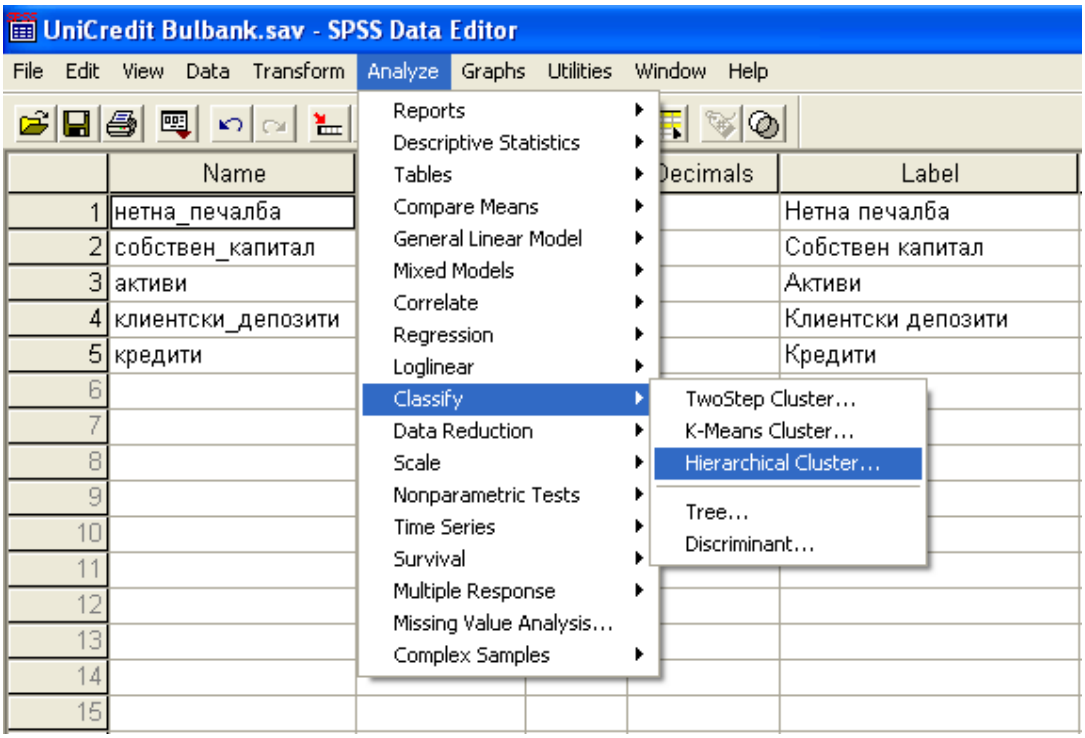
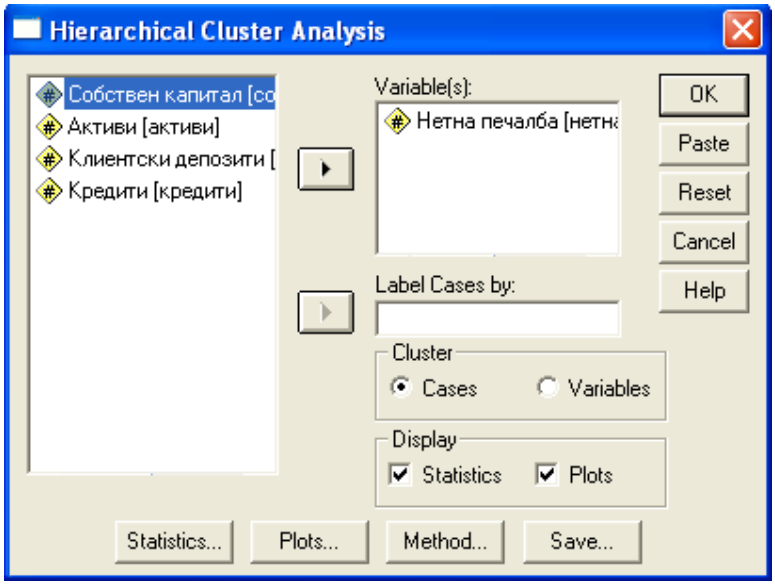
Клъстерен анализ с SPSS: йерархичен клъстерен анализ

Избира се от главното меню последователно **Analyze** → **Classify** →**Hierarchical Cluster**.



Фигура 1.

Появява се следния диалогов прозорец:



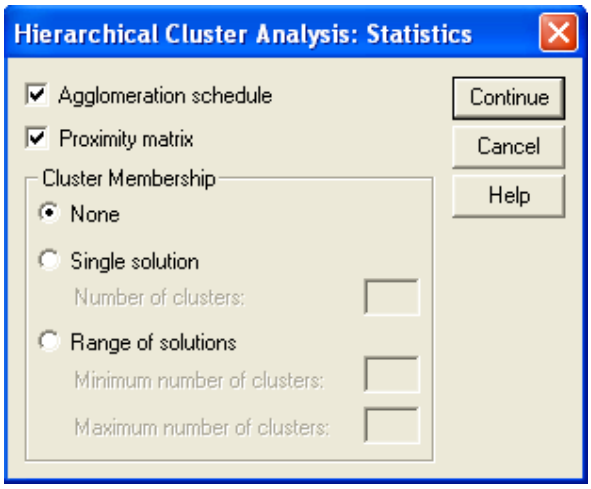
Фигура 2.

Маркират се една по една променливите, които ще се разглеждат и се изпращат в полето **Variables**. Действията по-нататък в значителна степен зависят от това, кой тип клъстеризация ще се избере тук. За тази цел в полето **Cluster** има избор между **Cases**, ако се извършва клъстеризация на обектите и **Variables** – ако е на променливите. По-нататък следва да се зададе начина за идентификация на обекта. Полето **Label Cases by** служи за въвеждане на стрингова променлива величина за маркиране на единиците. Важно е да отбележим, че тази променлива трябва да е от тип String (низ), за да може да се въведе в това поле.

Ако вместо Cases (обекти) в полето Cluster се установи Variables (променливи), то в списъка Variable(s) се изисква да укажем променливите, а полето Label Cases by ще остане празно. В полето **Display** по подразбиране са маркирани **Statistics** – за екранизиране на статистическите резултати от анализа и **Plots** – за екранизиране на графиките. И в двата случая не е необходимо да се махат отметките.

В долната част на диалоговия прозорец са разположени четири бутона, предназначени за въвеждане на допълнителни команди. Активизира се клавиша **Statistics** (фигура 2), който служи за определяне на статистическите

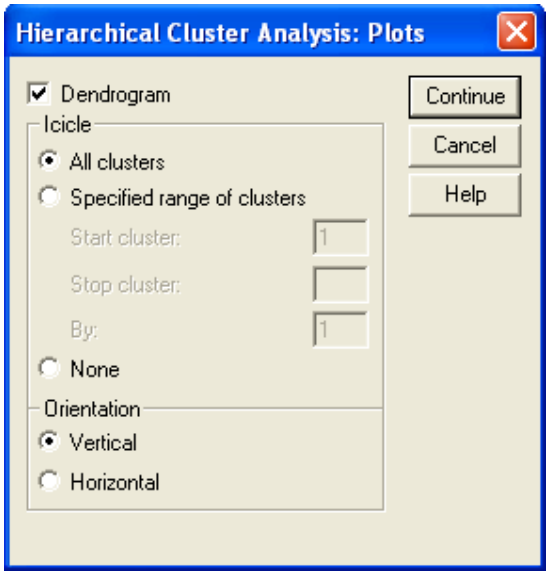
резултати, които да бъдат изведени на екрана. Тук може да се сложи отметка в полето на **Agglomeration schedule** (агломеративно включване) за показване на агломерационния график, както и в полето на **Proximity Matrix** за екранизиране на матрицата на близостта, която отразява информацията за разстоянията между обектите и клъстерите. В първия случай се визуализира последователността на обединяване на обектите в клъстери, като в началото всеки обект се възприема като отделен клъстер и след това започва окрупняване. По-долу в полето **Cluster Membership** (принадлежност към клъстер) може да се избере: **None** ако не се изисква показване на принадлежността на обектите към клъстерите, **Single solution** (единично решение) – като се укаже точният брой на клъстерите и **Range of solutions** (ранг на решенията) – като се определи диапазонът на желаните клъстери – от колко до колко клъстера искаме да получим. С Continue се продължава.



Фигура 3.

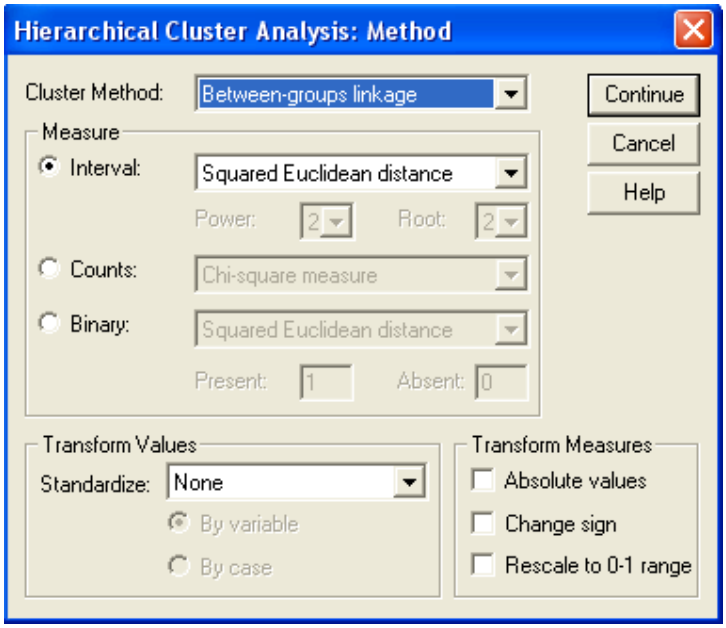
Когато се активира клавиша **Plots** може да се маркира **Dendrogram**, в случай че искаме графична визуализация на резултатите от йерархичната клъстеризация. Дендрограмата е граф-дърво, в което всеки възел отразява една стъпка от процеса на обединяване. Той носи и допълнителна информация за величината на разстоянието между двата клъстера в момента на обединение. Пунктираната хоризонтална линия на дендрограмата показва

рескалираното разстояние, при което са формирани клъстерите. В полето **Icicle** (висяща диаграма) с **All clusters** се определя диаграмата да обхване всички клъстери, със **Specified range of clusters** може да се уточни диапазона от клъстери, а с **None** се отказваме от висяща диаграма. В полето **Orientations** чрез превключване на радиобутоните може да се избере вида на диаграмата – дали да е вертикална или хоризонтална.



Фигура 4.

Когато се щракне на **Method** в прозореца **Hierarchical Cluster Analysis** се открива следния диалогов прозорец:



Фигура 5.

Тук най-напред с **Cluster Method** се определя клъстерния метод, който ще се ползва. При SPSS са възможни 7 метода:

- Метод на междугрупово свързване (Between-groups linkage)
- Метод на вътрешногрупово свързване (Within-groups linkage)
- Метод на най-близкия съсед (Nearest neighbor)
- Метод на най-отдалечения съсед (Furthest neighbor)
- Центроиден метод (Centroid clustering)
- Медианен метод (Median clustering)
- Метод на Вард (Ward's method)

Всеки от тези методи води до различна клъстеризация. Не може да се даде оценка коя е най-добрата, но все пак е препоръчително, ако се търсят клъстери във формата на „верига" да се използват методите на „междугрупово свързване" и „на най-близкия съсед". А методите на „вътрешногрупово свързване" и „на най-отдалечения съсед", когато се търсят клъстери във вид на „грозд". Важно е да се отбележи, че при верижния тип клъстери броят на обектите в различните клъстери е съществено различен за разлика от типа „грозд".

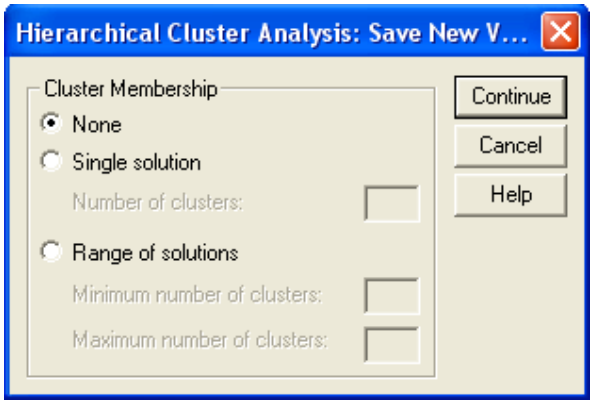
В полето **Measure** (мярка) трябва да се определи мярката за сходство, т.е. методът за измерване на подобие или различие между единиците. Тя се избира в зависимост от скалата на измерване на използваните променливи – дали е интервална (**Interval**), категорийна (**Counts**) или дихономна (**Binary**). С други думи дали ще се определят измерители за подобие и различие при метрирани, неметрирани или алтернативни променливи величини. За променливите с номинален тип SPSS предлага два метода Chi-square measure и Phi-square measure. Препоръчително е да се използва първия метод като по-разпространен. Дихономната скала включва такива променливи, отразяващи настъпване или ненастъпване на дадено събитие (купил/некупил, да/не и т.н.). Другите типове дихономни променливи (например мъж/жена) следва да се разглеждат и анализират като номинални. Най-често използвания метод за определяне на разстоянието на интервалните променливи е квадратичното евклидово разстояние, но все пак за дихономни променливи, където наблюдението представлява само две значения (например 1 и 0), даденият метод не е подходящ. Най-комплексния метод за измерване на разстояния, взимащ под внимание всички важни видове взаимодействия между две дихономни променливи се явява метода Lambda. При определяне на метода за бинарни променливи и необходимо в съответното поле да се укаже конкретното значение, което приема изследваната дихономна променлива: в полето Present – например кодировката на Да и в Absent – тази за Не. Аналогично е и за настъпване/ненастъпване на дадено събитие.

В полето **Transform Values** чрез **Standardize** трябва да се определи начина за стандартизиране на променливите величини. Различните методи бяха описани по-рано.

В полето **Transform Measures** чрез **Absolute Values** се елиминира посоката на връзката, когато се използва коефициентът на корелация при групиране на променливите величини. Посредством **Change Sign** се сменя знака пред измерителите на подобие, като ги превръща в измерители на

различие и обратно, а с **Rescale to 0-1 Range** се стандартизират измерителите на подобие и различие в интервала от 0 до 1, когато е необходимо.

С клавиша **Save** се записват резултатите от клъстеризацията, т.е. принадлежността на всеки обект към съответния клъстер, като отделна променлива във файла с данни. Тук може да се определя начина на съхраняване на номерата на клъстерите, към които принадлежат единиците. С **None** се отказваме от съхраняване на номерата на клъстерите, към които принадлежат единиците. Със **Single Solution** може да се съхранят номерата на клъстерите, към които принадлежат единиците, при предварително определен брой клъстери, а с **Range of Solutions** се съхраняват номерата на клъстерите, към които принадлежат единиците, при предварително определена поредица от клъстерни решения.



Фигура 6.

Да се върнем на примера с основните показатели на UniCredit Bulbank и да извършим йерархична клъстеризация посредством метода на средното свързване между групите. В Таблица 1 е представено резюме за случаите, т.е. налични, липсващи и общо стойности.

Таблица 1.

Case Processing Summary(a)

Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
7	100,0	0	,0	7	100,0

(a) Average Linkage (Between Groups)

В Таблица 2 (Proximity Matrix), която е получена директно със SPSS е представена матрицата, която съдържа квадратичните евклидови разстояния (Squared Euclidean Distance) с измерител на различие по данните от примера. Например квадратичното разстояние между първите две години се получава по следния начин: $s_{il}^{E2} = \sum_{j=1}^{p} (X_{ij} - X_{lj})^2$, $i, l = 1, ..., n$.

$s_{12} = (160{,}065 - 68{,}912)^2 + (602{,}776 - 490{,}479)^2 + (2\,559{,}476 - 2\,731{,}686)^2 +$
$(1692{,}270 - 2021{,}634)^2 + (316{,}380 - 362{,}353)^2 = 161\,169{,}931.$

Таблица 2.

Proximity Matrix

Case	Squared Euclidean Distance						
	1	2	3	4	5	6	7
1	,000	161169,931	230196,576	674488,593	3781328,705	3633840,451	9195794,008
2	161169,931	,000	42154,582	344379,088	2653215,196	2731459,825	7490302,184
3	230196,576	42154,582	,000	152904,699	2241708,704	2206177,741	6768568,041
4	674488,593	344379,088	152904,699	,000	1398061,171	1244224,157	5140974,232
5	3781328,705	2653215,196	2241708,704	1398061,171	,000	207861,100	1260287,862
6	3633840,451	2731459,825	2206177,741	1244224,157	207861,100	,000	1457084,166
7	9195794,008	7490302,184	6768568,041	5140974,232	1260287,862	1457084,166	,000

This is a dissimilarity matrix

Евклидовото разстояние се влияе в значителна степен от мярката и мащаба на отделните променливи величини. По-голяма тежест при изчисляването му се дава на променливата величина, която се изразява с по-големи числа.

Да разгледаме по-подробно процеса на йерархична клъстеризация по метода на средното свързване между групите, след като сме получили матрицата на разстоянията.

	2000	2001	2002	2003	2004	2005	2006
2000	0,000	161 169,931	230 196,576	674 488,593	3 781 328,705	3 633 840,451	9 195 794,008
2001	161 169,931	0,000	42 154,582	344 379,088	2 653 215,196	2 731 459,825	7 490 302,184
2002	230 196,576	42 154,582	0,000	152 904,699	2 241 708,704	2 206 177,741	6 768 568,041
2003	674 488,593	344 379,088	152 904,699	0,000	1 398 061,171	1 244 224,157	5 140 974,232

2004	3 781 328,705	2 653 215,196	2 241 708,704	1 398 061,171	0,000	207 861,100	1 260 287,862
2005	3 633 840,451	2 731 459,825	2 206 177,741	1 244 224,157	207 861,100	0,000	1 457 084,166
2006	9 195 794,008	7 490 302,184	6 768 568,041	5 140 974,232	1 260 287,862	1 457 084,166	0,000

На първия етап от клъстеризацията се обединяват втората и третата година, защото разстоянието между тях е най-малко s_{23} = 42 154,582. Размерността на матрицата с разстоянията се редуцира с единица и има следните елементи:

	2000	2001, 2002	2003	2004	2005	2006
2000	0,000	391 366,507	674 488,593	3 781 328,705	3 633 840,451	9 195 794,008
2001,2002	391 366,507	0,000	497 283,787	4 894 923,900	4 937 637,566	14 258 870,226
2003	674 488,593	497 283,787	0,000	1 398 061,171	1 244 224,157	5140 974,232
2004	3 781 328,705	4 894 923,900	1 398 061,171	0,000	207 861,100	1 260 287,862
2005	3 633 840,451	4 937 637,566	1 244 224,157	207 861,100	0,000	1 457 084,166
2006	9 195 794,008	14 258 870,226	5 140 974,232	1 260 287,862	1 457 084,166	0,000

На следващия етап се обединяват първия и втория клъстер (2000 и 2001, 2002) защото съгласно теорията се получава най-малко средно разстояние: s_{12} = 391 366,507/2 = 195 683,253. Размерността на матрицата с разстоянията отново се редуцира с единица:

	2000, 2001, 2002	2003	2004	2005	2006
2000,2001,2002	0,000	1 171 772,381	8 676 252,605	8 571 478,016	23 454 664,234
2003	1 171 772,381	0,000	1 398 061,171	1 244 224,157	5 140 974,232
2004	8 676 252,605	1 398 061,171	0,000	207 861,100	1 260 287,862
2005	8 571 478,016	1 244 224,157	207 861,100	0,000	1 457 084,166
2006	23 454 664,234	5 140 974,232	1 260 287,862	1 457 084,166	0,000

На третия етап се обединяват третия и четвъртия клъстер (2004 и 2005), където: s_{34} = 207 861,100.

	2000,2001,2002	2003	2004,2005	2006
2000,2001,2002	0,000	1 171 772,381	17 247 730,621	23 454 664,234
2003	1 171 772,381	0,000	2 642 285,328	5 140 974,232
2004,2005	17 247 730,621	2 642 285,328	0,000	2 717 372,028
2006	23 454 664,234	5 140 974,232	2 717 372,028	0,000

При следващия етап обединяваме първи и втори клъстер (2000,2001,2002 и 2003 година). При този случай s_{12} = 1 171 772,381/3 = 390 590,794.

	2000,2001,2002,2003	2004,2005	2006
2000,2001,2002,2003	0,000	19 890 015,949	28 595 638,465
2004,2005	19 890 015,949	0,000	2 717 372,028
2006	28 595 638,465	2 717 372,028	0,000

На петия етап обединяваме втория и третия клъстер, т.е. 2004,2005 и 2006, където средното разстояние е най-малко: s_{23} = 2 717 372,028/2 = 1 358 686,014.

	2000, 2001, 2002, 2003	2004, 2005, 2006
2000, 2001, 2002, 2003	0,000	48 485 654,415
2004,2005,2006	48 485 654,415	0,000

На последния етап обединяваме останалите два клъстера, където средното разстояние е s_{12} = 48 485 654,415/12= 4 040 471,201.

Резултатите от различните етапи на йерархичната клъстеризация в SPSS се обобщават и извеждат в таблица, която се нарича **агломеративна схема** (**Agglomeration Schedule**). В тази таблица можем да видим и колона с изчислените до тук средни разстояния. В случая за измерител на подобие е използвано квадратичното евклидово разстояние.

Таблица 3.
Агломеративна схема

Agglomeration Schedule

	Cluster Combined			Stage Cluster First Appears		
Stage	Cluster 1	Cluster 2	Coefficients	Cluster 1	Cluster 2	Next Stage
1	2	3	42154,582	0	0	2
2	1	2	195683,253	0	1	4
3	5	6	207861,100	0	0	5
4	1	4	390590,794	2	0	6
5	5	7	1358686,014	3	0	6
6	1	5	4040471,201	4	5	0

При агломеративната схема в първата колона **Stage** са посочени номерата на отделните етапи. Като при последния етап са обединени всички изследвани обекти в един клъстер. В общия случай те са *n*-1. В колоните с общо заглавие **Cluster Combined** са посочени номерата на клъстерите, които се обединяват на отделните етапи. Например при първия етап са обединени втория и третия клъстер. В колоната **Coefficients** са дадени осреднените разстояния (от теорията), при които се обединяват клъстерите. Като тези коефициенти зависят от избрания метод за формиране на клъстера. Показателите в тази колона могат да се използват за приблизителна оценка на степента на еднородност на клъстерите, които се формират на всеки етап. Големите коефициенти (при измерители на различие) или малките коефиценти (при измерители на подобие) означават, че клъстерът е относително еднороден и съдържа в себе си единици, които са близки в значителна степен помежду си. Коефициентите в тази колона могат да се използват и за приблизителна ориентация относно броя на клъстерите, които трябва да бъдат профилирани от практическа гледна точка. За тази цел може да се проучи етапът, в който се забелязва рязка промяна в коефициентите. В колоните със заглавие **Stage Cluster First Appears** са посочени етапите, в които съответните клъстери са се появили за първи път, а в колоната **Next Stage** е изведен номерът на етапа, в който съответният клъстер ще се появи следващия път при обединение с друг клъстер. Например на първия етап при обединението на втория и третия клъстер се създава нов, на който присвояваме номер 2. Създаденият клъстер 2 се обединява с клъстер 1 на втория етап и т.н.

Резултатите от различните етапи на йерархичната клъстеризация могат да се илюстрират и с т.нар. **висяща диаграма** (**Icicle Plot**). Тя може да бъде вертикална или хоризонтална. На фигура 7 е представена вертикалната висяща диаграма на резултатите от клъстеризацията на годините на развитие на UniCredit Bulbank.

Vertical Icicle

Number of clusters	Case 7		6		5		4		3		2		1
1	X	X	X	X	X	X	X	X	X	X	X	X	X
2	X	X	X	X	X		X	X	X	X	X	X	X
3	X		X	X	X		X	X	X	X	X	X	X
4	X		X	X	X		X		X	X	X	X	X
5	X		X		X		X		X	X	X	X	X
6	X		X		X		X		X	X	X		X

фигура 7.

Всеки ред във вертикалната висяща диаграма (**Vertical Icicle**) съответства на броя на възможните клъстери. В общия случай те са *n*-1. За всяка единица съответства отделна колона, която е запълнена със знака X до последния ред. Между отделните колони, съответстващи на единиците, има други колони, които са запълнени със знака X. Диаграмата се разглежда отдолу нагоре. Например при 6 клъстера са обединени 2001 и 2002 година – колоната между тях е запълнена до последния ред, т.е. има 6 знака X. При 5 клъстера към 2001 и 2002 се присъединява и 2000 година. Колоната между 2000 и 2001 има пет знака X и т.н.

При относително големи съвкупности вертикалната висяща диаграма може да се окаже в недостатъчна степен информативна. В този случай може да се използва хоризонтална висяща диаграма или да се представят графично само част от клъстерите.

За графична визуализация на резултатите от йерархичната клъстеризация може да се използва и т.нар. **дендрограма** (dendrogram).

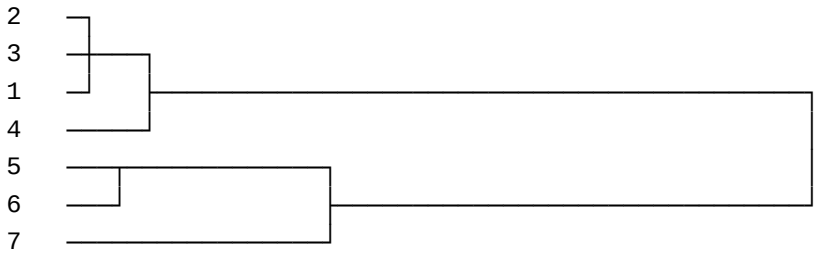
Пунктираната хоризонтална линия на дендрограмата, представена на фигура 8, показва рескалираното разстояние, при което са формирани клъстерите. Най-малкото разстояние – в случая 42 154,582 отговаря на 1, а най-голямото – 4 040 471,201 на 25. Дендрограмата позволява да се формулират следните резултати:

```
* * * * * * H I E R A R C H I C A L  C L U S T E R  A N A L Y S I S * * * * * *

 Dendrogram using Average Linkage (Between Groups)

                     Rescaled Distance Cluster Combine

  C A S E    0         5        10        15        20        25
  Label  Num  +---------+---------+---------+---------+---------+
```



Фигура 8.

- 2000, 2001 и 2002 година са обединени в общ клъстер при сравнително малко разстояние, т.е. клъстерът е относително еднороден;
- 2003 година образува отделен клъстер, който в последствие се присъединява към клъстера на 2000, 2001 и 2002 с относително еднакви показатели и почти двойно кредити;
- 2004 и 2005 година формират отделен клъстер и се обединяват със собствения клъстер на 2006, които са на значително разстояние от останалите. Те са с относително по-високи стойности на всички основни показатели, особено за 2006 година.

Забелязва се, че резултатите получени от йерархичния клъстерен анализ съвпадат с тези получени от клъстерния анализ с K-средни величини при промяна на центровете на клъстерите след присъединяването на всеки обект към даден клъстер.

Съставил: Десислава Войникова

Литература:

1. Манов, А. Б., Статистика със SPSS, изд. Тракия, София, 2001.
2. Гоев, В. Д. , Статистическа обработка и анализ на информацията от социологически, маркетингови и политически изследвания със SPSS, УНСС, София, 1996.