени карти, които дават допълнителна позватателна стойност на получените резултати от клъстерния анализ и позволяват тези резултати да бъдат по лесно интерпретирани.

Пространствена карта № 1

Пространствено разположение на групите през 2006 г.



50 trials completed, Maximum 5000 Iterations per trial, 47278 proximities were read,
Minimum stress = 0.8080% in trial 17, Fit is excellent
Източник: Собствени изчисления

Оценяване надеждността на модела на Многомерно скалиране. Оценката на надеждността на модела на многомерно скалиране в използвания софтуер се прави с критерия - Standardized Residual Sum of Square (STRESS). При метричния метод на скалиране Джозеф Бернард Кръскал препоръчва използването именно на този критерии. Той въвежда (1964a) следните граници за качеството на получените резултати[18]. По принцип STRESS критерият е мярка за липсата на съответствие – той показва каква част от вариацията на оптимално скалираните данни не може да бъде обяснена от модела. Колкото по-висока е стойността на STRESS критерия, толкова по-ниско е равнището на съответствие на модела. В настоящото изследване в модела на многомерно скалиране за 2000 г. стойността на STRESS е приблизително 0.89% и показва високата му надеждност. Според таблицата на Кръскал ако стойността на STRESS критерия е 2.5% моделът е отличен, а когато клони към нула се приема за идеален. Изводът за качеството на резултатите за модела на многомерно скалиране от 2000 г. който може да се направи е че той е идеален. В модела от 2006 г. стойността на STRESS е още по-ниска – около 0.81%, което показва и по-високата му надеждност спрямо модела от 2000 г. С други думи през 2000 г. 0.89%, а през 2006 г. 0.81% от вариацията на обектите по класификационните критерии не е обяснена от пространствените карти. И двата модела на многомерно скалиране са получени чрез приложението на програмата Clustan Graphics, като за всеки са направени 50 опита с максимален брой от 5000 итерации на опит. Най-ниската стойност на STRESS критерия за модела от 2000 г. е получена от опит 34, а на модела от 2006 г. от опит 17. В заключение основният извод, които се налага по отношение на получените данни, показват че резултатите от изчислението на двата модела са отлични.

[18] Kruskal, Joseph Bernard. (1964a). Multidimensional scaling by optimizing Goodness of fit to a Nonmetric Hipothesis, Psychometrika, 29, 1-28. Kruskal, Joseph Bernard. (1964b). Nonmetric multidimensional scaling: A numerical method, Psychometrika, 29, 115-129.

Стойности на Standardized Residual Sum of Square (STRESS)

Минимална стойност на индекса в %	**20**	**10**	**5**	**2.5**	**0**
Качество на получения резултат	Слаб	Задоволителен	Добър	Отличен	Идеален

Оценка на валидността на групите

Етапът на оценяване валидността на формираните групи обекти във всяко едно изследване представлява една от най-трудните стъпки в анализа. Въпреки, че теорията и практиката препоръчват определени статистически методи и инструменти за валидизиране на получените групи, много от тези методи са трудно приложими (поради липсата на необходимите условия за това) или резултатите, които дават не са достатъчно надежни.

До тук в клъстерния анализ бе доказана статистическата значимост на броя на групите посредством бутстрап процедурата по валидизация към клъстерния метод. На този етап ще бъде извършена валидизация на групите от отрасли по отношение на класификационните променливи за 2000 г. и 2006 г. Същността му се състои в измерване на статистическата значимост на разликите между получените групи на база на същите тези променливи. За да се постигне тази цел е необходимо на първо място да се създаде променлива, която съдържа груповата принадлежност на отраслите на преработващата промишленост по области в страната през 2000 г. и 2006 г. За тази променлива през двете години е характерно, че принадлежи към слабите скали. За да се направят изводи за статистическата значимост на разликите между получените групи е необходимо да се изследва взаимодействието между груповата принадлежност от една страна и класификационните променливи от друга за всяка от двете години. При изследване на това взаимодействие променливата, която показва груповата принадлежност на отраслите по области се разглежда като независима променлива, а класификационните критерии се изследват в качеството им на зависими променливи. Преди да се пристъпи към осъществяване на самия анализ е необходимо да се отбележи, че всички класификационни променливи са количествени и принадлежат на силните скали. В такива случай (когато независимата променлива принадлежи на слабите скали, а зависимите променливи принадлежат към силните скали) подходящ статистически инструмент за изследване на посочените зависимости се явява

дисперсионният анализ. От своя страна дисперсионният анализ[19] дава достоверни резултати, само когато са спазени следните условия:

Първо: Данните са получени от извадка, осъществена чрез прост случаен избор;

Второ: Честотното разпределение на резултативните променливи Y при различните равнища на факторната променлива X са приблизително нормални;

Трето: Дисперсиите в разпределенията на зависимите променливи Y при различните равнища на факторната променлива X са еднакви.

Данните включени в настоящото изследване са представителни защото обхващат всички области в България, т.е. тук не става въпрос за направата на прост случаен избор, защото от такъв просто няма нужда - налице са всички данни за изследваните явления. Тези данни са предоставени от Националния статистически институт, който е гарант за качество и изчерпателност. Поради това първото изискване за приложимост на параметричния дисперсионен анализ може да се счита за изпълнено.

Второто необходимо условие за приложимост на параметричния дисперсионен анализ е наличие на нормално разпределение на променливите, което в случая е изследвано с коефициентите на асиметрия[20] и ексцес[21]. След това е определено равнището на значимост, което в случая е: $\alpha = 0.01$ и за двата коефициента. По статистическите таблици е определена т.н. критична или теоретична стойност на критерия. Те показват, че нито една от променливите през 2000 г. и 2006 година няма нормално разпределение, защото за вясяка една променлива емпиричната стойност е по-голяма от критичната. Ако за променливите емпиричната стойност бе по-малка или равна на критичната стойност щеше да се приеме, че разпределението не се различава съществено от нормалното, но се наблюдава точно обратното. Изводът който се налага, е че второто необходимо условие за приложение на параметричния дисперсионен анализ не е изпълнено.

Дисперсионният анализ е много чувствителен към третото условие - изискването за наличие на хомоскедастицитет. Като цяло в случаите на нарушаване на това условие нулевата хипотеза се отхвърля трудно въпреки че не е вярна. За да се избегне това в статистическата теория са разработени различни тестове, но по-голямата част от тях поставя задължителното изискване за нормалност на разпределението, което както вече бе показано отсъства. В приложните изследвания често се използва теста на Ливин, защото се счита за устойчив при нарушаване на условието за нормалност на разпределението. В изследването тестът на Ливин е приложен в следната последователост:

На първо място: Определени са Нулевата H0 и алтернативната H1 хипотези.

H0: Между дисперсиите на независимата и зависимите променливи не съществува статистически значима разлика;

H1: Между дисперсиите съществува статистически значима, закономерна разлика.

На второ място е определено равнището на значимост: $\alpha = 0.05$

На трето място е направено сравнение на наблюдаваното и критичното равнище на значимост. Равнището на значимост на теста на Ливин за всички променливи и през 2000 г. и 2006 г. е $\alpha s = 0.00$ и е по-малко от грешката $\alpha = 0.05$, което означава, че за всички променливи нулевата хипотеза се отхвърля и се приема за вярна алтернативната т.е. не може да се твърди, че съществува статистическо значимо равенство между дисперсиите на променливите[22]. От направените изчисления следва, че е нарушено изискването за хомоскедастицитет за всички променливи през 2000 г. и 2006 г.

Процедурите по установяване доколко изискванията за приложение на параметричния дисперсионен анализ са удовлетворени,

[19] Вж. Гоев, В. „Статистическа обработка и анализ на информацията от социологически, маркетингови и политически изследвания", София, УИ „Стопанство", 1996, с. 133 – 146
Манов Арсо, „Статистика със SPSS", И. Рикол – Б, 2000, с. 190 – 200

[20] Коефициент на асиметрия характеризира формата по отношение на характера и степента на отклонение от симетричността. Степента на симетричност на изследваните разпределения може да бъде различна в зависимост от концентрацията на единиците от двете страни на центъра на разпределението.

[21] Ексцесът характеризира външната източеност на разпределението, т.е. изследва дали по-голямата част от единиците са концентрирани близо до центъра или обратно.

[22] Вж. Гоев Валентин, „Статистическа обработка и анализ на информацията от социологически, маркетингови и политически изследвания със SPSS", УИ „Стопанство", София, 1996, с. 140
Манов Арсо, „Статистика със SPSS", София, И. Рикол – Б, 2000, с. 196

показват че две (наличието на нормално разпределение и равенство на дисперсиите) от трите условия за приложението на този анализ не са изпълнени. Това налага използването на непараметрични тестове при три и повече независими извадки за проверка на статистическата значимост на разликите между получените групи от обекти и класификационните променливи за 2000 г. и 2006 г. От тях, тестът на Кръскал-Уолис[23], който представлява непараметричен аналог на еднофакторния дисперсионен анализ е неприложим, защото изискването за хомоскедастицитет е нарушено[24]. В същото време неприложим е и Медианния тест при три и повече независими извадки за данните от 2000 г. и 2006 г., защото същите данните не удовлетворяват условията за приложение на теста (резултативните таблици в някои от клетките има очаквана честота, която е по-малка от единица и на второ място в повече от 20% от клетките има очаквани честоти, които са по-малки от пет[25]).

В тази ситуация за измерване на зависимостите между променливата, показваща груповата принадлежност на отраслите от една страна и класификационните променливи за същите отрасли е използван друг непараметричен метод – Коефициент на корелация на Спирман. Той представлява алтернатива на ленейния коефициент на корелация, защото е приложим не само в случайте, когато изследваните променливи са метрирани и имат двумерно нормално разпределение, но и когато „не са изпълнени посочените условия, т.е. когато променливите величини са ординални или когато тези величини са метрирани, но няма доказателства, че двумерното им разпределение е достатъчно близко до нормалното.” . Коефициентът на корелация на Спирман (ρ_s) има същата интерпретация като линейния коефициент на корелация r. Той променя стойността си в границите от -1 до 1.

За формиране на изводи, коефициент на корелация на Спирман може да се разглежда като оценка на неизвестен параметър с помощта на коефициента ρ. За тази цел анализът ще премине в следната последователност:

На първо място са формулирани две хипотези: нулева H0 и алтернативна H1.

Нулевата хипотеза гласи, че коефициента ρ е статистически незначим, (H0: ρ = 0), т.е. несъществува значима връзка между груповата принадлежност и класификационните променливи.

Алтернативната хипотеза гласи, че коефициента ρ е статистически значим (H1: ρ≠0), т.е. между груповата принадлежност и класифи-кационните променливи съществува значима връзка.

На второ място: определен е риска за грешка: **α = 0.05**

На трето място емпиричната характеристика (t_{emp}) за проверка на хипотезата е получена по уравнение (5).

$$t = \rho_s \cdot \sqrt{\frac{n-2}{1-\rho_s^2}},$$

където: ρs- коефициента на корелация;

v – степен на свобода, v = (n - 2);

n – общият брой на случаите.

На четвърто място: критичната област е определена като двустранна.

На пето място: Теоретичната характеристика (критичната стойност - $t_{\alpha\upsilon}$) е определена с помощта на таблиците за t разпределението при двустранна критична област, критично равнище на значимост α= 0.05 и степени на свобода v.

На шесто място: сравнението на емпиричната с теоретичната характеристики за 2000 г. и 2006 г. за всяка една от променливите при двустранна критична област показва, че $t_{emp} > (005,)t_{\alpha=0=0,5\upsilon}$ т.е. че нулевата хипотеза се отхвърля, а коефициентите на корелация се приемат за статистически значими. До същия извод се стига и при сравнение на равнищата на значимост. За всяка една променлива през 2000 г. и 2006 г. е изпълнено условието $\alpha_s /2 < \alpha$, където α_s е наблюдаваното равнище на значимост, а α е критичното равнище на значимост. Това дава основание да се приеме алтернативната хипотеза, която гласи, че коефициентите на корелация са статистически значими, т.е. между груповата принадлежност и класифика-ционните променливи е налице силна статистически значима връзка. Наличието на

[23] Гоев Валентин, „Статистическа обработка и анализ на информацията от социологически, маркетингови и политически изследвания със SPSS”, УИ „Стопанство”, София, 1996, с. 141

[24] Манов Арсо, „Статистика със SPSS”, София, И. Рикол – Б, 2000, с. 386

[25] Манов Арсо, „Статистика със SPSS”, София, И. Рикол – Б, 2000, с. 388

статистически значима връзка е индикатор за валидността на получените групи от отрасли по отношение на класификационните променливи. Получените коефициенти на корелация на Спирман в изследването са интерпретирани с помощта на коефициенти на детерминация, което повишава тяхната точност. В изследването е използван коригиран коефициент на детерминация, който е изчислен с помощта на уравнение (6).

$$\rho^2_{s(adjusted)} = \rho_s{}^2 - \frac{p-1}{n-p}\frac{1-\rho_s{}^2}{},$$

където p е броят на оценяваните параметри в модела. Общите промени в класификационните променливи, които не могат да се обяснят с промени в груповата принадлежност [29] се дължат на други фактори, които са могли да оказват влияние върху тях, но не са включени в анализа.

Въз основа на резултатите от корелационния анализ и коефициентите на детерминация е направено заключението, че формираните групи от отрасли през 2000 г. и 2006 г. се различават значително един спрямо друг откъм стойност на класификационните променливи. В същото време валидността на получените групи от обекти е потвърдена още веднъж и при сравнение на получените резултати през 2000 г. и 2006 г.

До тук процесът на валидизация на получените групи от отрасли на преработващата промишленост по области бе извършена посредством методи за изследване на значимостта на разликите между тях, т.е. изследвано бе различието между групите.

Успоредно с този подход за валидизация е приложен и втори подход, чиято задача е свързана с изследване на хомогенността на отделните групи от отрасли през двете години. Един от широко използваните методи за изследване на хомогенността на групите е коефициентът F, емпиричната стойност на който се получава по уравнение (7).

$$F_{jg} = \frac{\sigma^2_{(j,g)}}{\sigma^2_{(j)}},$$

където: $\sigma_{(j,g)}{}^2$ е дисперсията на променлива j на група g;

$\sigma_{(j)}{}^2$ е дисперсията на променлива j за всички данни (общо всички групи).

От уравнението е видно, че с коефициентът F_{jg} се прави съпоставка между вътрешно груповата дисперсия и общата дисперсия. Най-висока хомогенност имат получените групи в случаите когато стойността на $F_{jg}<1$ за всички класификационни променливи.

Изчислението на F коефициента за групите от 2000 г. по всички променливи позволява да се направят следните изводи: група № 1, група № 5 и група № 6 са напълно хомогенни. Стойността на F коефициента за всички променливи в тях е по-ниска от единица. Като цяло група № 6 е най-хомогенната група от отрасли спрямо всички групи през 2000 г. Стойността на F коефициента за всички променливи в нея е по-ниска от 0.45. За група № 4 може да се твърди, че има високо равнище на хомогенност, въпреки, че стойпността на F за една променлива е по-висока от 2.1. За група № 2 и група № 3 през 2000 г. равнището на хомогенност е умерено. При група № 2 стойността на F за две от класификационните променливи е много по-висока от 1, за една от класификационните променливи стойността на F е около 1, а за останалите три класифика-ционни променливи е по-ниска от 1. Група № 3 има най-ниска хомогенност през 2000 г. Стойността на F за една от класифика-ционните променливи е над 3, за четири от класифика-ционните променливи между 1 и 2, и само за една класификационна променлива под 1.

През 2006 г. резултатите показват, че напълно хомогенни (по всички класифика-ционни променливи) са групи № 1, № 2, № 7 и № 8. От тях най-висока хомогенност има група № 8 – стойността на F за всяка една променлива е под 0.4 десети. На второ място групи № 3, № 4, № 5 и № 6 могат да се определят като относително хомогенни. Като цяло за всяка една от тях половината от променливите имат стойност на F по-висока от 1.

Направения преглед недвусмислено показва, че формираните групи от отрасли са вътрешно хомогенни. Групите, които имат много висока хомогенност са формирани от

[29] Тук е необходимо да се посочи, че връзката е изследвана при равни други условия, т.е. други фактори (освен клъстерната принадлежност), които са могли да оказват влияние върху нея не са включени в анализа, поради високото равнище на корелация.

отрасли, които са изостанали в своето развитие (или са слабо развити). На противоположния полюс се намират групите, чиято вътрешната хомогенност е по-ниска. Те са съставени преди всичко от по-силно развити отрасли. По част от класификационните променливи тези отрасли в отделните области имат високи или много високи стойности, докато по други по-ниски, което е и причината за по-ниската хомогенност в тези групи.

Извършените процедури по установяване валидността на броя на групите от отрасли от преработващата промишленост в България по области през 2000 г. и 2006 г. недвусмислено показа че такива групи реално съществуват. Валидността на групите от отрасли за всяка от двете години бе установена на три равнища: **първо** - валидизация на броя на групите чрез бутстрап процедура; **второ** – валидизация чрез корелационен анализ и коефициенти на детерминация, които показаха, че формираните групи от отрасли се различават значително един спрямо друг откъм стойност на класификационните променливи; **трето** – валидизация чрез изследване на вътрешната хомогенност на получените от клъстерния анализ групи от отрасли. Всичко това дава основание получените групи от отрасли от преработващата промишленост по области да се приемат за валидни.

Профилиране на групите

Последният етап от клъстерният анализ поставя акцент върху процеса на профилиране на получените групи от отрасли. Този процес включва характеристика на отделните групи чрез „измерване на статистическата значимост на разликите, при която груповата принадлежност на отделните обекти се използва като независима променлива.[30]" В зависимост от характеристиките на изходните данни в съвкупностите за изчисляване на значимостта на разликите са използвани различни статистически инструменти. Така например за една част от променливите значимостта на разликите е получена чрез сравнение на средните стойности на съвкупностите за всяка една група със средните стойности на общата съвкупност (данните за всички групи от отрасли).

При сравняване на средни стойности на съвкупностите са използвани следните статистически методи:

I. Метод. В изследването методът е приложен в следната проследователност:

Първо: Дефиниране на нулева (H0) и алтернативна (H1) хипотези.

Нулева хипотеза: Между $\overline{X_1}$ на променлива V в група C_n и между $\overline{X_2}$ на променлива V за всички данни от генералната съвкупност не съществува статистическа значима разлика, т.е

$$\mu_1 = \mu_2 \left(\mu_1 - \mu_2 = 0\right);$$

Алтернативна хипотеза: Между $\overline{X_1}$ на променлива V в група C_n и между $\overline{X_2}$ на променлива V за всички данни от генералната съвкупност съществува статистическа значима разлика, т.е. $\mu_1 > \mu_2$ (при дясностранна алтернативна хипотеза) и $\mu_1 < \mu_2$ (при лявостранна алтернативна хипотеза).

Второ: Определяне на теоретична стойност на риска за грешка. В зависимост от данните и променливите са използвани три стойности на α: Минимална: α = 0.01 (риск за грешка от 1%); Стандартна: α = 0.05 (риск за грешка от 5%) и Максимална: α = 0.1 (риск за грешка от 10%)

Трето: Избрани са критерии за изчисляване на емпиричната характеристика за проверка на хипотезите. Емпиричната характеристика е изчислена по уравнението[31] за статистически заключения между две средни величини при големи извадки ($n_1 \geq 30$, $n_2 \geq 30$) – уравнение[32] (8).

$$U' = n_1 n_2 + \frac{n_1 \ n_1 + 1}{2} - S_1,$$

където: μ_1, μ_2 - неизвестни средни величини в двете съвкупности (на дадена група и общо за всички групи);

n_1, n_2- брой на наблюденията в двете съвкупности;

X_1, X_2- оценка на средните величини на двете съвкупности;

S_1, S_2- оценка на стандартните

[30] Желев, С., „Маркетингови изследвания за маркетингови решения", София, И. Тракия-М, 2000, с. 179

[31] Манов Арсо, „Статистика със SPSS", София, И. Рикол – Б, 2000, с. 149 - 152

[32] Манов Арсо, „Статистика със SPSS", София, И. Рикол – Б, 2000, с. 150 - 151

отклонения на двете съвкупности.

Четвърто: Определена е критичната област. Критичните области в изследването са дясно-странни или ляво-странни, защото се разполага с информация за посоката им.

Пето: Определени са теоретичните характеристики на z.

Шесто: Сравнени са емпиричната с теоретичната характеристика.

Седмо: Взето е решение за приемане или отхвърляне на нулевата хипотеза. При едностранна критична област H0 е отхвърлена, когато |z| > |z (α)|, т.е. когато емпиричната характеристика е по-малка от теоретичната, определена с помощта на таблиците показващи функцията на стандартизираното нормално разпределение (площите под стандартизираната нормална крива) в интервала от (-z) до (z) и интервала (-∞) до (z).

Методът е приложим за групите, за които е изпълнено условието $n_1 \geq 30$, $n_2 \geq 30$. През 2000 г. на това условие отговорят Групи № 5 и № 6, а през 2006 г. Групи № 7 и № 8.

За преобладаваща част от средните на променливите за тези групи резултатите показват, че съществува статистически значима разлика спрямо средната на променливите от генерална съвкупност. Като цяло разликата в стойността между средните на променливите на групите, за които се приема нулевата хипотеза, спрямо средната на променливите за генералната съвкупност е много малка.

II. Метод. В случаите когато изходните данни не отговарят на изискването за приложимост на емпиричната характеристика представено чрез уравнение (8) е използван непараметричния тест на Ман-Уитни[33] при две извадки. Той се явява алтернатива на t-критерия [34] за формиране на статистически заключения за две средни величини при независими извадки. Този тест не изисква нормално разпределение и може да се прилага за метрирани и ординални данни. В изследването методът е приложен в следната проследователност:

Първо: Дефинирани са нулева (H0) и алтернативна хипотези (H1).

Нулевата хипотеза (H0) гласи: Между $\overline{X}_1$ на променлива V в група C_n и между $\overline{X}_2$ на променлива V за генералната съвкупност не съществува статистическа значима разлика, т.е

$$\mu_1 = \mu_2 \; (\mu_1 - \mu_2 = 0);$$

Алтернативната хипотеза (H1) гласи: Между $\overline{X}_1$ на променлива V в група C_n и между $\overline{X}_2$ на променлива V за генералната съвкупност съществува статистическа значима разлика, т.е. $\mu_1 > \mu_2$ (при дясно-странна алтернативна хипотеза), $\mu_1 < \mu_2$ (при ляво-странна алтернативна хипотеза) и $\mu_1 \neq \mu_2$ (при двустранна алтернативна хипотеза).

Второ: Определена е теоретична стойност на риска за грешка. Използвани са три равнища на стойността на α: Минимална: α = 0.01 (риск за грешка от 1%), Стандартна: α = 0.05 (риск за грешка от 5%) и Максимална: α = 0.1 (риск за грешка от 10%).

Трето: Избрани са критерии за изчисляване на емпиричната характеристика за проверка на хипотезите. Емпиричната характеристика за проверка на хипотезите е получена по уравнение (9).

$$U' = n_1 n_2 + \frac{n_1 \; n_1 + 1}{2} - S_1,$$

където: n_1, n_2 - брой на наблюденията в двете съвкупности (на дадена група и на генералната съвкупност); S_1, S_2 – сума на ранговете на двете съвкупности.

В случаите когато: $U = n_1 n_2 / 2$

емпиричната характеристика е изчислена по следния начин:

$$U' = n_1 n_2 - U,$$

Четвърто: Определени са критичните области. Емпиричните характеристики са изчислени с програмния продукт SPSS 17.0 при двустранна критична област.

Пето: Определени са теоретичната характеристика на U.

Шесто: Сравнени са емпиричната с теоретичната характеристика.

Седмо: Взето е решение за приемане или отхвърляне на нулевата хипотеза. За изчисление на емпиричната характеристика при двустранна критична област е използвана програмата SPSS 17.0, като H0 е отхвърлена, когато $\alpha_s < \alpha$, т.е. когато теоретичното равнище на значимост е по-голямо от равнището на значимост на

[33] Манов Арсо, „Статистика със SPSS", София, И. Рикол – Б, 2000, с. 377 - 379

[34] В настоящия случай z- критерия и t- критерия са неприложим, защото 1). $n_1 < 30$ и 2). разпределението в генералната съвкупност е различно от нормалното – виж Манов Арсо, „Статистика със SPSS", София, И. Рикол – Б, 2000, с. 136, Таблица 6.6

емпиричната характеристика U.

Резултатите показват че между средните на променливите за групите, при които е приложен теста (през 2000 г. това са Групи № 1, № 2, № 3 и № 4, а през 2006 г. това са Групи № 1, № 2, № 3, № 4, № 5 и № 6), от една страна и средните на променливите за всички групи за 2000 г. и 2006 г. съществува статистически значима разлика. Този извод е валиден за преобладаващата част от средните на променливите на групите. При останалите променливи за вярна е приета Нулевата хипотеза (Нулевата хипотеза е приета за тези средни на променливите, при които разликата в стойността спрямо средните на променливите за всички групи е много малка). Резултатите от приложението на двата метода потвърждават разликите в стойността между груповите средни и средните за променливите общо за всички отрасли от преработващата промишленост в България по области.

Изводи и резултати от приложението на клъстерния анализ

В статията клъстерният анализ е приложен съобразно описаните в литературата последователност и изисквания. Получените резултати за отраслите от преработващата промишленост по области в България показват съществуването на 6 валидни групи от отрасли през 2000 г. и 8 валидни групи от отрасли през 2006 г. Логическата интерпретация показва, че разликата между броя на групите през 2000 г. и 2006 г. се дължи на развитието на отраслите от преработващата промешленост по области, което се изразява в промените настъпили в класификационните критерии (променливи). Тук е мястото да се отбележи, че три от класификационните критерии, които са използвани в клъстерния анализ през 2000 г., са индентични с три от критериите използвани в клъстерния анализ през 2006 г., което показва валидността на логическата интерпретация на разликата в броя на групите. Останалите три класификационни критерия (променливи) са различни (външни) спрямо класификационните критерии използвани в клъстрният анализ през 2000 г. Според специалистите в областта „Валидизирането на клъстерите на база на външни променливи е един от най-надежните методи на валидизация. Неговата същност се състои в оценка на клъстерната валидност на базата на променливи, които не са участвали в процеса на формиране на клъстерите."[35] По нататък един от изследователите в този област казва, че „Финей и Муус (1979) посочват още по-силен подход на валидизация. … те установяват наличието на 8 клъстера, които валидизират на базата на 5 външни променливи. Интересното в случая е, че информацията за тези променливи е набавена шест мсеца след първоначалното изследване … . Осемте клъстера се оказват валидни по отношение на всичките пет външни променливи."[36]. В настоящото изследване е използван сходен подход на валидизация. През 2000 г. са използвани шест променливи, на база на които са получени 6 валидни групи от отрасли. През 2006 г. на базата на шест променливи (три от които са еднакви и три от които са различни (външни) спрямо променливите използвани в клъстерния анализ за данните от 2000 г.), са получени 8 валидни групи от отрасли, т.е с две повече от тези за 2000 г. Разликата в промяната в броя на получените групи се обяснява с развитието на отраслите от преработващата промишленост по области през изследвания период, с промяната на техните количествени и качествени характеристики. Отхвърлена е вероятността, че отраслите от преработващата промишленост по области в България не са претърпели развитие за периода 2000-2006 г. Получените на четвъртия етап от приложението на клъстерния анализ, дендограми показват, че увеличаването на групите се дължи на отраслите с високи стойности по класификационните променливи, т.е. на по-развитите отрасли. Първите четири групи през 2000 г. са съставени именно от такива отрасли. Получените дендограми показват още, че развитието на отраслите от тези четири групи през 2000 г. водят в крайна сметка до формирането на шест групи през 2006 г. (т.е. на две допълнителни групи от отрасли през 2006 г. спрямо 2000 г.). И за четирите групи през 2000 г. и за шесте групи през 2006 г. е характерно, че включват малък брой отрасли. Освен това

[35] Катранджиев Христо, „Сегментиране на телевизионната аудитория на основата на зрителските навици", дисертационен труд за присъждане на образователна и научна степен „Доктор", 2004 г., стр. 164

[36] Вж. Finney, J.W., R.H.Moos, Treatment and Outcome for Empirical Subtypes of Alcoholic Patients, Jornal of Comsulting and Clinical Psychology, 47, pp. 25-35

пространствените карти за двете години показват, че тези групи имат по-ниска хомогенност, а отраслите в тях са по-силно разпръснати в пространството. Логиката подсказава, че между развити отрасли е нормално да съществуват по-големи различия, а техният брой да е значително по-малък. Останалите две групи през 2000 г. и 2006 г. имат висока хомогенност, съставени са от много голям брой отрасли, между които съществуват много малки различия по класификационните променливи (в пространствените карти тези две групи са ясно обособени, а от отраслите в тях не са разпръснати). Логиката подсказва, че колкото по-слабо развити са отраслите, толкова по-малки са различията между тях.

Въз основа на всичко посочено до тук е прието, че броят на получените групи от отрасли от приложението на клъстерния анализ за 2000 г. и 2006 г. е валиден.

Груповите средни за 2006 г. от петия етап на клъстерния анализ позволяват да се направи обобщена характеристика на всяка една от групите от отрасли на преработващата промишленост по области в България (Таблица № 2).

Таблица № 2.

Групи	Характеристика на групите според груповите средни
Група № 4	Отрасли с много голямо значение * за промишлеността в об ластта и с много голям брой предприятия ** от малък мащаб ***. (предприятията имат среден брой наети **** с много ниска въоръженост на труда с ДМА*****)
Група № 1	Отрасли с голямо значение за промишлеността в областта и с малък брой предприятия от малък мащаб (ДМА). (предприятията имат малко наети с ниска въоръженост на труда с ДМА)
Група № 2	Отрасли с голямо значение за промишленоста в областта и с голям брой предприятия от малък мащаб (ДМА). (предприятията имат висок брой наети с много ниска въоръженост на труда с ДМА)
Група № 3	Отрасли със средно значение за промишлеността в областта и много голям брой предприятия от микро мащаб (ДМА). (предприятията имат много малко наети със средна въоръженост на труда с ДМА)
Група № 5	Отрасли със средно значение за промишлеността в областта и среден брой предприятия от висок мащаб (ДМА). (предприятията имат много висок брой наети с много висока въоръженост на труда с ДМА)
Група № 6	Отрасли със средно значение за промишлеността в областта и малък брой предприятия от среден мащаб (ДМА). (предприятията имат среден брой наети с много висока въоръженост на труда с ДМА)
Група № 7	Отрасли с малко значение за промешлеността в областта и малък брой предприятия от малък мащаб (ДМА). (предприятията имат среден брой наети с умерена въоръженост на труда с ДМА)
Група № 8	Отрасли с много малко значение за промишлеността в областта и много малък брой предприяитя от микро мащаб (ДМА). (предприятията имат много малко наети с много ниска въоръженост на труда с ДМА)

* Значението на отръслите за област та е определено според средният им дял в преработващата промишленост на същата област . За целта е използвана следнат а скала: 4% до 10% - малко значение; от 10% до 15% - средно значение; от 15% до 21% - голямо значение и над 21% много голямо значение.

** За класиране на отарслите в групите според броя на предприятията в областите е използвана следната скала: от 1 до 20 много малък брой; от 20 до 40 малък брой; от 40 до 70 среден брой; от 70 до 100 голям брой над 250 много голям брой.

*** За класиране на отарслите в групите според мащабът на предприятията е използвана следната скала: до 1500 хил. лв. ДМА – микро; от 1500 до 5000 хил. лв. ДМА - малък; от 5000 до 10000 хил. лв. ДМА – среден и над 10 000 хил. лв. ДМА - висок.

**** За класиране на отарслите в гру пите според броя на наетите в областите е използвана следната скала: до 30 наети – много малък брой; от 30 до 50 наети - малък брой; от 50 до 70 наети - среден брой; от 70 до 100 наети - висок брой; над 100 наети - много висок брой.

***** За класиране на отарслите в групите според въоръженоста на труда на 1 нает в предприятията е използвана следната скала: до 20 хил. лв. ДМА на 1 нает – изключително ниска; от 20 до 30 хил. лв. ДМА на 1 нает – ниска въоръженост на труда; от 30 до 60 хил. лв. ДМА на 1 нает – умерена; от 60 до 90 хил. лв ДМА на1 нает – висока; над 90 хил. лв. ДМА на 1 нает – много висока.

Източник: Собствени изчисления

Представената характеристиката на отраслите от преработващата промишленост през 2006 г. позволява разработката на специфични мерки на икономическата политика по групи отрасли. Когато номерата на групите се разположат по териториален и отраслов признак в Матрица № 1 се получава статусът на регионалната икономика на България по групи отрасли от преработващата промишленост. Така Матрица № 1 позволява координиране на мерките на икономическа политика не само по отраслов, но и по териториален признак. Нещо повече Матрица № 1 позволява да се изведат изискванията на развитето на територията към икономическата политика в областта на отраслите от преработващата промишленост. С други думи Матрица № 1 представлява основа за разработване на регионална икономическа политика в областта на отраслите от преработващата промишленост в България.

Матрица № 1

Групи от отрасли (от преработващата промишленост) по области в България през 2006 г.

NUTS 2	NUTS 3	DA	DB	DD	DE	DG	DH	DI	DJ	DK	DL	Dr.
СЗР	ВИДИН	1	1	8	8	8	6	7	8	8	7	8
	ВРАЦА	1	1	8	8	6	8	7	1	7	8	8
	МОНТАНА	1	1	8	8	8	8	7	7	6	7	7
	ЛОВЕЧ	1	7	6	8	7	8	7	8	7	7	1
	ПЛЕВЕН	1	2	8	7	8	8	7	7	7	8	8
СЦР	В. ТЪРНОВО	2	7	7	7	7	7	8	1	7	7	8
	ГАБРОВО	8	1	8	8	8	6	7	1	2	7	7
	РУСЕ	1	2	8	8	7	8	8	7	7	8	2
	РАЗГРАД	1	1	8	8	7	8	7	8	8	8	7
	СИЛИСТРА	1	7	6	8	8	8	8	8	7	7	8
СИР	ВАРНА	1	8	7	7	5	8	7	1	7	8	2
	ДОБРИЧ	1	2	8	8	8	8	8	8	8	8	1
	ТЪРГОВИЩЕ	1	1	8	8	8	8	6	8	8	7	8
	ШУМЕН	1	7	7	8	8	8	6	1	8	8	8
ЮЗР	БЛАГОЕВГРАД	7	4	7	8	8	8	8	8	8	7	2
	КЮСТЕНДИЛ	8	1	8	8	6	8	8	8	8	7	1
	ПЕРНИК	1	1	8	8	8	8	8	6	6	8	8
	СОФИЯ	3	3	8	3	8	7	8	3	3	3	3
	СОФИЯ-ГРАД	5	2	8	2	7	7	7	5	8	2	7
ЮЦР	КЪРДЖАЛИ	1	2	8	8	8	8	8	7	7	8	8
	ПАЗАРДЖИК	1	7	2	6	7	7	8	8	8	7	2
	ПЛОВДИВ	4	4	8	2	7	2	7	3	2	7	2
	СМОЛЯН	8	2	6	8	8	8	8	8	8	7	8
	ХАСКОВО	1	2	8	8	7	8	7	7	7	8	8
ЮИР	БУРГАС	2	7	6	8	8	7	7	1	8	8	5
	СЛИВЕН	1	6	7	8	8	8	8	8	7	7	8
	ЯМБОЛ	1	7	8	8	8	8	7	7	7	8	7
	С. ЗАГОРА	2	7	8	8	8	7	8	1	2	7	8

Източник: Собствени изчисления

Библиография

1. Goev, V., Statisticheska obrabotka i analiz na informatshiyata ot sotshiologicheski, marketingovi i politicheski izsledvaniya, Sofiya, UI Stopanstvo, 1996
2. Zhelev, S., Marketingovi izsledvaniya, Sofiya, UI Stopanstvo, 1999
3. Zhelev, S., Marketingovi izsledvaniya za marketingovi resheniya, Sofiya, I. Trakiya-M, 2000
4. Manov, A., Statistika sas SPSS, Sofiya, I. Rikol – B, 2000
5. Katrandzhiev, Hristo, Segmentirane na televizionnata auditoriya na osnovata na zritelskite navitshi, disertatshionen trud za prisuzhdane na obrazovatelnata i nauchna stepen "doktor", 2004
6. Krumov, Kalin, Evropeyskoto budeshte na regionalnata politika na Bulgariya, disertatshionen trud za prisuzhdane na obrazovatelnata i nauchna stepen "doktor", 2012
7. Aldenderfer M. and Blashfield R., "Cluster Analysis", USA, Sage Publication, 1984
8. Baxter J., Exploratory Multivariate Anaysis in Archaeology, Edinburgh, Edinburgh University Press, 1994
9. Blashfield R. K., The growth of cluster analysis: Tryon, Ward and Johnson, Multivariate Bechavioral Research, 1980
10. Brian S. Everitt, Sabine Landau, Morven Leese, "Cluster Analysis", London, "Arnold", 2001
11. Cox T. F., Cox A.A. M., "Multidimensional scaling", Second edition, Chapman & Hall, 2000
12. Duflou H., Maenhaut W., Application of principal component and cluster analysis to the study ot the distribution of minor and trace element in the normal human brain, Cheomometrics and Intelligent Laboratory Systems, 1990
13. Felsenstein D. and Portnov B., "Regional Disparities in small Countries", Berlin, Springer-Verlag, 2005
14. Fratianni Michaele and more, "Regional Economic Integration", "Elsevier JAI", Amsterdam, 2006
15. Kruskal, J.B., Wish M., "Multidimensional scaling", Quantative application in the Social Sciences, SAGE Publication Inc., 1978
16. Kruskal, Joseph Bernard, Multidimensional scaling by optimizing Goodness of fit to a Nonmetric Hipothesis, Psychometrika, 1964
17. Kruskal, Joseph Bernard, Nonmetric multidimensional scaling: A numerical method, Psychometrika, 1964
18. Milligan G., An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms, Psychometrika, 1980
19. Rombesburg C. H., "Cluster Analysis for Researcher", Belmont, Lifetime Learning Publications, 1984
20. Jeffrey A. Mills, Sourushe Zandvakili, "Statistical Inference via Botstrapping for Measures of Ineguality", University of Cincinati, Departament of Economics, 1995
21. Ward J., Herarchical grouping to optimizean objective function, Journal of the American statistical Association, 1963

INFORMATION SOURCES

NATIONAL STATISTICAL INSTITUTE

Nomenclatures and classifications

1. National classification of the economic activities, Version 2003 (NCEA-2003)
2. National classification of the products in respect of the economic activities, Version 2003 (NCPEA-2003)

Applied software products

1. SPSS 17
2. Clustan Graphics 6