New Forms for Business Intelligence

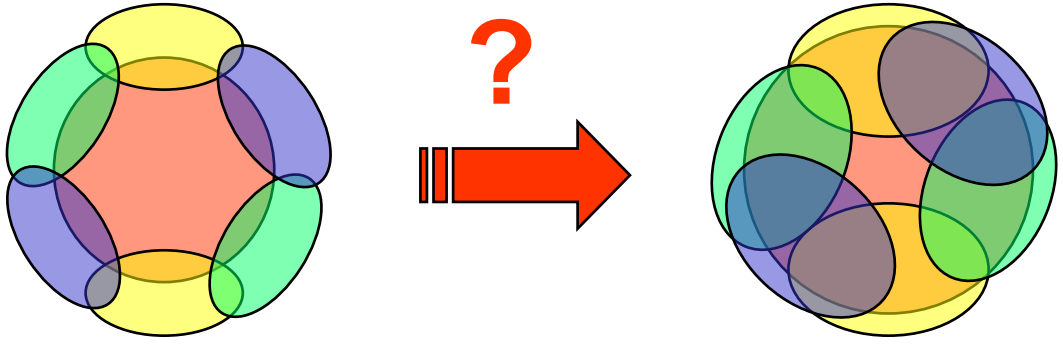
A concrete approach to the use of data to improve and quicken business decisions

Agenda

Module 4: Demand Segmentation

- Introduction
- The Multidensional Classfication Issue
- The Behavioral Segmentation
- The Analytical Process

Business Intelligence & Customer Profiling



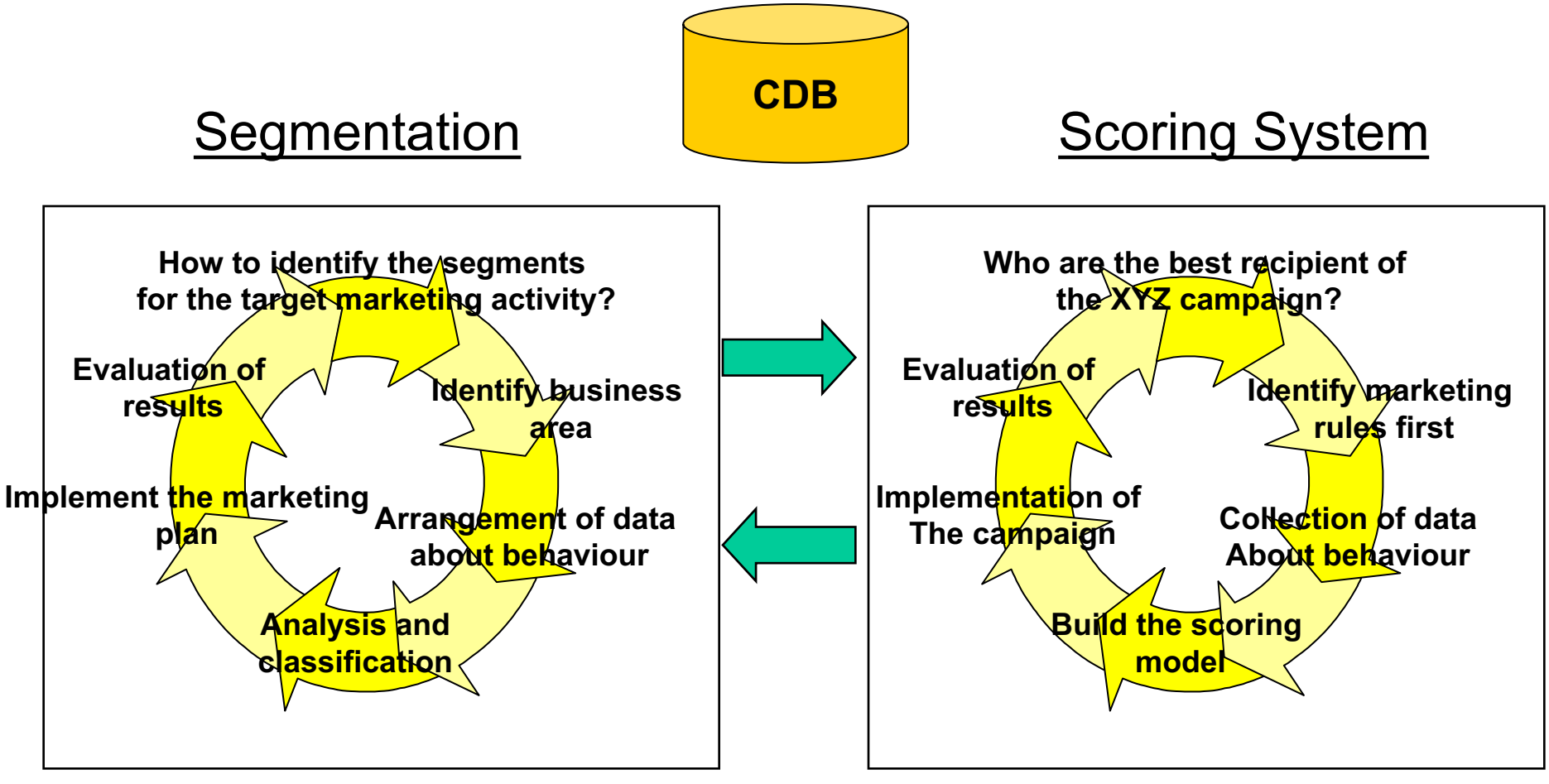
Identification of ‘business’ segments for planning the operations

➔ Segmentation ←

Identification of most likelihood clients to accept a cross/up selling offer

➔ Propensity models←

Goals of a CP project



The Active segments

HNFHC Active Customer Segments	% of AC	Brief description
Active Segments:		
"A1 - Young consumers"	**24,0%**	Almost young, significant propensity towards traditional products; small amount frequent withdrawal transactions. Low contribution amount and the lowest prodtype density.
"A2 - Young warriors"	**12,6%**	Young customers, extremely high percentage of Loans and Stocks. High propensity towards innovative products and accordingly high-risk propensity.
"A3 - High potential"	**13,7%**	Lowest contribution amount and high deposit balance (2° best). Predominantly low risk product type (time deposits and insurances). High percentage of exclusive customers and average low prodtype density. They make more deposit than withdrawal transactions, significant amounts.
"A4 - Frequent flyers"	**3,1%**	Heavy consumers, spending very frequently a significant amount. Relatively high % of Credit Card holders and ATM. Good % of investment products (Stocks, Funds, Futures). They provide an important contribution amount to the bank (2° best). Good prodtype holding profile.
"A5 - Got-it-all"	**16,9%**	Highest prodtype density with a percentage of exclusive customers close to zero. Frequency of transactions is high, but - given their rich portfolio - they show high dormant rates (especially for credit cards). Good average contribution amount.
"A6 - Abandoned hounds"	**10,7%**	Eldest active customers with an extremely long relationship with the company. Most of them have not been opening any new accounts since at least 10 years. Even if they show the highest deposit dormant rate within active customers, they have to be considered loyal and valuable customers
"A7 - High Rollers"	**2,0%**	Simply the best customers for HN: contribution amount much higher than the other segments, relevant percentage of aggressive and value added products. Mostly adult with a 10 years old relationship with the bank.
"A8 - Conventional savers"	**17,0%**	Quite young, significant propensity towards traditional products, small transactions to deposit money or pay car insurance premiums.
Total active customers	***100,0%***	

Agenda

Module 4: Demand Segmentation

- Introduction
- The Multidensional Classfication Issue
- The Behavioral Segmentation
- The Analytical Process

YXY Newspaper

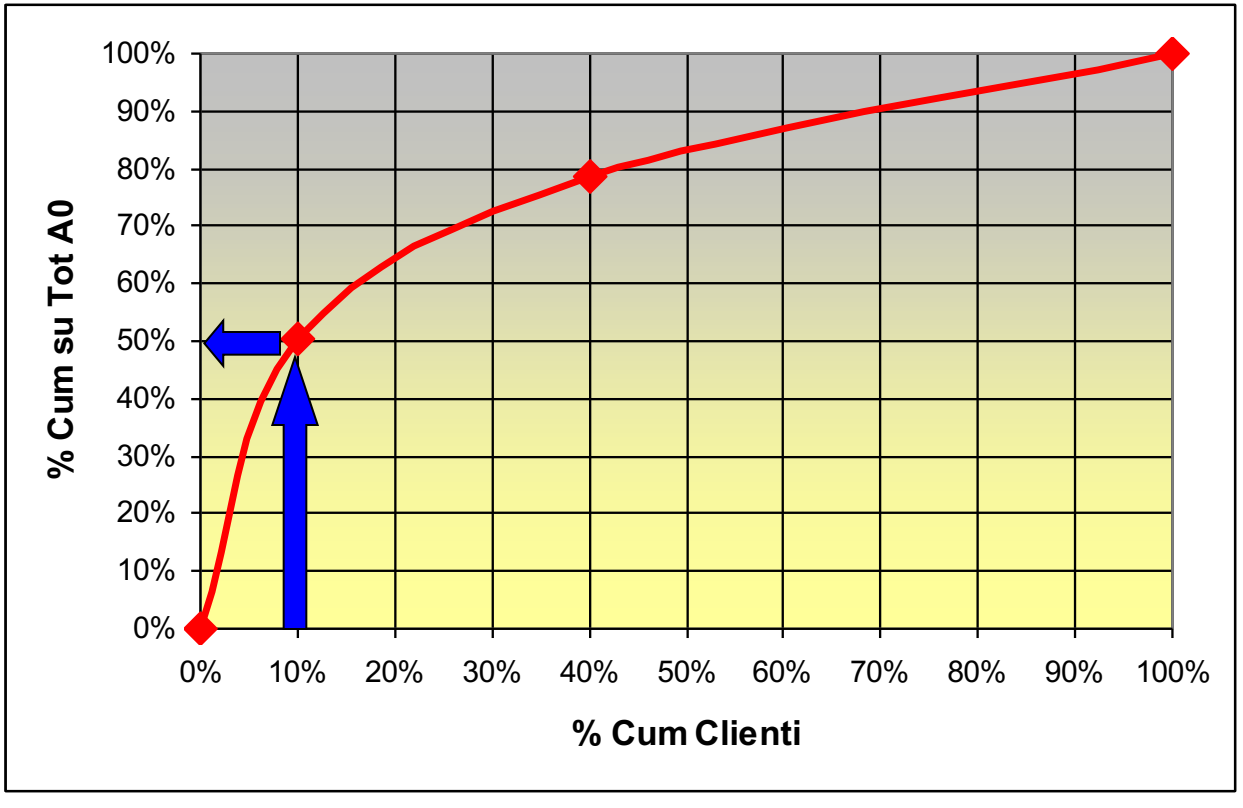
Active subscribers to YXY:
50.650
(52.837 active subscriptions)

YXY		
Average amount A0	**Average amount A1**	**Average amount A2**
182,97	183,40	149,01

Whole product portfolio		
Average amount A0	**Average amount A1**	**Average amount A2**
449,74	436,27	360,57

YXY Newspaper

Slot	% Clients	% Cum Clients	% on Tot A0	% Cum on Tot A0
1	10%	10%	50,3%	50,3%
2	30%	40%	28,2%	78,5%
3	60%	100%	21,5%	100,0%



YXY Newspaper

Slot	Average amount A0	% Imp. PER	% Imp. QUO	% Imp. OS	% Imp. BD
1	2.263,86	46,3%	9,1%	9,7%	34,9%
2	405,94	87,8%	2,8%	8,8%	0,6%
3	164,53	96,9%	0,0%	3,0%	0,1%
Totale Clienti	**449,74**	**87,8%**	**2,1%**	**5,8%**	**4,3%**

Slot	Seniority in years	Recency in months	Contacts for attempted sale	Contacts for solicitude at the renewal of GD
1	7,1	3,3	7,4	6,2
2	6,7	4,3	7,1	6,1
3	5,7	7,2	6,2	6,8
Total Clients	**6,1**	**5,9**	**6,6**	**6,5**

Multidimensional classification

- Use of elementary variables to describe the portfolio
- Dimensional cardinality problem:
  - Variables number
  - Modalities number
- Build summary indicators about behaviour
- Do not overlook different behavioural dimensions
- Behavioural segmentation

Agenda

Module 4: Demand Segmentation

- Introduction
- The Multidensional Classfication Issue
- The Behavioral Segmentation
- The Analytical Process

Behavioural segmentation

Segmentation consists of:

- **Identify groups** in which to segment customers with multivariate analysis techniques;

- **Classify every single person** as a memeber of a clients group;

- **Describe the groups** identified with their behavioural and social-demographic features;

- **Crate an assignment rule** to the groups identified, applicable over the time.

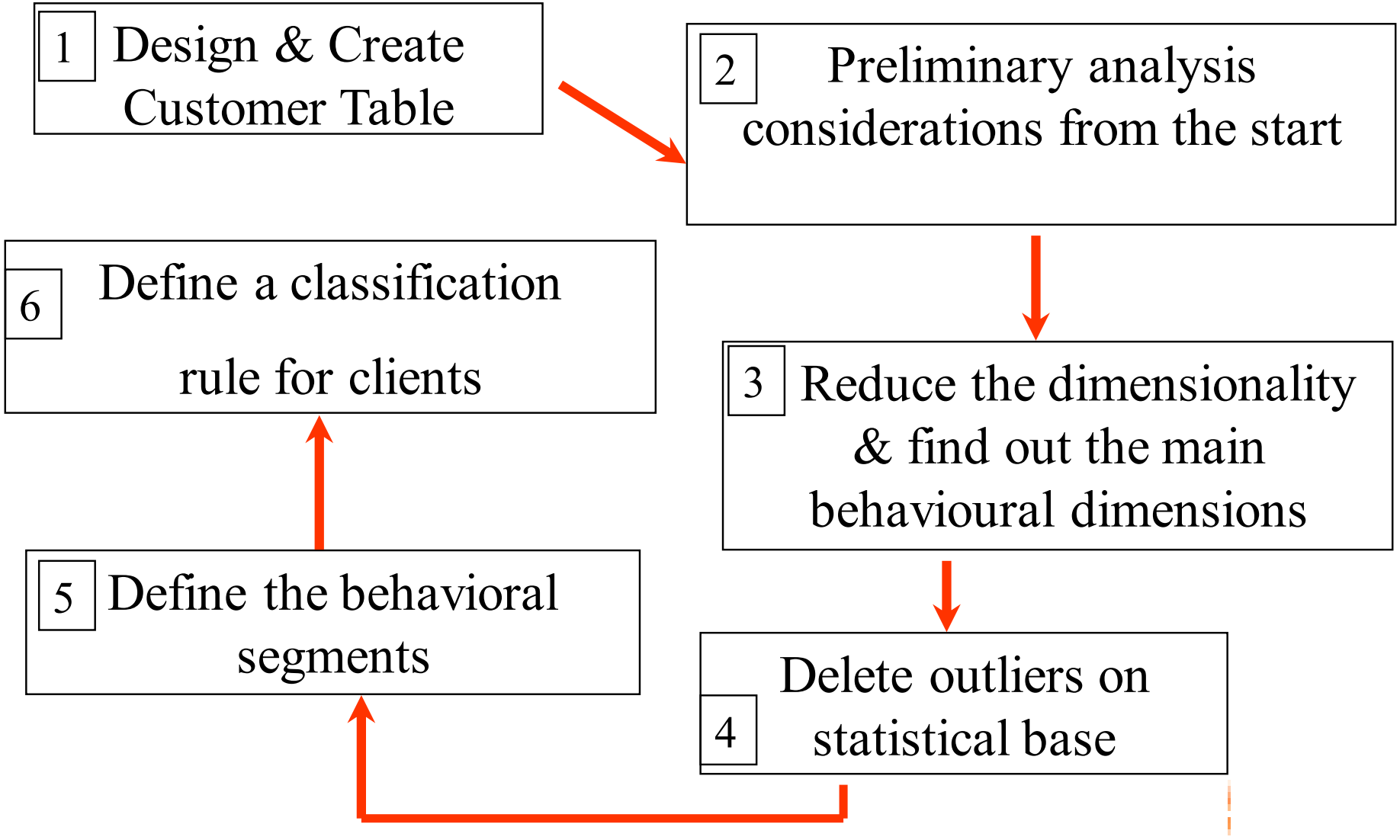
Behavioural segmentation

The main areas where the segmentation is used are:

- product development and assessment of potential
- commercial plan definition
- monitoring & directional reporting

- cross-selling activity;
- operative CRM activity

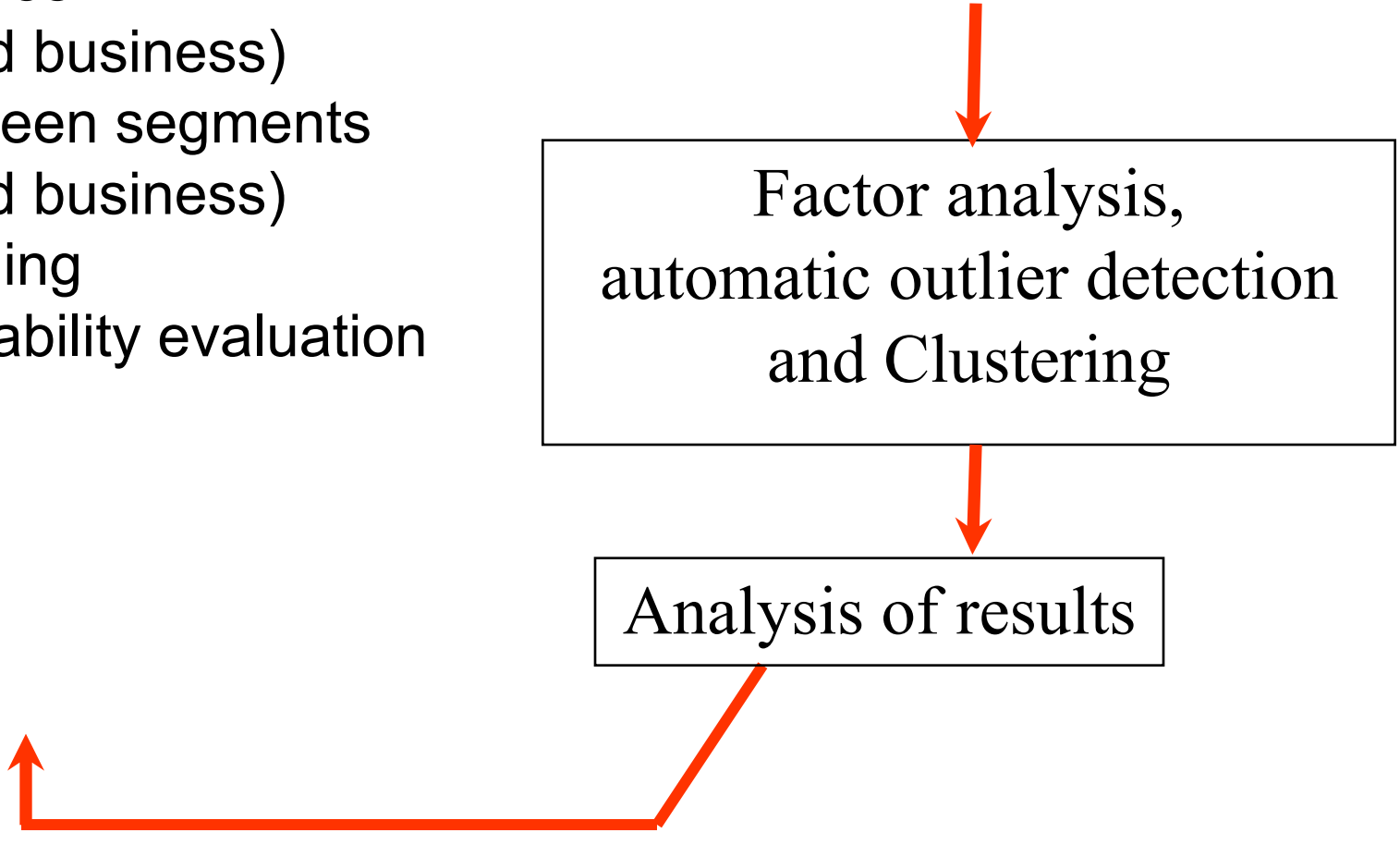
Analysis path



Analysis path - 1

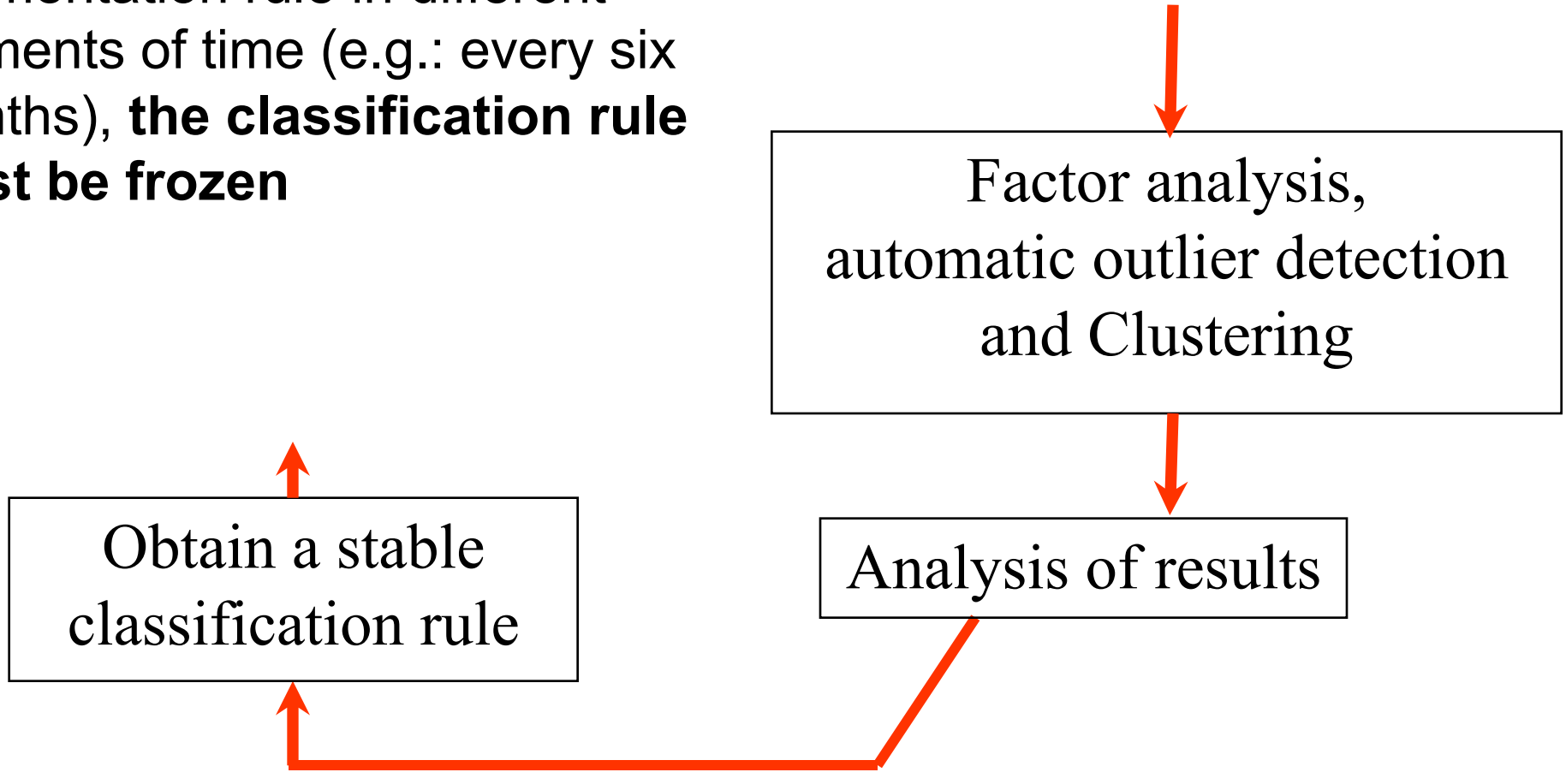
Analysis of results:

-% of customers in different segments
- segment profiles
  (analytical and business)
- distance between segments
  (analytical and business)
- business naming
- robustness/stability evaluation

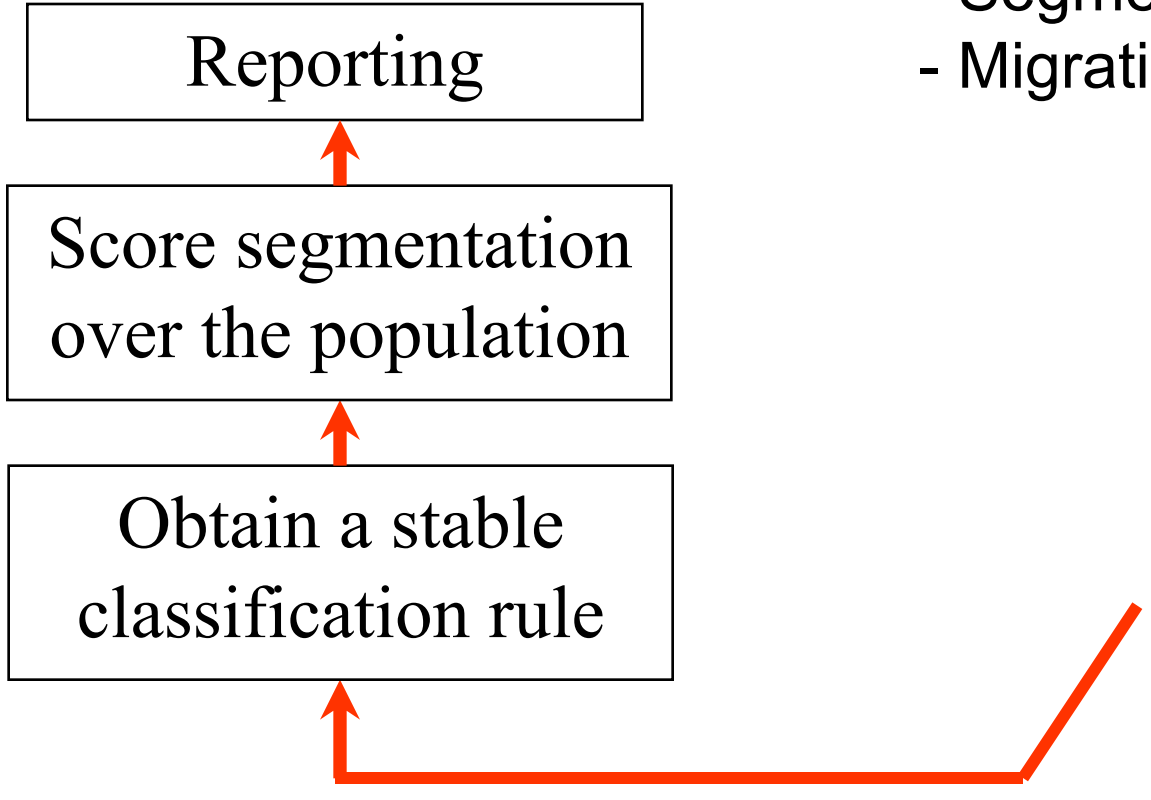


Analysis path - 2

Stable classification rule:
to be able to re-apply the same segmentation rule in different moments of time (e.g.: every six months), **the classification rule must be frozen**



Analysis path - 3



Reporting:

- Segment distribution
- Profiling maps
- Segment profile indexes
- Migration matrix

Segmentation project at HNFHC

Final Release Executive Summary

Prepared by *Guido Cuzzocrea, Salvina Piccarreta (Nunatac)*
for *Hua Nan Financial Holdings Co., Ltd.*

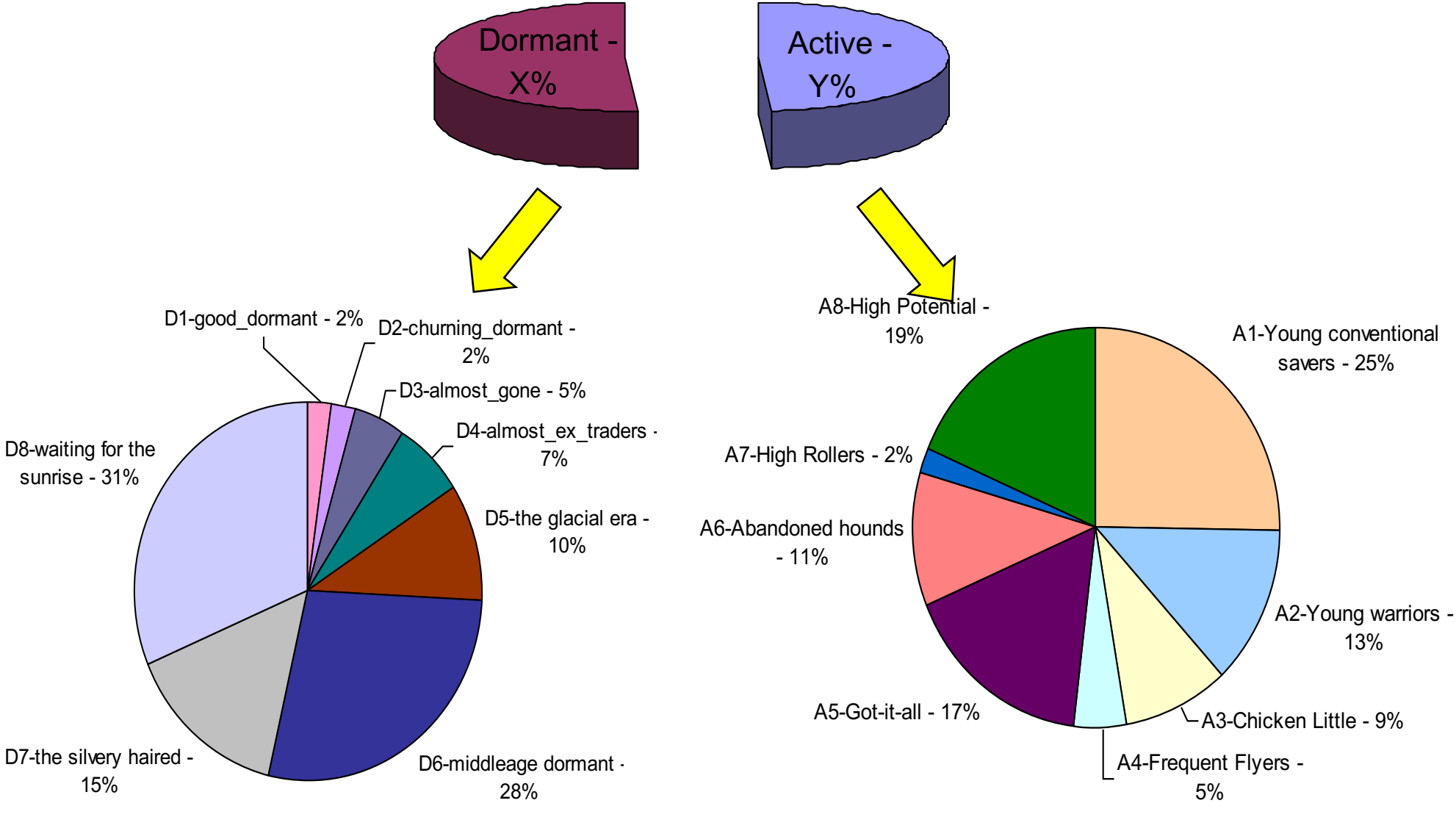
sas e-Intelligence

Segmentation: 1° step – active clients/sleeper

To avoid insignificant results, a classification rules has been previously applied to the CB, defined using business criteria, that identifies two macro segments: active clients vs. sleeper

Segmentation: 2° step – Clusters

Whole Customer Base of HNFHC ➔ XXX clients identified

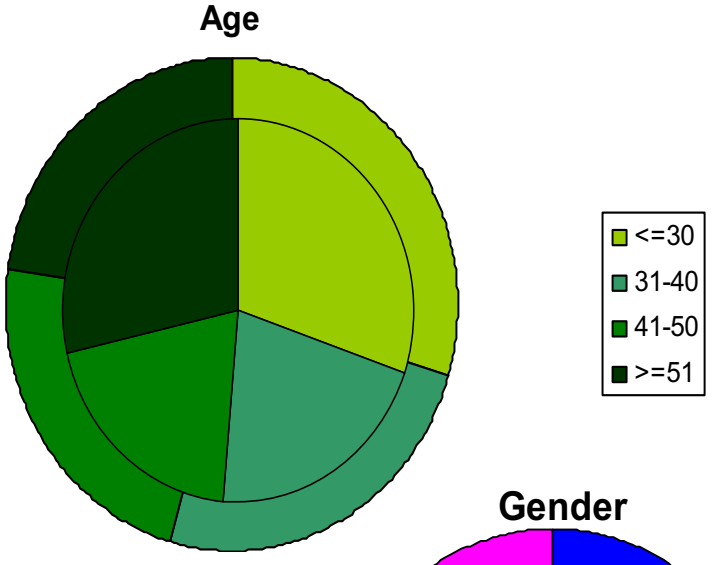


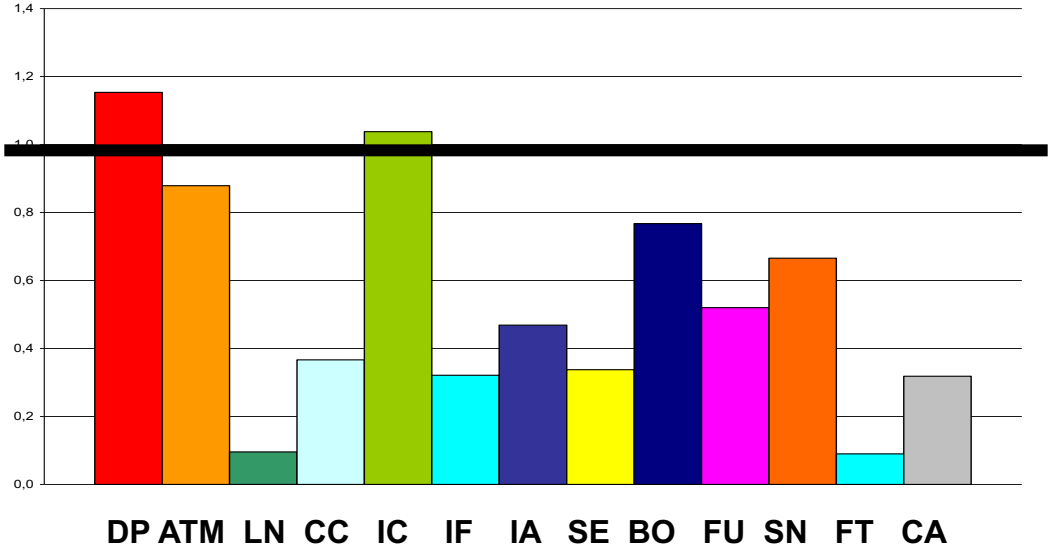
The Active segments

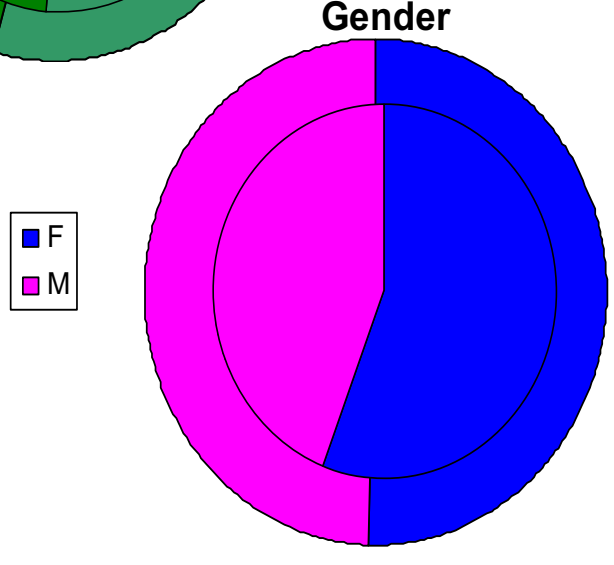
HNFHC Active Customer Segments	% of AC	Brief description
Active Segments:		
"A1 - Young consumers"	**24,0%**	Almost young, significant propensity towards traditional products; small amount frequent withdrawal transactions. Low contribution amount and the lowest prodtype density.
"A2 - Young warriors"	**12,6%**	Young customers, extremely high percentage of Loans and Stocks. High propensity towards innovative products and accordingly high-risk propensity.
"A3 - High potential"	**13,7%**	Lowest contribution amount and high deposit balance (2° best). Predominantly low risk product type (time deposits and insurances). High percentage of exclusive customers and average low prodtype density. They make more deposit than withdrawal transactions, significant amounts.
"A4 - Frequent flyers"	**3,1%**	Heavy consumers, spending very frequently a significant amount. Relatively high % of Credit Card holders and ATM. Good % of investment products (Stocks, Funds, Futures). They provide an important contribution amount to the bank (2° best). Good prodtype holding profile.
"A5 - Got-it-all"	**16,9%**	Highest prodtype density with a percentage of exclusive customers close to zero. Frequency of transactions is high, but - given their rich portfolio - they show high dormant rates (especially for credit cards). Good average contribution amount.
"A6 - Abandoned hounds"	**10,7%**	Eldest active customers with an extremely long relationship with the company. Most of them have not been opening any new accounts since at least 10 years. Even if they show the highest deposit dormant rate within active customers, they have to be considered loyal and valuable customers
"A7 - High Rollers"	**2,0%**	Simply the best customers for HN: contribution amount much higher than the other segments, relevant percentage of aggressive and value added products. Mostly adult with a 10 years old relationship with the bank.
"A8 - Conventional savers"	**17,0%**	Quite young, significant propensity towards traditional products, small transactions to deposit money or pay car insurance premiums.
Total active customers	***100,0%***	

A8: High potential (19%)

Low profitability. High deposit balance (2° higher). Low product density, high percentage of exclusives clients. Few transaction (of relevant amount).



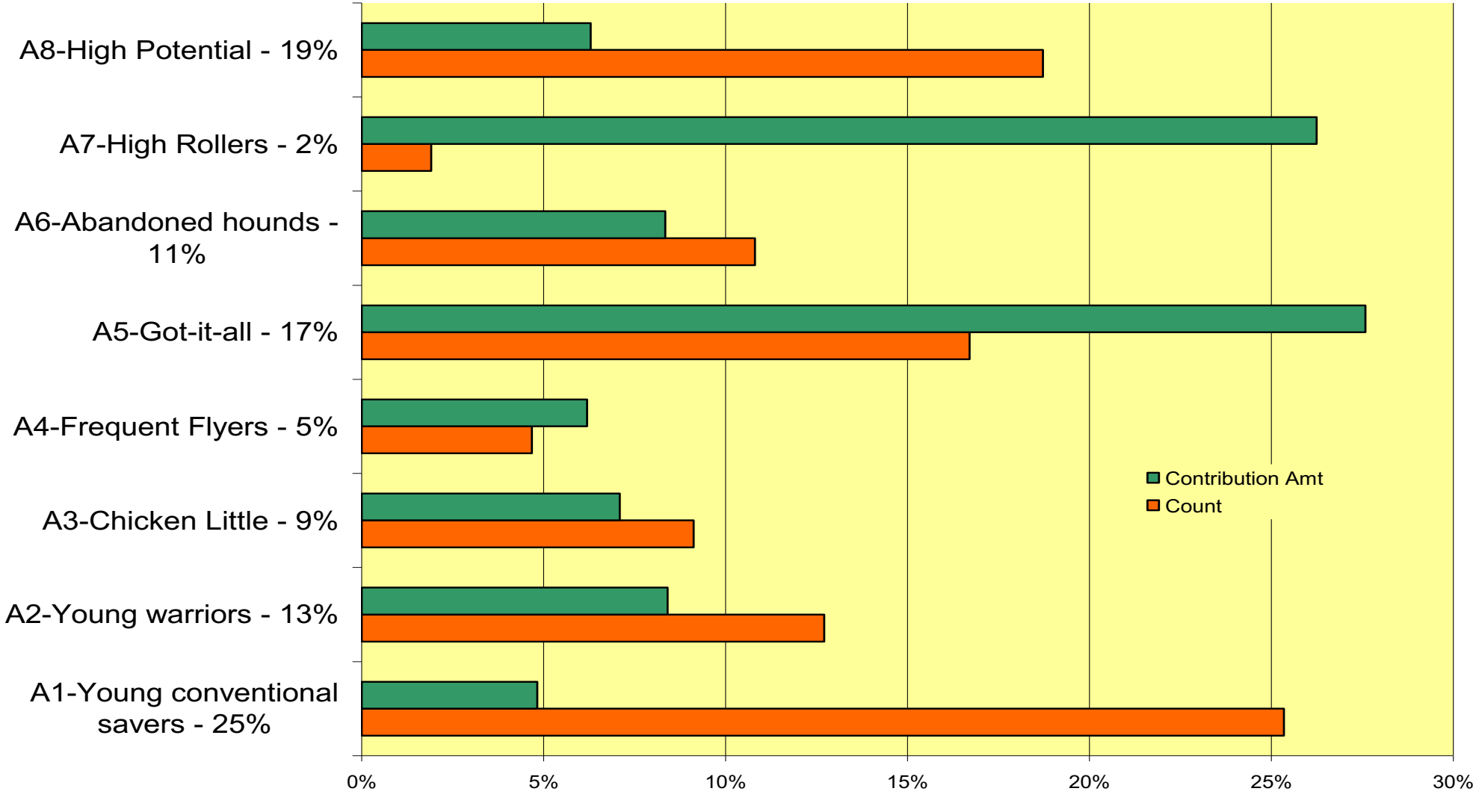




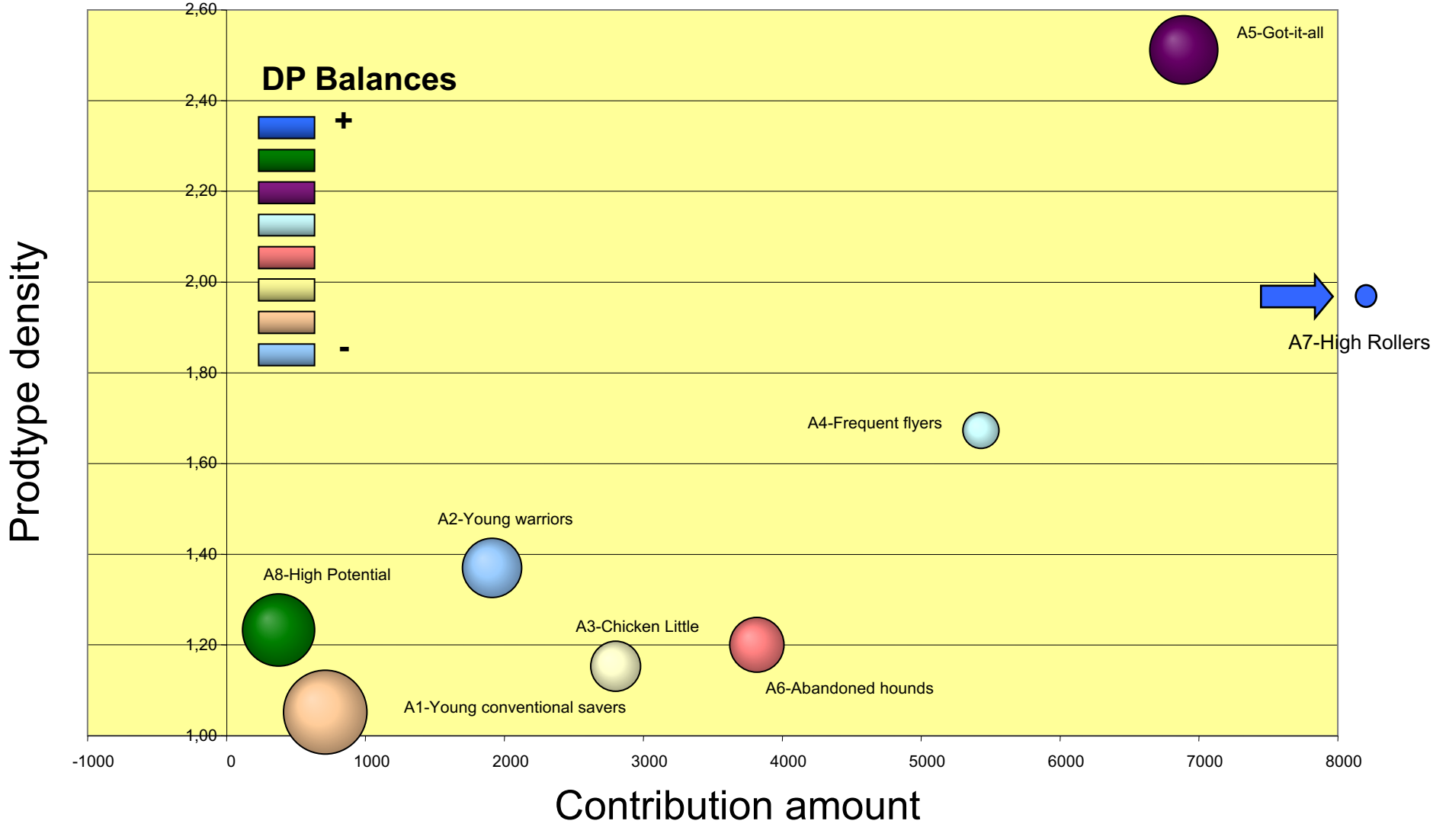
	A8	Best 10%	T. Active
Prodtype Density	1.23	2.00	1.44
% Exclusive	77%		66%
Contributed Amt	375	2,736	3,667
Balance DP	544,656	1,613,001	297,285
Tenure (year)	7.34		7.00
Recency (year)	-1.51		-2.74
Ratio Open Last Year	▲		◄►

Active segments(1/2)

% of clients vs. % of contributed amount



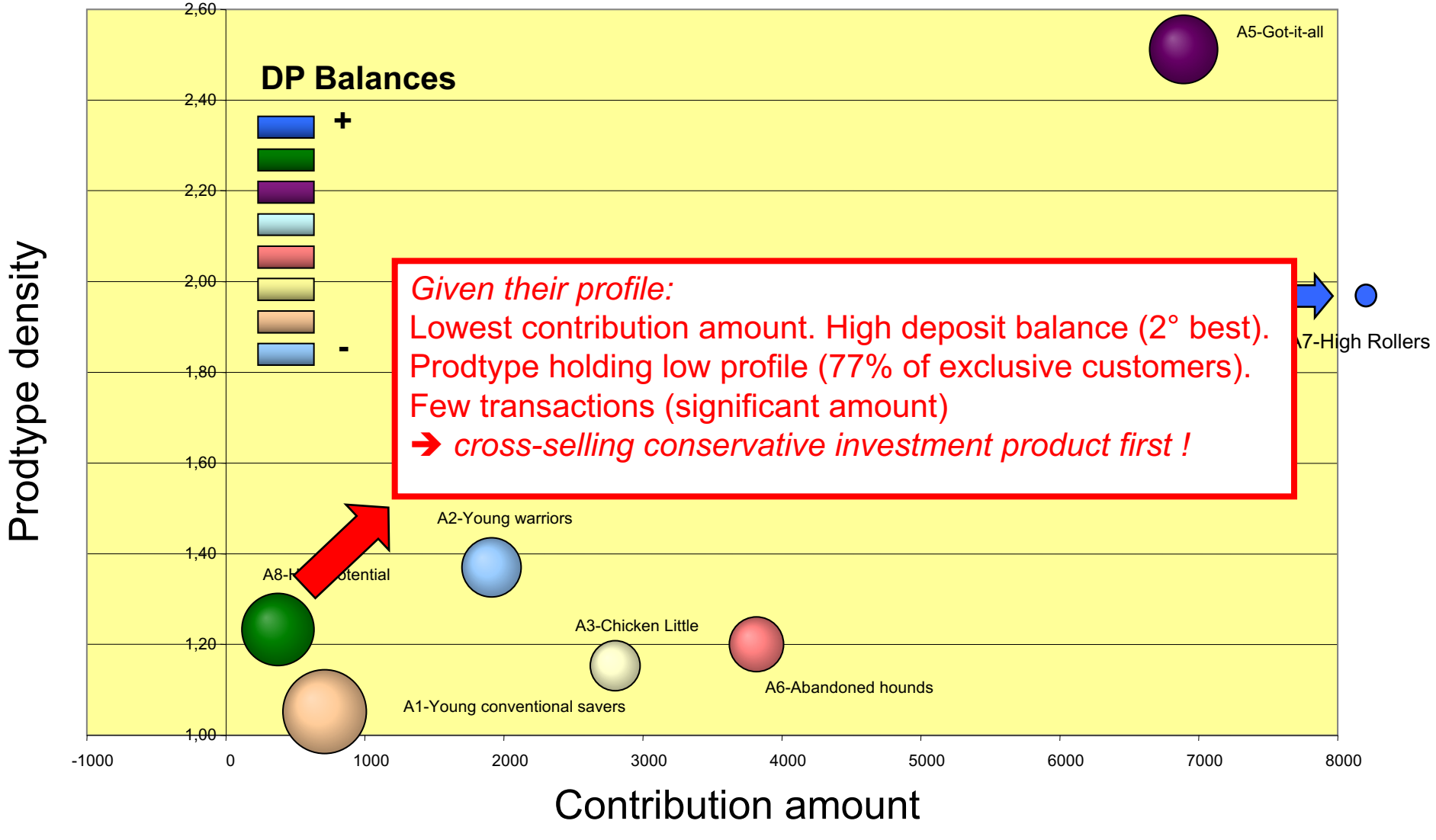
Active segments(2/2)



Scenario of marketing strategy- Active

Marketing	Active Segments	%	#	CRM Actions
VIP	High Rollers	1.9%	XXX	• **To Take Care (one-to-one)** • Maximum Time Consuming • Share-of-wallet Retention using Business and Management Rules
Premium	• Got-it-all • Frequent Flyers	21.4%	YYY	• **To Invest** • High Time Consuming • Up-selling using Business Rules and Propensity Models • Portfolio Retention using Churn Models
High Potential Customer	• …	X%	…	• **To Improve** • …
Basic Customers	• …	X%	…	• **To Manage** • ….

Cross-selling towards “High potential”



example

Segment profiling

	TOT. CLIENTI	CLUSTER 3	N. INDICE
C/C	79,26%	37,86%	47,77%
CARTE DI DEBITO	40,45%	12,43%	30,74%
CARTE DI CREDITO	23,48%	6,17%	26,28%
DEPOSITI A RISPARMIO	23,98%	74,10%	308,95%
CERTIFICATI DI DEPOSITO	2,19%	47,17%	2156,07%
OBBLIGAZIONI DELLA BANCA	1,07%	0,56%	52,69%
RISPARMIO AMMINISTRATO	22,26%	25,81%	115,93%
FONDI COMUNI INVESTIMENTO	18,75%	19,42%	103,54%
GESTIONI PATRIMONIALI	8,13%	3,54%	43,51%
ASSICURAZIONI VITA+ DANNI	5,30%	3,48%	65,56%
PRESTITI PERSONALI	6,51%	1,88%	28,89%
MUTUI E PRESTITI	3,88%	0,19%	4,84%
ACCREDITO STIPENDI	22,84%	13,75%	60,19%
ADDEBITO UTENZE	39,04%	14,03%	35,94%

Traditionalist

Cross-selling

example

Client	Services	Target
no →	utilities	73%
no →	leasing	57%
no	plastic cards	13%
no	loans	23%
no	policies	05%
yes	equity	34%
2	accounts	1.9

Mr. Verdi

classified as utilitarian

Utilitarians

Agenda

Module 4: Demand Segmentation

- Introduction
- The Multidensional Classfication Issue
- The Behavioral Segmentation
- The Analytical Process
  - Introduction
  - Factor Analysis
  - Cluster Analysis

Behavioral Segmentation

In the behavioral segmentation analysis (or homogeneity segmentation) the elements of a target population are gathered with regard to their similarity to a set of variables: needs, attitudes, benefits, reasons behind the use of goods, opportunities to use the items or life styles.

The groups thus formed must be characterized by:

- a low "within variability"
- a high "between variability"

From an application standpoint, we basically have two protocols of analysis for homogeneity segmentation:

- Traditional

✓ *combination of Factor and Cluster*

- Flexible

✓ *combination of Conjoint and Cluster*

Behavioral Segmentation

Traditional protocol entails:

- The collection of significance assessment on or about specific features of the products/services surveyed

- The synthesis in underline variables through the Factor analysis

- The built-up of homogeneous groups through the Cluster analysis using the underlines variables obtained in the previous step

- The cross between the clusters and the social and demographic data of the target population

Factor Analysis

We used the Factor Analysis in order to summarize and reduce the different variables into a lower number trying to loose the least number of information possible.

VARIABLES OF ANALYSIS

- Reasons that drive you to check facebook?
  - Make new friends
  - Keep in touch with friends
  - Reconnect with old classmates
  - Have news about products
  - Share photos and videos
  - Curiosity
  - Discuss interest and hobbies
  - Plan Parties and events
- Which features do you use?
  - Wall
  - Photo & Video
  - Private Messaging
  - Events Creation
  - Group Affiliation

Number of starting variables= 13

Factor Analysis

facebook

Final Factors

Behavioral Segmentation

facebook

Most Suitable Segmentation

K-Means with 5 Clusters

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Spying	20,723	4	,580	188	35,707	,000
Broadening	29,283	4	,398	188	73,533	,000
Keeping Up	23,121	4	,529	188	43,680	,000
Public Relations	20,953	4	,575	188	36,411	,000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Number of Cases in each Cluster

Cluster	1	54,000
	2	13,000
	3	38,000
	4	35,000
	5	53,000
Valid		193,000
Missing		,000

Final Cluster Centers

	Cluster				
	1	2	3	4	5
Spying	,11836	,25028	-,42659	-1,05320	,81938
Broadening	1,17308	-,46080	,03264	-,79728	-,57909
Keeping Up	,30329	-2,01646	-,71115	,42419	,41535
Public Relations	,31027	,93554	-1,17595	,63099	-,11916

Spying	3	2	4	5-	1+
Broadening	1++	3	2	5	4
Keeping Up	3	5--	4	1	2
Public Relations	3	1	5--	2	4

Behavioral Segmentation

facebook

The 5 Clusters

- **Cool Hunters (28%)**: More than all, they are users absolutely interested on Broadening.
- **PR's (7%)**: Interested above all in Public Relations and express some attachment to Spying, but not related at all with Keeping Up.
- **Detached (20%)**: Apart from some light interest on Broadening, they do not express any involvement with the Facebook use (in particular with Public Relations).
- **Functional (18%)**: Above all, interested in Keeping up with their network of friends and use Public Relations inside this network. Besides, they do not care at all about Spying and Broadening.
- **Gossipers (27%)**: They are also interested in Keeping up, but above all in Spying their network. Furthermore, they are not interested in Public Relations and Broadening.

Each single Cluster was then crossed with socio-demographic and usage variables, through the contingency table tool, in order to better understand their main characteristics. The following slides sum-up the most relevant results of these crossings for each single cluster.

Agenda

Module 4: Demand Segmentation

- Introduction
- The Multidensional Classfication Issue
- The Behavioral Segmentation
- The Analytical Process
  - Introduction
  - Factor Analysis
  - Cluster Analysis

Factor Analysis

12 Assign to each one of the following options a percentage of: Where do you usually use internet? (Sum percentages =100)

a. Home	
b. Work	
c. University	
	0%

13 What do you use internet for?

	Low			Medium			High		
	1	2	3	4	5	6	7	8	9
a. Sources of information									
b. Work									
c. Friendship									
d. Buy and sell									
e. University									
f. Organizing events									

Factor Analysis

Correlations

		Internet_ Home	Internet_Work	Internet_ University	Internet for_ Information	Internet for_Work	Internet for_ Friendship	Internet for_Buy&Sell	Internet for_University	Internet for_Organize Events
Internet_Home	Pearson Correlation	1	-.541**	-.746**	-.094	-.188**	-.031	-.085	.028	-.074
	Sig. (2-tailed)		.000	.000	.194	.009	.671	.239	.695	.309
	N	193	193	193	193	193	193	193	193	193
Internet_Work	Pearson Correlation	-.541**	1	-.156*	.077	.195**	-.031	.095	-.272**	.015
	Sig. (2-tailed)	.000		.030	.288	.007	.666	.188	.000	.836
	N	193	193	193	193	193	193	193	193	193
Internet_University	Pearson Correlation	-.746**	-.156*	1	.048	.067	.061	.025	.182*	.077
	Sig. (2-tailed)	.000	.030		.511	.357	.397	.735	.011	.289
	N	193	193	193	193	193	193	193	193	193
Internet for_Information	Pearson Correlation	-.094	.077	.048	1	.367**	.085	.154*	.135	.045
	Sig. (2-tailed)	.194	.288	.511		.000	.242	.033	.061	.534
	N	193	193	193	193	193	193	193	193	193
Internet for_Work	Pearson Correlation	-.188**	.195**	.067	.367**	1	.113	.031	-.029	.088
	Sig. (2-tailed)	.009	.007	.357	.000		.119	.669	.688	.222
	N	193	193	193	193	193	193	193	193	193
Internet for_Friendship	Pearson Correlation	-.031	-.031	.061	.085	.113	1	.025	.227**	.314**
	Sig. (2-tailed)	.671	.666	.397	.242	.119		.732	.001	.000
	N	193	193	193	193	193	193	193	193	193
Internet for_Buy&Sell	Pearson Correlation	-.085	.095	.025	.154*	.031	.025	1	.063	.208**
	Sig. (2-tailed)	.239	.188	.735	.033	.669	.732		.383	.004
	N	193	193	193	193	193	193	193	193	193
Internet for_University	Pearson Correlation	.028	-.272**	.182*	.135	-.029	.227**	.063	1	.167*
	Sig. (2-tailed)	.695	.000	.011	.061	.688	.001	.383		.020
	N	193	193	193	193	193	193	193	193	193
Internet for_Organize Events	Pearson Correlation	-.074	.015	.077	.045	.088	.314**	.208**	.167*	1
	Sig. (2-tailed)	.309	.836	.289	.534	.222	.000	.004	.020	
	N	193	193	193	193	193	193	193	193	193

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Factor Analysis

If the information is spread among many correlated variables:

$\Rightarrow$ *we may have several different problems.*

- Apparent information;
- Miss- understanding;
- Difficulties in the interpretation phase;
- Robustness of the results;
- Efficiency of the estimates;
- Degrees of freedom;
- …..

Factor Analysis

The high number and the correlation between variables lead to analysis problems:

=> *it's necessary to reduce their number, however making sure not to loose any valuable information.*

The Factor Analysis (FA) is a multivariate technique used to perform the analyses of correlation between quantitative variables.

Considering a data matrix: $X_{(nxp)}$, with "n" observations and "p" original variables, the use of the FA allows to summarize the information within a restricted set of transformed variables (the so called Factors or latent factors).

Factor Analysis

We used the Factor Analysis in order to summarize and reduce the different variables into a lower number trying to loose the least number of information possible.

VARIABLES OF ANALYSIS

- Reasons that drive you to check facebook?
  - Make new friends
  - Keep in touch with friends
  - Reconnect with old classmates
  - Have news about products
  - Share photos and videos
  - Curiosity
  - Discuss interest and hobbies
  - Plan Parties and events
- Which features do you use?
  - Wall
  - Photo & Video
  - Private Messaging
  - Events Creation
  - Group Affiliation

Number of starting variables= 13

Factor Analysis

facebook

Final Factors

Factor Analysis

Input: “p” quantitative variables characterized by significant level of correlations

Output: “k” new quantitative variables characterized by optimal properties (“k”<<“p”)

Method: Principal Components

Application field:

- Cluster Analysis – “p” is high and/or Correlation
- Regression Analysis – Correlation is significant

Factor Analysis

Pricipal Components Methods

One of the methods used for the creation of the Factors is the so called Method of the Principal Components (PC).

Using this method entails assuming the specific information contribution of the input variables is very low, whereas the shared information contribution is the highest and explained through the common factors.

Factor Analysis

Pricipal Components Methods

- The Factors computed trough the PC method are linear combinations of the original variables.

$$CP_j = s_{j1}x_1 + s_{j2}x_2 + \ldots\ldots\ldots\ldots + s_{jp}x_p$$

- They are new standardized variables
- They are orthogonal between each other (non correlated)
- Altogether they explain the variability of the “p” original variables
- They are listed in descending order related to the explained variability

Factor Analysis

Pricipal Components Methods

The maximum number of Factors (or principal components) is equal to the number of the original variables (in our case “p”).

The first Factor (or principal component) is a linear combination of the original “p” variables and it is characterized by the highest variability, all the way down to the last Factor, which is again a linear combination of the “p” input variables and it has assigned the lowest level of variability.

If the correlation between the “p” variables is high, it is enough to consider a few components, such as k<<p (k much lower than p) in order to adequately represent the original data. These principal components sum up a large part of the total variability.

Factor Analysis

Issues of the Factor Analysis are the following:

a) How many Factors (or components) need to be considered

1. The ratio between the number of components and the variables;
2. The percentage of the explained variance;
3. The scree plot;
4. Eigenvalue;
5. The communalities

b) How to interpret

6. The Component Matrix
7. The Rotated Component Matrix

Factor Analysis

- Twenty items related to the product “Biscuits” have been indicated
- The interviewees have been asked to make a judgment with regard to the importance of each of these elements upon buying the product biscuits

1. Quality of ingredients
2. Genuineness
3. Lightness
4. Taste
5. Nutritional features
6. Specific needs
7. Natural yeast process
8. Handicraft
9. Shape
10. Reference to tradition
11. Size of the package
12. Functionality of the package
13. Attractiveness of the package
14. Expire date
15. Brand name
16. Advertising
17. Promotions
18. Various tips
19. Price
20. Brand awareness

1. Quality of ingredients – Qualità degli ingredienti
2. Genuineness – Genuinità
3. Lightness – Leggerezza
4. Taste – Sapore/Gusto
5. Nutritional features – Caratteristiche nutrizionali
6. Specific needs – Attenzione a bisogni specifici
7. Natural yeast process – Lievitazione naturale
8. Handicraft – Produzione artiginale
9. Shape – Forma e stampo
10. Reference to tradition – Richiamo alla tradizione
11. Size of the package – Grandezza della confezione
12. Functionality of the package – Funzionalità della confezione
13. Attractiveness of the package – Estetica della confezione
14. Expire date – Scadenza
15. Brand name – Nome del biscotto
16. Advertising – Pubblicità e comunicazione
17. Promotions – Promozione e offerte speciali
18. Various tips – Consigli per l'utilizzo
19. Price – Prezzo
20. Brand awareness – Notorietà della marca

Factor Analysis

Correlations

		Qualità degli ingredienti	Genuinità	Leggerezza	Sapore/gusto	Caratteristiche nutrizionali
Qualità degli ingredienti	Pearson Correlation	1	.629**	.299**	.232**	.234**
	Sig. (2-tailed)		.000	.000	.001	.001
	N	220	220	218	220	214
Genuinità	Pearson Correlation	.629**	1	.468**	.090	.354**
	Sig. (2-tailed)	.000		.000	.181	.000
	N	220	220	218	220	214
Leggerezza	Pearson Correlation	.299**	.468**	1	.030	.460**
	Sig. (2-tailed)	.000	.000		.657	.000
	N	218	218	219	219	213
Sapore/gusto	Pearson Correlation	.232**	.090	.030	1	-.015
	Sig. (2-tailed)	.001	.181	.657		.823
	N	220	220	219	221	215
Caratteristiche nutrizionali	Pearson Correlation	.234**	.354**	.460**	-.015	1
	Sig. (2-tailed)	.001	.000	.000	.823	
	N	214	214	213	215	215

**. Correlation is significant at the 0.01 level (2-tailed).

Total Variance Explained

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	4.171	20.853	20.853
2	2.678	13.389	34.241
3	1.843	9.216	43.457
4	1.376	6.879	50.336
5	1.129	5.643	55.979
6	1.016	5.079	61.057
7	.937	4.684	65.741
8	.881	4.405	70.146
9	.781	3.907	74.054
10	.751	3.756	77.810
11	.682	3.412	81.222
12	.592	2.960	84.183
13	.568	2.838	87.021
14	.550	2.750	89.771
15	.453	2.267	92.038
16	.386	1.930	93.968
17	.376	1.880	95.848
18	.324	1.621	97.470
19	.270	1.352	98.822
20	.236	1.178	100.000

Extraction Method: Principal Component Analysis.

1. The ratio between the number of components and the variables:

One out of Three

20 original variables
6-7 Factors

Total Variance Explained

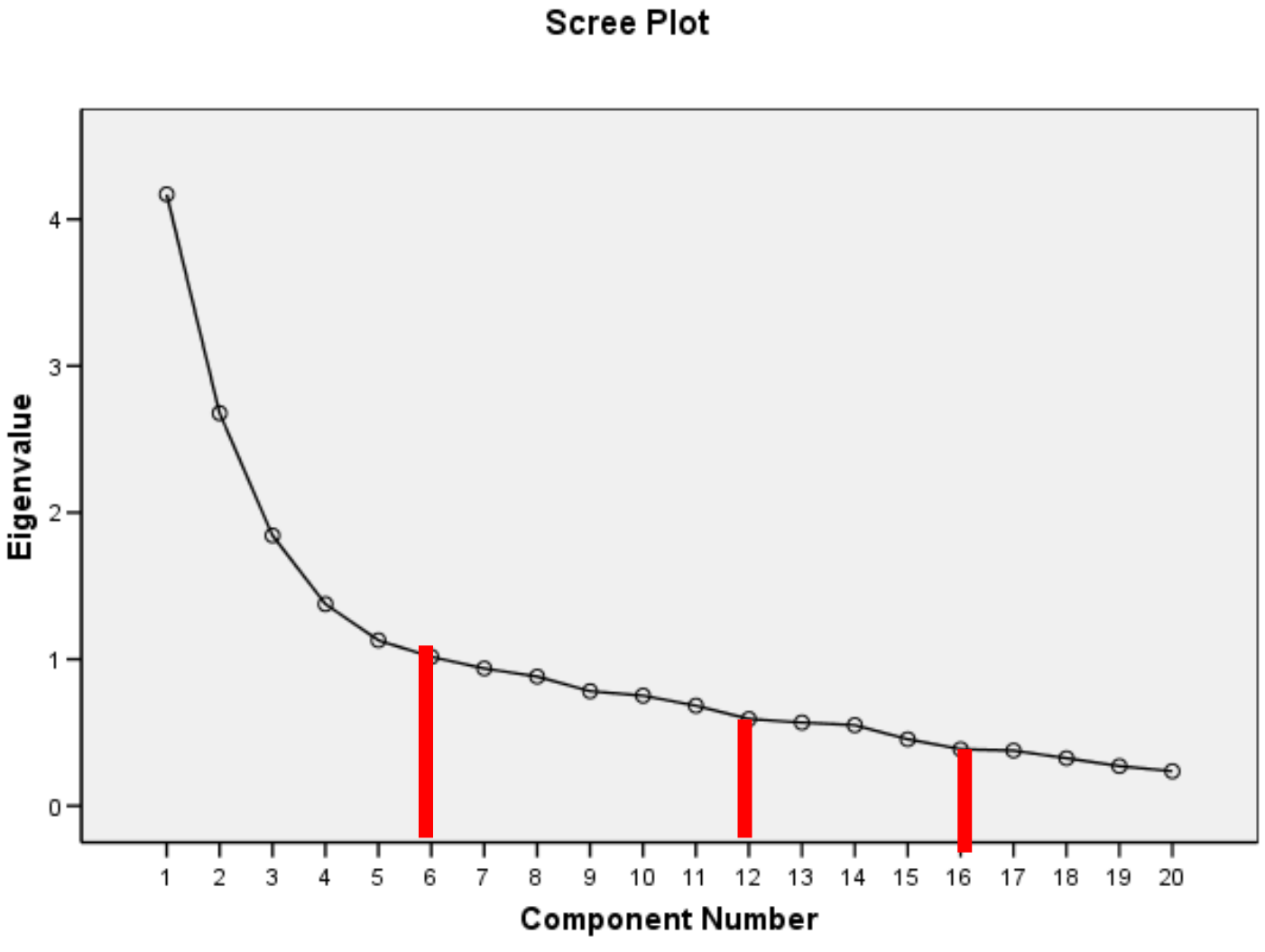
Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	4.171	20.853	20.853
2	2.678	13.389	34.241
3	1.843	9.216	43.457
4	1.376	6.879	50.336
5	1.129	5.643	55.979
6	1.016	5.079	61.057
7	.937	4.684	65.741
8	.881	4.405	70.146
9	.781	3.907	74.054
10	.751	3.756	77.810
11	.682	3.412	81.222
12	.592	2.960	84.183
13	.568	2.838	87.021
14	.550	2.750	89.771
15	.453	2.267	92.038
16	.386	1.930	93.968
17	.376	1.880	95.848
18	.324	1.621	97.470
19	.270	1.352	98.822
20	.236	1.178	100.000

Extraction Method: Principal Component Analysis.

2. The percentage of the explained variance:

Between 60%-75%

Factor Analysis



3. The scree plot :

The point at which the scree begins

Total Variance Explained

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	4.171	20.853	20.853
2	2.678	13.389	34.241
3	1.843	9.216	43.457
4	1.376	6.879	50.336
5	1.129	5.643	55.979
6	1.016	5.079	61.057
7	.937	4.684	65.741
8	.881	4.405	70.146
9	.781	3.907	74.054
10	.751	3.756	77.810
11	.682	3.412	81.222
12	.592	2.960	84.183
13	.568	2.838	87.021
14	.550	2.750	89.771
15	.453	2.267	92.038
16	.386	1.930	93.968
17	.376	1.880	95.848
18	.324	1.621	97.470
19	.270	1.352	98.822
20	.236	1.178	100.000

Extraction Method: Principal Component Analysis.

4. Eigenvalue:

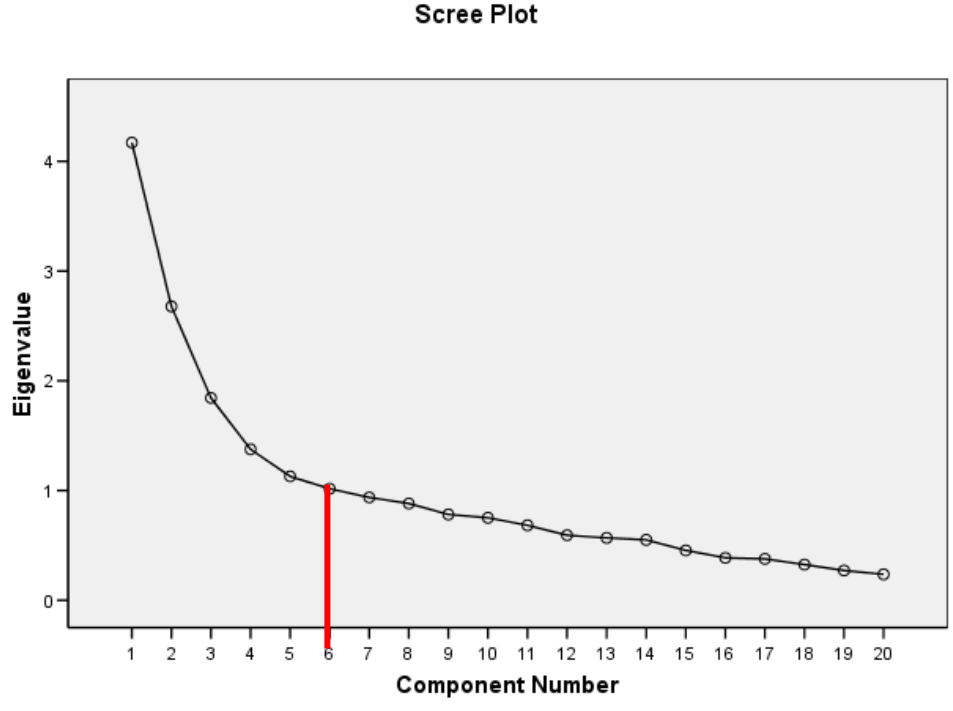
Eigenvalues>1

Factor Analysis

Total Variance Explained

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	4.171	20.853	20.853
2	2.678	13.389	34.241
3	1.843	9.216	43.457
4	1.376	6.879	50.336
5	1.129	5.643	55.979
6	1.016	5.079	61.057
7	.937	4.684	65.741
8	.881	4.405	70.146
9	.781	3.907	74.054
10	.751	3.756	77.810
11	.682	3.412	81.222
12	.592	2.960	84.183
13	.568	2.838	87.021
14	.550	2.750	89.771
15	.453	2.267	92.038
16	.386	1.930	93.968
17	.376	1.880	95.848
18	.324	1.621	97.470
19	.270	1.352	98.822
20	.236	1.178	100.000

Extraction Method: Principal Component Analysis.



Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.171	20.853	20.853	4.171	20.853	20.853
2	2.678	13.389	34.241	2.678	13.389	34.241
3	1.843	9.216	43.457	1.843	9.216	43.457
4	1.376	6.879	50.336	1.376	6.879	50.336
5	1.129	5.643	55.979	1.129	5.643	55.979
6	1.016	5.079	61.057	1.016	5.079	61.057
7	.937	4.684	65.741			
8	.881	4.405	70.146			
9	.781	3.907	74.054			
10	.751	3.756	77.810			
11	.682	3.412	81.222			
12	.592	2.960	84.183			
13	.568	2.838	87.021			
14	.550	2.750	89.771			
15	.453	2.267	92.038			
16	.386	1.930	93.968			
17	.376	1.880	95.848			
18	.324	1.621	97.470			
19	.270	1.352	98.822			
20	.236	1.178	100.000			

Extraction Method: Principal Component Analysis.

Communalities

	Initial	Extraction
Qualità degli ingredienti	1.000	.717
Genuinità	1.000	.746
Leggerezza	1.000	.588
Sapore/gusto	1.000	.670
Caratteristiche nutrizionali	1.000	.631
Attenzione a bisogni specifici	1.000	.332
Lievitazione naturale	1.000	.674
Produzione artigianale	1.000	.762
Forma e stampo	1.000	.689
Richiamo alla tradizione	1.000	.600
Grandezza della confezione (peso netto)	1.000	.579
Funzionalità della confezione	1.000	.414
Estetica della confezione	1.000	.599
Scadenza	1.000	.432
Nome del biscotto	1.000	.494
Pubblicità e comunicazione	1.000	.717
Promozioni e offerte speciali	1.000	.736
Consigli per l'utilizzo	1.000	.463
Prezzo	1.000	.653
Notorietà della marca	1.000	.716

Extraction Method: Principal Component Analysis.

5. Communalities:

The quote of explained variability for each input variable must be *satisfactory (higer than 0.3-0.4)*

In the example the overall explained variability (which represents the mean value) is 0.61057

Factor Analysis

Issues of the Factor Analysis are the following:

a) How many Factors (or components) need to be considered

b) How to interpret

6. The Component Matrix
   - *The correlation between the principal components and the original variables*
7. The Rotated Component Matrix
   - *A mathematical transformation to facilitate the interpretation*

Factor Analysis

- 6. The Component Matrix *(factor loadings)*
  - The most relevant output of a factorial analysis is the so called “component matrix”, which shows the correlations between the original input variables and the obtained components (factor loadings)
  - Each variable is associated specifically to the factors (components) with which there is the highest correlation
  - The interpretation of the each factor has to be guided considering the variables with the highest correlations related to single factor

Component Matrix[a]

	Component					
	1	2	3	4	5	6
Qualità degli ingredienti	.418	-.513	.072	.099	.375	.353
Genuinità	.383	-.717	.082	-.080	.137	.231
Leggerezza	.426	-.478	.136	-.349	.162	.105
Sapore/gusto	.163	-.079	.195	.671	.229	.310
Caratteristiche nutrizionali	.410	-.364	.298	-.417	.100	-.240
Attenzione a bisogni specifici	.410	-.220	-.214	-.197	-.032	-.172
Lievitazione naturale	.624	-.360	-.309	.019	-.228	-.083
Produzione artigianale	.573	-.339	-.160	.377	-.374	-.109
Forma e stampo	.482	.320	-.272	.202	.430	-.234
Richiamo alla tradizione	.615	.046	-.269	.372	-.082	-.045
Grandezza della confezione (peso netto)	.403	.287	.461	.196	.209	-.197
Funzionalità della confezione	.483	.131	.162	-.123	.081	-.340
Estetica della confezione	.463	.439	-.383	-.026	.174	-.118
Scadenza	.390	-.158	.100	.088	-.473	-.118
Nome del biscotto	.416	.306	-.383	-.126	.252	.032
Pubblicità e comunicazione	.421	.525	-.145	-.331	-.062	.361
Promozioni e offerte speciali	.340	.419	.660	-.062	-.025	.073
Consigli per l'utilizzo	.629	.123	.093	-.173	-.058	.104
Prezzo	.429	.265	.594	.129	-.166	-.047
Notorietà della marca	.413	.434	-.115	-.121	-.305	.486

Extraction Method: Principal Component Analysis.

a. 6 components extracted.

6. Interpretation:

Correlation between Input Vars & Factors

The new Factors must have a meaning based on the correlation structure

Rotated Component Matrix[a]

	Component					
	1	2	3	4	5	6
Genuinità	.795	-.089	-.123	.237	-.051	.178
Leggerezza	.748	.072	-.007	.096	.050	-.104
Qualità degli ingredienti	.716	-.026	.078	.080	.007	.437
Caratteristiche nutrizionali	.619	.312	.009	.111	-.127	-.349
Attenzione a bisogni specifici	.327	-.054	.243	.324	.020	-.239
Promozioni e offerte speciali	.002	.799	-.052	-.111	.286	.035
Prezzo	-.015	.764	-.063	.180	.154	.092
Grandezza della confezione (peso netto)	.017	.697	.250	.006	-.067	.159
Funzionalità della confezione	.158	.448	.334	.165	-.028	-.219
Forma e stampo	-.011	.163	.799	.070	-.024	.137
Estetica della confezione	-.096	.065	.704	.107	.268	-.076
Nome del biscotto	.071	-.040	.624	.005	.309	-.047
Produzione artigianale	.158	.028	.083	.836	-.023	.172
Lievitazione naturale	.369	-.103	.224	.681	.094	-.065
Scadenza	.066	.211	-.137	.593	.078	-.086
Richiamo alla tradizione	.023	.082	.439	.566	.132	.251
Notorietà della marca	-.083	.108	.103	.161	.811	.051
Pubblicità e comunicazione	-.002	.139	.310	-.055	.764	-.119
Consigli per l'utilizzo	.282	.342	.228	.234	.394	-.064
Sapore/gusto	.048	.163	.025	.083	-.074	.793

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 6 iterations.

6. Interpretation:

The correlation structure between Input Vars & Factors

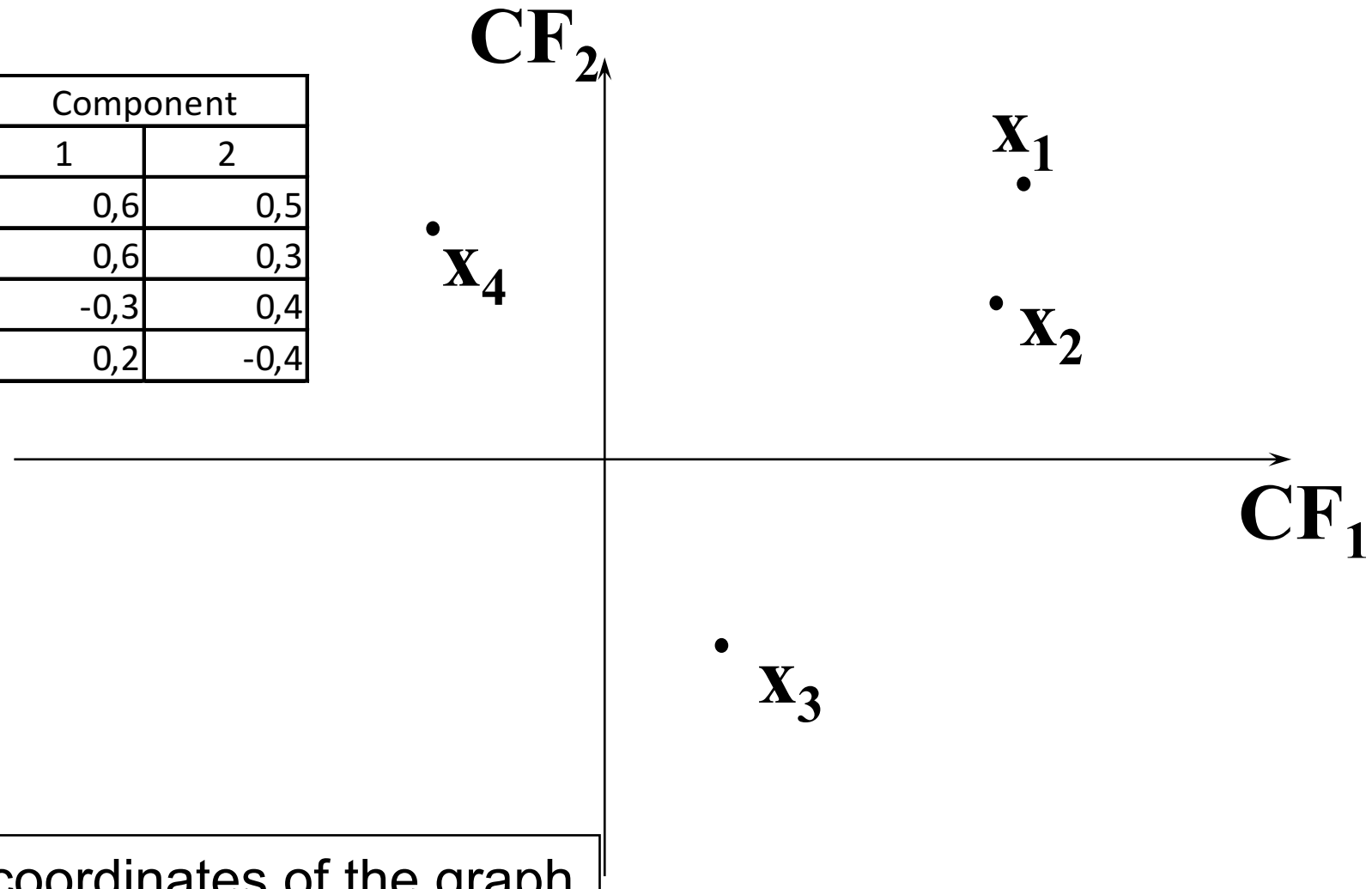
In this case the correlation structure is well defined and the interpretation phase is easier

Factor Analysis

- 7. The Rotated Component Matrix (The rotation of factors)
  - There are numerous outputs of factorial analysis which can be produced through the same input data
  - These numerous outputs don't provide interpretation that are remarkably different from one another, as matter of fact they differ only slightly and there are areas of ambiguity

Factor Analysis

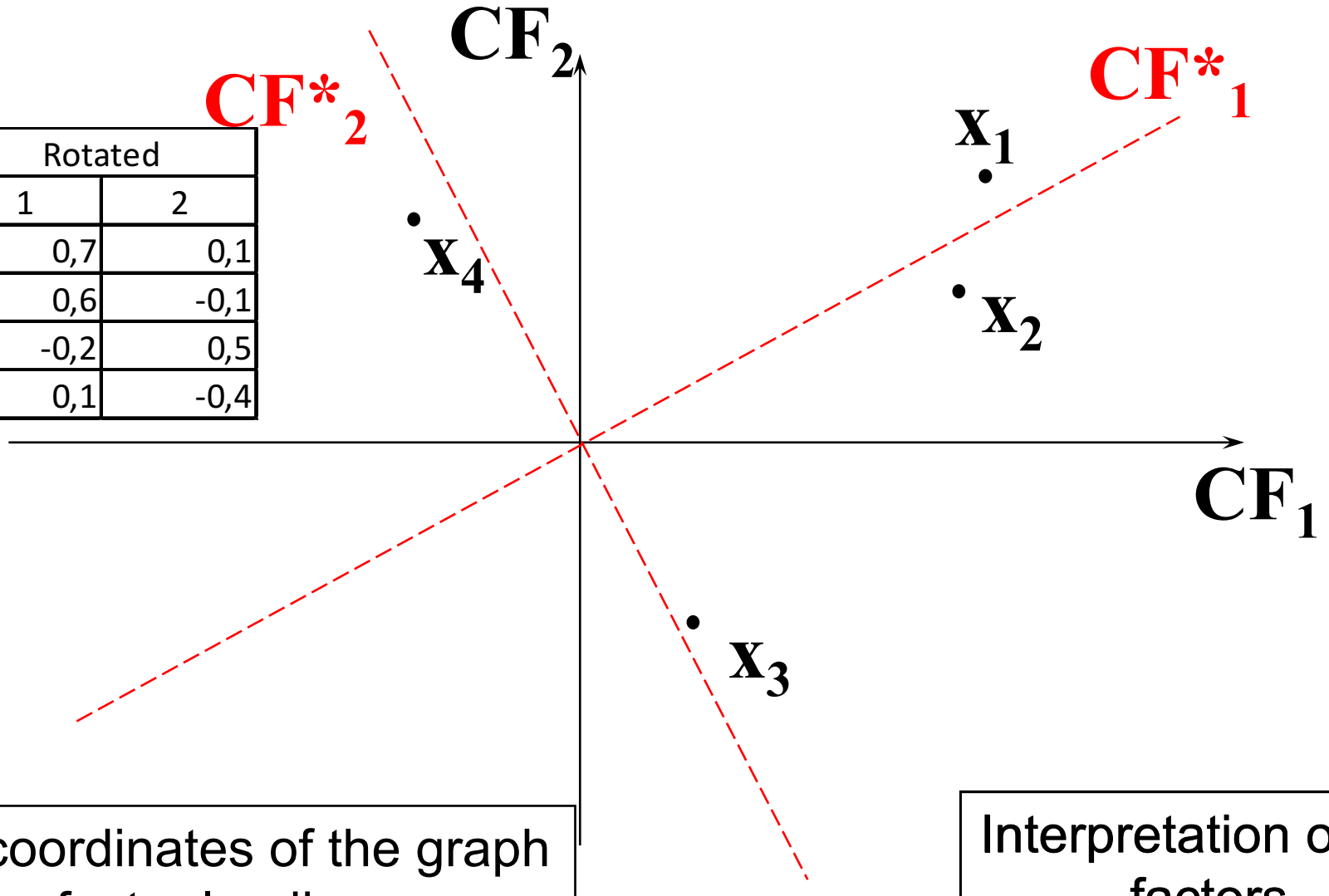
	Component	
	1	2
X1	0,6	0,5
X2	0,6	0,3
X3	-0,3	0,4
X4	0,2	-0,4



The coordinates of the graph are the factor loadings

Factor Analysis

	Rotated	
	1	2
X1	0,7	0,1
X2	0,6	-0,1
X3	-0,2	0,5
X4	0,1	-0,4



The coordinates of the graph are the factor loadings

Interpretation of the factors

Factor Analysis

- 6. Interpretation: The rotation of factors
  - The Varimax method of rotation, suggested by Kaiser, has the purpose of minimizing the number of variables with high saturations (correlations) for each factor
  - The Quartimax method attempts to minimize the number of factors tightly correlated to each variable
  - The Equimax method is a cross between the Varimax and the Quartimax
  - The percentage of the overall variance of the rotated factors doesn't change, whereas the percentage of the variance explained by each factors shifts

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.171	20.853	20.853	4.171	20.853	20.853
2	2.678	13.389	34.241	2.678	13.389	34.241
3	1.843	9.216	43.457	1.843	9.216	43.457
4	1.376	6.879	50.336	1.376	6.879	50.336
5	1.129	5.643	55.979	1.129	5.643	55.979
6	1.016	5.079	61.057	1.016	5.079	61.057
7	.937	4.684	65.741			
8	.881	4.405	70.146			
9	.781	3.907	74.054			
10	.751	3.756	77.810			
11	.682	3.412	81.222			
12	.592	2.960	84.183			
13	.568	2.838	87.021			
14	.550	2.750	89.771			
15	.453	2.267	92.038			
16	.386	1.930	93.968			
17	.376	1.880	95.848			
18	.324	1.621	97.470			
19	.270	1.352	98.822			
20	.236	1.178	100.000			

Extraction Method: Principal Component Analysis.

Before the rotation step

Total Variance Explained

Component	Initial Eigenvalues			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.171	20.853	20.853	2.490	12.448	12.448
2	2.678	13.389	34.241	2.294	11.468	23.917
3	1.843	9.216	43.457	2.214	11.068	34.984
4	1.376	6.879	50.336	2.203	11.016	46.000
5	1.129	5.643	55.979	1.736	8.680	54.680
6	1.016	5.079	61.057	1.276	6.378	61.057
7	.937	4.684	65.741			
8	.881	4.405	70.146			
9	.781	3.907	74.054			
10	.751	3.756	77.810			
11	.682	3.412	81.222			
12	.592	2.960	84.183			
13	.568	2.838	87.021			
14	.550	2.750	89.771			
15	.453	2.267	92.038			
16	.386	1.930	93.968			
17	.376	1.880	95.848			
18	.324	1.621	97.470			
19	.270	1.352	98.822			
20	.236	1.178	100.000			

Extraction Method: Principal Component Analysis.

After the rotation step

Communalities

	Initial	Extraction
Qualità degli ingredienti	1.000	.717
Genuinità	1.000	.746
Leggerezza	1.000	.588
Sapore/gusto	1.000	.670
Caratteristiche nutrizionali	1.000	.631
Attenzione a bisogni specifici	1.000	.332
Lievitazione naturale	1.000	.674
Produzione artigianale	1.000	.762
Forma e stampo	1.000	.689
Richiamo alla tradizione	1.000	.600
Grandezza della confezione (peso netto)	1.000	.579
Funzionalità della confezione	1.000	.414
Estetica della confezione	1.000	.599
Scadenza	1.000	.432
Nome del biscotto	1.000	.494
Pubblicità e comunicazione	1.000	.717
Promozioni e offerte speciali	1.000	.736
Consigli per l'utilizzo	1.000	.463
Prezzo	1.000	.653
Notorietà della marca	1.000	.716

Extraction Method: Principal Component Analysis.

5. Communalities:

The communalities don't change after the Rotation Step

Rotated Component Matrix[a]

	Component					
	1	2	3	4	5	6
Genuinità	.795	-.089	-.123	.237	-.051	.178
Leggerezza	.748	.072	-.007	.096	.050	-.104
Qualità degli ingredienti	.716	-.026	.078	.080	.007	.437
Caratteristiche nutrizionali	.619	.312	.009	.111	-.127	-.349
Attenzione a bisogni specifici	.327	-.054	.243	.324	.020	-.239
Promozioni e offerte speciali	.002	.799	-.052	-.111	.286	.035
Prezzo	-.015	.764	-.063	.180	.154	.092
Grandezza della confezione (peso netto)	.017	.697	.250	.006	-.067	.159
Funzionalità della confezione	.158	.448	.334	.165	-.028	-.219
Forma e stampo	-.011	.163	.799	.070	-.024	.137
Estetica della confezione	-.096	.065	.704	.107	.268	-.076
Nome del biscotto	.071	-.040	.624	.005	.309	-.047
Produzione artigianale	.158	.028	.083	.836	-.023	.172
Lievitazione naturale	.369	-.103	.224	.681	.094	-.065
Scadenza	.066	.211	-.137	.593	.078	-.086
Richiamo alla tradizione	.023	.082	.439	.566	.132	.251
Notorietà della marca	-.083	.108	.103	.161	.811	.051
Pubblicità e comunicazione	-.002	.139	.310	-.055	.764	-.119
Consigli per l'utilizzo	.282	.342	.228	.234	.394	-.064
Sapore/gusto	.048	.163	.025	.083	-.074	.793

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 6 iterations.

6. Interpretation:

The correlation structure between Input Vars & Factors improves after the rotation step

Rotated Component Matrix[a]

	Component					
	1	2	3	4	5	6
Genuinità	.795	-.089	-.123	.237	-.051	.178
Leggerezza	.748	.072	-.007	.096	.050	-.104
Qualità degli ingredienti	.716	-.026	.078	.080	.007	.437
Caratteristiche nutrizionali	.619	.312	.009	.111	-.127	-.349
Attenzione a bisogni specifici	.327	-.054	.243	.324	.020	-.239
Promozioni e offerte speciali	.002	.799	-.052	-.111	.286	.035
Prezzo	-.015	.764	-.063	.180	.154	.092
Grandezza della confezione (peso netto)	.017	.697	.250	.006	-.067	.159
Funzionalità della confezione	.158	.448	.334	.165	-.028	-.219
Forma e stampo	-.011	.163	.799	.070	-.024	.137
Estetica della confezione	-.096	.065	.704	.107	.268	-.076
Nome del biscotto	.071	-.040	.624	.005	.309	-.047
Produzione artigianale	.158	.028	.083	.836	-.023	.172
Lievitazione naturale	.369	-.103	.224	.681	.094	-.065
Scadenza	.066	.211	-.137	.593	.078	-.086
Richiamo alla tradizione	.023	.082	.439	.566	.132	.251
Notorietà della marca	-.083	.108	.103	.161	.811	.051
Pubblicità e comunicazione	-.002	.139	.310	-.055	.764	-.119
Consigli per l'utilizzo	.282	.342	.228	.234	.394	-.064
Sapore/gusto	.048	.163	.025	.083	-.074	.793

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 6 iterations.

6. Interpretation:

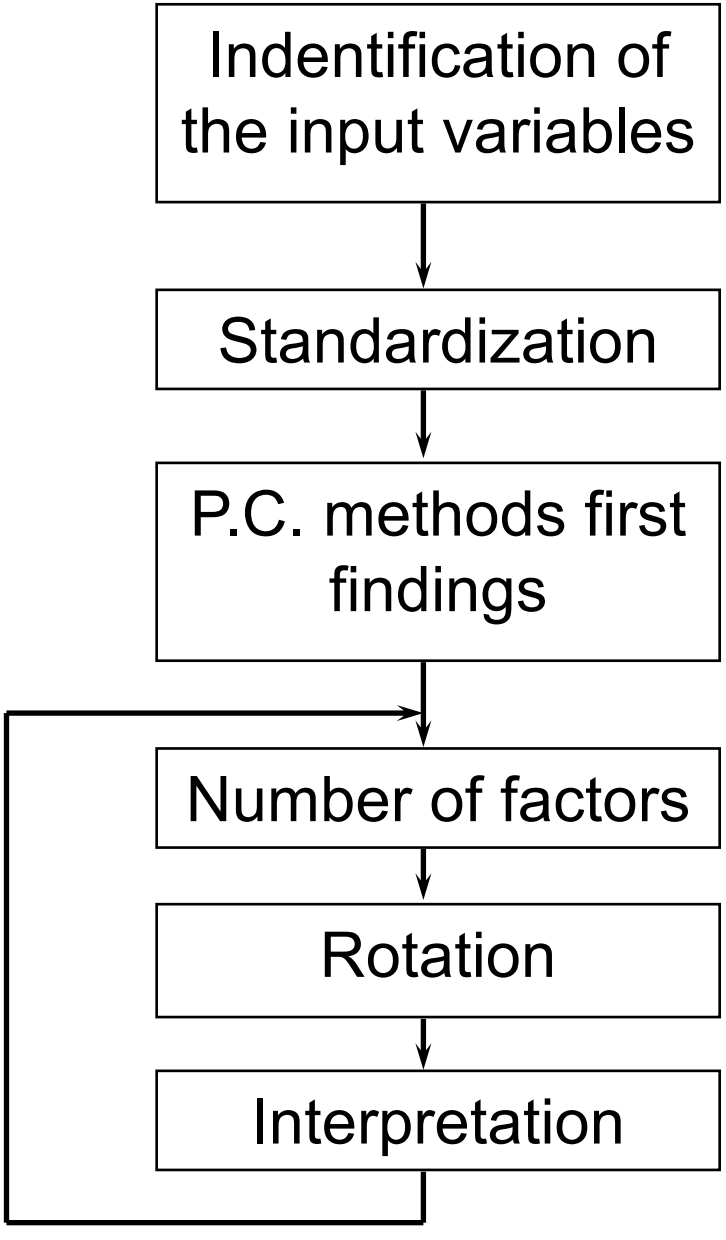
The correlation structure between Input Vars & Factors

The variable with the lowest communality is not well explained by this solution

Factor Analysis

- Once an adequate solution is found, it is possible to use the obtained factors as new macro variables to consider for further analyses on the phenomenon under investigation, thus replacing the original variables;
- Again taking into consideration the example, we may add six new variables into the data file, as follows:
  - Health,
  - Convenience & Practicality,
  - Image,
  - Handicraft,
  - Communication,
  - Taste.
- They are standardized variables: zero mean and variance equal to one.
- They will be the input for further analyses of Dependence or/and Interdependence.

Factor Analysis



Factor Analysis

The hypotheses of the FACTORIAL MODEL

Quantitative Variables $x_1, x_2, ..., x_i, ... x_p$

Info	x_i	=	Shared Info	+	Specific Info
Var	x_i	=	Communality	+	Specific Variability
	x_i	=	$f(CF_1, ...,CF_k)$	+	UF_i

$i = 1, ..., p$
$k << p$

CF_i = Common Factor$_i$
UF_i = Unique Factor$_i$

$Corr (UF_i , UF_j) = 0$ per $i \neq j$
$Corr (CF_i , CF_j) = 0$ per $i \neq j$
$Corr (CF_i , UF_j) = 0$ per ogni i,j

Factor Analysis

Factor Loadings & Factor Score Coefficients

$x_i = l_{i1}CF_1 + l_{i2}CF_2 + ... + l_{ik}CF_k + UFi$

$l_{i1}, l_{i2},.....,l_{ik}$ factor loadings

$i = 1, ..., p$ *factors meaning*

$CF_j = s_{j1}x_1 + s_{j2}x_2 + ... + s_{jp}x_p$

$s_{j1}, s_{j2},.....,s_{jp}$ factor score coefficients

$j = 1, ..., k << p$ *factors built-up*

Factor Analysis

Pricipal Components Methods

One of the methods used for the estiamation of the coefficients (LOADINGS) is the so called Method of the Principal Components (PC).

Using this method entails assuming the specific information contribution of the input variables is very low, whereas the shared information contribution is the highest and explained through the common factors.

As for the estimate of the loadings, the Eigenvalues and the Eigenvectors of the correlation matrix of the input variables are needed: actually the loadings coincide with the coefficients of correlation between the input variables and the Factors (so called Principal Components).

Factor Analysis

Pricipal Components Methods

- The Factors computed trough the PC method are linear combinations of the original variables.

$$CP_j = s_{j1}x_1 + s_{j2}x_2 + \ldots\ldots\ldots\ldots + s_{jp}x_p$$

- They are new standardized variables
- They are orthogonal between each other (non correlated)
- Altogether they explain the variability of the “p” original variables
- They are listed in descending order related to the explained variability

Keywords

- Multivariate Analysis
- Analysis of Dependence
- Analysis of Inter-Dependence
- Factor Analysis
  - Principal Components Method
  - Factors
    - Linear combination
    - Standardized
    - Orthogonal
  - Variability
  - Factor Loadings
  - Factor Score coefficients
  - Eigenvalues
  - Scree Plot
  - Communalities
  - Component Matrix
  - Factor Rotation
    - Varimax
    - Equimax
    - Quartimax
  - Rotated Component Matrix
  - Factor Interpretation

Text Book

Naresh K. Malhotra, "*Marketing Research – An Applied Orientation*", Pearson – Prentice Hall, 2010

- Chapter 19 – pag 634-659

Agenda

Module 4: Demand Segmentation

- Introduction
- The Multidensional Classfication Issue
- The Behavioral Segmentation
- The Analytical Process
  - Introduction
  - Factor Analysis
  - Cluster Analysis

Cluster analysis

The Cluster analysis is an automatic classification technique which classifies the statistical units into groups or clusters.

The clusters created as such are internally homogeneous, the statistical units belonging to the same cluster are similar, but different among them.

In the marketing activities the Cluster analysis is used to support strategic as well as operative decisions:

1) Targeting

2) *New product developments*

3) *Market test areas*

Cluster analysis

Objectives

Given a set $I=\{I1, I2,..., In\}$ consisting of n statistical units (products, brands, individuals, …), we assume we have collected the values of p variables $X1, X2,..., Xp$ related to each of the objects considered.

The main goal of a Cluster analysis procedure is to obtain a partition of the set I in k subsets $C1, C2,..., Ck$, the *clusters*, in such a way that some requirements are met.

Cluster analysis

Objectives

1) $k < n$: the analysis must lead to a synthesis of the units

2) $C_h \bigcap C_j = \varnothing \, h, j = 1,2,...,k$: the intersection of two clusters is equal to the empty set, that is every “object” must belong just to one cluster

3) $\bigcup_{i=1}^{k} C_i = I$: the join of the k clusters is the set of the n original elements

Cluster analysis

Objectives

The clusters C_1, C_2,..., C_k must have two essential characteristics:

- Inner homogeneity, in that the elements belonging to the same cluster should be as homogeneous as possible,
- External heterogeneity, in that the elements belonging to each cluster should be as dis-homogeneous to one another as possible

In Cluster Analysis algorithms is necessary to define a measure of homogeneity among the “objects”. Specifically, the homogeneity is defined in terms of lower distance or higher similarity among the objects themselves.

Cluster analysis

Distance function

The most common distance used is the Euclidean distance. Let **X** be a data matrix $_{n \times p}$ with rows **x**1',..,**x***n*'. The squared Euclidean distance between **x***i* and **x***j* (*dij*) is defined as below:

$$d_{ij}^2 = \sum_{s=1}^{p} (x_{is} - x_{js})^2$$

In case the variables used to carry out the clusterization have a different unit of measurement, it's necessary to standardize the variables before beginning the analysis by using the Euclidean distance.

Cluster analysis

Distance function

Several algorithms have been suggested which allow to get to a solution that is the closest possible solution to the optimal one by considering a limited number of possible alternatives.

Overall we have two major groups of classification algorithms:

- Non Hierarchical or Direct classification algorithms, whose purpose is to minimize a given distance function in order to create the clusters.
- Hierarchical algorithms, with an iterative procedure which gives rise to a hierarchy in the partitions.

Behavioral Segmentation Analysis

Non Hierarchical Cluster Analysis: *K-Means algorithm*

- Part I: methodological aspects

k-means algorithm

This type of algorithm is largely used particularly in order to analyze big sets of data, streamlining computational time and hardware resources.

k-means algorithms are characterized by an iterative procedure which optimizes the process of data partitioning.

k-means algorithm

Such methods assume that the number of clusters be set in advance. This assumption may be easily modified and tested by repeating the same procedure in a very simple and efficient way.

To perform the analysis we need to set three different criteria:

1. the choice of centers of the initial clusters;
2. the allocation of the elements in the initial clusters;
3. the criteria to quit the iterative procedure.

The differences between the various k-means algorithms are found especially in (1) and (3).

k-means algorithm

Let's assume we want clsssify a set *I* of *n* "objects", with "*p*" variables. In addition, let's consider:

- Euclidean distance, expressed with *d*
- a set of *k* clusters.

The procedure is iterative and is carried out through *m* steps:

- STEP 0.

Initially temporary *k* centers are set; the choice may be random or may the result of a predefined streamlining criteria. Such *k* centers, which we indicate with

$$Cen_1^0, Cen_2^0, \ldots, Cen_k^0$$

Create a partition P_0 in *k* clusters

$$C_1^0, C_2^0, \ldots, C_k^0$$

k-means algorithm

The rule of appointment is the following: an object belongs to the i-th cluster if it's the closest (that is with the shorter distance) to Cen_j than all the other centers.

- STEP 1.

The *k* centers are rearrenged by computing the average values on all the variables within the groups $C_1^0, C_2^0, \ldots, C_k^0$

$$Cen_1^1, Cen_2^1, \ldots, Cen_k^1$$

created in step 0. Such new centers determine a second partition P_1, built up following the very same criteria used for P_0 and made up of clusters

$$C_1^1, C_2^1, \ldots, C_k^1$$

k-means algorithm

- STEP *m*.

The k centers $Cen_1^m, Cen_2^m, \ldots, Cen_k^m$

created in step *m*-1 are rearranged; such new centers in turn determine a new partition P_m made up of clusters $C_1^m, C_2^m, \ldots, C_w^m$

The algorithm stops

1) When two consecutive iterations lead to the same partition
2) When the number of itrerations previously set has been reached
3) According to a loss fuction previosly defined

k-means algorithm

It is possible to prove that the K-means procedure tends to minimize the Within Variance of the final clusters.

The final partition may sometimes depend on the choice of the initial centers during step 0. For this reason it may be advisable to repeat the analysis by changing the initial centers in order to check how stable the solution is.

Behavioral Segmentation Analysis

Non Hierarchical Cluster Analysis: *K-Means algorithm*

- Part II: application aspects

Traditional approach

- Input Data Matrix Definition: $X_{(n,p)}$
- Factor analysis
- Cluster analysis
- Cross-Tabulation

1. Quality of ingredients – Qualità degli ingredienti
2. Genuineness – Genuinità
3. Lightness – Leggerezza
4. Taste – Sapore/Gusto
5. Nutritional features – Caratteristiche nutrizionali
6. Specific needs – Attenzione a bisogni specifici
7. Natural yeast process – Lievitazione naturale
8. Handicraft – Produzione artiginale
9. Shape – Forma e stampo
10. Reference to tradition – Richiamo alla tradizione
11. Size of the package – Grandezza della confezione
12. Functionality of the package – Funzionalità della confezione
13. Attractiveness of the package – Estetica della confezione
14. Expire date – Scadenza
15. Brand name – Nome del biscotto
16. Advertising – Pubblicità e comunicazione
17. Promotions – Promozione e offerte speciali
18. Various tips – Consigli per l'utilizzo
19. Price – Prezzo
20. Brand awareness – Notorietà della marca

Rotated Component Matrix[a]

	Component					
	1	2	3	4	5	6
Genuinità	.795	-.089	-.123	.237	-.051	.178
Leggerezza	.748	.072	-.007	.096	.050	-.104
Qualità degli ingredienti	.716	-.026	.078	.080	.007	.437
Caratteristiche nutrizionali	.619	.312	.009	.111	-.127	-.349
Attenzione a bisogni specifici	.327	-.054	.243	.324	.020	-.239
Promozioni e offerte speciali	.002	.799	-.052	-.111	.286	.035
Prezzo	-.015	.764	-.063	.180	.154	.092
Grandezza della confezione (peso netto)	.017	.697	.250	.006	-.067	.159
Funzionalità della confezione	.158	.448	.334	.165	-.028	-.219
Forma e stampo	-.011	.163	.799	.070	-.024	.137
Estetica della confezione	-.096	.065	.704	.107	.268	-.076
Nome del biscotto	.071	-.040	.624	.005	.309	-.047
Produzione artigianale	.158	.028	.083	.836	-.023	.172
Lievitazione naturale	.369	-.103	.224	.681	.094	-.065
Scadenza	.066	.211	-.137	.593	.078	-.086
Richiamo alla tradizione	.023	.082	.439	.566	.132	.251
Notorietà della marca	-.083	.108	.103	.161	.811	.051
Pubblicità e comunicazione	-.002	.139	.310	-.055	.764	-.119
Consigli per l'utilizzo	.282	.342	.228	.234	.394	-.064
Sapore/gusto	.048	.163	.025	.083	-.074	.793

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 6 iterations.

6. Interpretation:

The correlation structure between Input Vars & Factors improves after the rotation step

Factor Analysis

- Once an adequate solution is found, it is possible to use the obtained factors as new macro variables to consider for further analyses on the phenomenon under investigation, thus replacing the original variables;
- Again taking into consideration the example, we may add six new variables into the data file, as follows:
  - Health,
  - Convenience & Practicality,
  - Image,
  - Handicraft,
  - Communication,
  - Taste.
- They are standardized variables: zero mean and variance equal to one.
- They will be the input for the next step: Cluster analysis.

Traditional approach

- Input Data Matrix definition: $X_{(n,p)}$
- Factor analysis
- Cluster analysis => K-Means Algorithm
- Cross-Tabulation

Interpretation of the output

- The elements needed for the choice and interpratation for the solution of a non hierachical closet analysis are:

1) *Number of observations (objects) in each cluster*
2) *Variance analysis chart*
3) *Characteristics of the final centers*

- In case of one or more of the above criteria should not provide adequate enough directions, we have to repeat the cluster again by changing the number of cluster needed.

Number of observations in each cluster

- The number of observations in each cluster must be possibly the same and in anycase not lower than a fixed threshold (i.e. not lower than5% of overall elements).

- The presence of clusters made of a very low number of units might indicate the presence of some outliers which should be eliminated before carrying out the procedure or be dealt separately

Number of observations in each cluster

- The chart below shows the number of cases in the clusters built up considering 6 factors and 6 clusters

Number of Cases in each Cluster

Cluster	1	47.000
	2	80.000
	3	31.000
	4	2.000
	5	18.000
	6	32.000
Valid		210.000
Missing		11.000

Number of observations in each cluster

- The chart below shows the number of cases in the clusters built up considering 6 factors and 5 clusters

Number of Cases in each Cluster

Cluster	1	2.000
	2	48.000
	3	69.000
	4	46.000
	5	45.000
Valid		210.000
Missing		11.000

Number of observations in each cluster

- We have to eliminate the detected outliers and to repeat the Factor Analysis

Rotated Component Matrix[a]

	Component					
	1	2	3	4	5	6
Genuinità	.806					
Qualità degli ingredienti	.782					
Leggerezza	.758					
Caratteristiche nutrizionali	.609	.328				
Promozioni e offerte speciali		.803				
Prezzo		.776				
Grandezza della confezione (peso netto)		.706				
Funzionalità della confezione		.438		.375		
Consigli per l'utilizzo		.378			.371	.312
Produzione artigianale			.851			
Lievitazione naturale	.328		.653			.360
Richiamo alla tradizione			.616	.452		
Scadenza			.549			
Forma e stampo				.818		
Estetica della confezione				.681		
Nome del biscotto				.592	.360	
Notorietà della marca					.809	
Pubblicità e comunicazione				.308	.772	
Sapore/gusto						-.799
Attenzione a bisogni specifici						.541

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 6 iterations.

Number of observations in each cluster

- The charts below show the analysis considering respectively 6 and 5 clusters eliminating the two outliers.

Number of Cases in each Cluster

Cluster	1	15.000
	2	47.000
	3	35.000
	4	33.000
	5	33.000
	6	45.000
Valid		208.000
Missing		.000

Number of Cases in each Cluster

Cluster	1	44.000
	2	37.000
	3	70.000
	4	40.000
	5	17.000
Valid		208.000
Missing		.000

Variance analysis output

- The variables whose p-value of the test F is lower than a pre-defined threshold (usually 5%) have means which are statistically very different in the final clusters

- A good suitable solution of a cluster analysis is when all the p-values of test F related to the input variables are lower than the pre-defined threshold

- Generally speaking, by encreasing the number of clusters the overall quality of the analysis improves, even though this does not necessary mean that the p-value of a specific variable decreases.

Variance analysis output

- **The variance analysis chart enables to assess the goodness of the clusterization, below is a case with 6 Clusters**

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
REGR factor score 1 for analysis 3	10.418	5	.767	202	13.586	.000
REGR factor score 2 for analysis 3	9.908	5	.780	202	12.710	.000
REGR factor score 3 for analysis 3	17.395	5	.594	202	29.275	.000
REGR factor score 4 for analysis 3	20.991	5	.505	202	41.553	.000
REGR factor score 5 for analysis 3	21.750	5	.486	202	44.716	.000
REGR factor score 6 for analysis 3	21.297	5	.498	202	42.798	.000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Variance analysis output

- **The variance analysis chart enables to assess the goodness of the clusterization, below is a case with 5 Clusters**

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
REGR factor score 1 for analysis 3	24.775	4	.532	203	46.611	.000
REGR factor score 2 for analysis 3	10.750	4	.808	203	13.307	.000
REGR factor score 3 for analysis 3	23.078	4	.565	203	40.847	.000
REGR factor score 4 for analysis 3	17.441	4	.676	203	25.798	.000
REGR factor score 5 for analysis 3	4.299	4	.935	203	4.597	.001
REGR factor score 6 for analysis 3	28.869	4	.451	203	64.029	.000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Variance analysis output

- If, in spite of increasing number of clusters, the p-value of a specific variable doesn't improve (decreases), that means such variable is not relevant in the analysis, therefore it is advisible to eliminate it in the clusterization process

- The variables whose p-values are the lowest are the most relevant in the marketing segmentation stage.

Variance analysis output

- **The variance analysis chart enables to assess the goodness of the clusterization, below is a case with 4 Clusters**

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
REGR factor score 1 for analysis 3	15.699	3	.784	204	20.028	.000
REGR factor score 2 for analysis 3	6.287	3	.922	204	6.817	.000
REGR factor score 3 for analysis 3	3.145	3	.968	204	3.247	.023
REGR factor score 4 for analysis 3	27.032	3	.617	204	43.800	.000
REGR factor score 5 for analysis 3	43.723	3	.372	204	117.622	.000
REGR factor score 6 for analysis 3	27.798	3	.606	204	45.878	.000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Final centers features

- The chart on the final centers shows the cluster means for each variable used in the process.

- Through this chart, it is possibile to spot the features of the clusters with regards to the used variables and therefore it is possible to evaluate the marketing significance of the built-up clusters.

Final centers features

Final Cluster Centers

	Cluster					
	1	2	3	4	5	6
REGR factor score 1 for analysis 3	-.85411	.32722	.02025	-.24193	-.70935	.62480
REGR factor score 2 for analysis 3	-.35454	.39045	-.23091	-.33065	-.70087	.64643
REGR factor score 3 for analysis 3	-.39857	.44949	-1.27680	.03443	.72998	.09590
REGR factor score 4 for analysis 3	.24098	.82788	-.09907	.75638	-.86399	-.78903
REGR factor score 5 for analysis 3	.07659	.81293	.52205	-1.24693	.27039	-.56449
REGR factor score 6 for analysis 3	2.32439	-.06192	-.22847	-.63170	-.46098	.26888

1. Health
2. Convenience & Practicality
3. Image
4. Handicraft
5. Communication
6. Taste

Final centers features

Final Cluster Centers

	Cluster				
	1	2	3	4	5
REGR factor score 1 for analysis 3	.39205	-.52138	.74112	-.83067	-.97710
REGR factor score 2 for analysis 3	-.49692	.42914	.44752	-.46137	-.40504
REGR factor score 3 for analysis 3	-1.16912	.22017	.45509	.51892	-.54813
REGR factor score 4 for analysis 3	.00707	1.07853	-.21895	-.74191	.28153
REGR factor score 5 for analysis 3	-.03331	-.39403	-.06841	.52609	-.01235
REGR factor score 6 for analysis 3	-.31293	-.76266	.29341	-.37423	2.14223

1. Health
2. Convenience & Practicality
3. Image
4. Handicraft
5. Communication
6. Taste

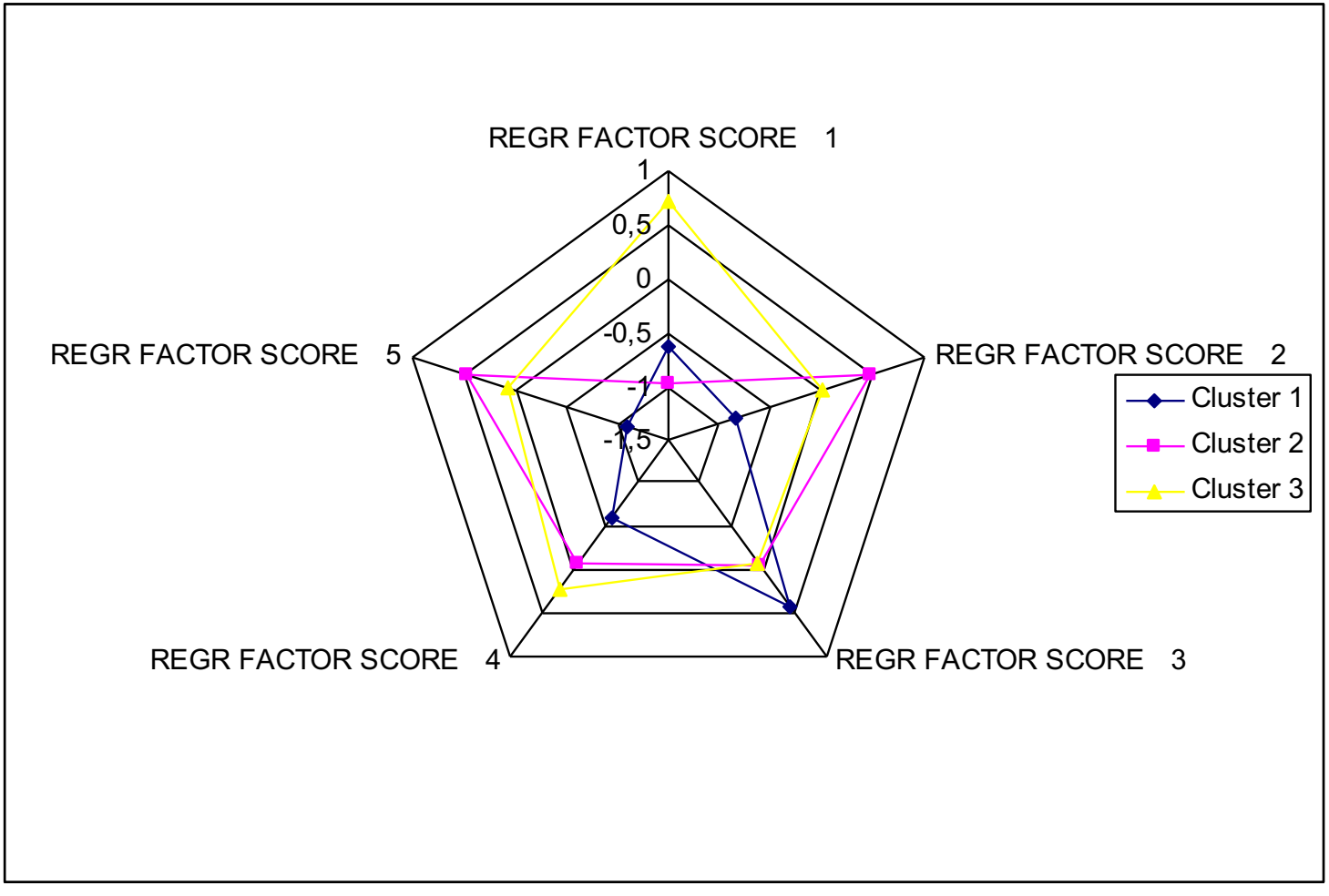
Final centers features

- The chart on the final centers can be easily summed-up by sorting out the clusters acorrding to the mean values that each variale (factor) assumes in the clusters. By so doing we highlight the cluster in which the variable (factor) is most significant. As for the interpretation we have to read through the chart by column.
- In the example below, cluster 3 is the one with the highest mean (ranking=1) on factors 1 and 4.

Final Cluster Centers / Ranking for each variable			
	Cluster		
	1	2	3
REGR FACTOR SCORE 1	2	3	1
REGR FACTOR SCORE 2	3	1	2
REGR FACTOR SCORE 3	1	2	3
REGR FACTOR SCORE 4	3	2	1
REGR FACTOR SCORE 5	3	1	2

Final centers features

- Another way of showing the final centers chart is through a radar-like graph



Cluster actionability

- Once we have secured the final solution, which best meets the three assessment criteria, we can create a new variable that indicates which cluster each unit belongs to.

- Such new variable may be used to describe the clusters by matching it with variables related to personal data in order to gauge the actual marketing effectiveness of those segments and consequently suggest the most adequate commercial strategies.

Cluster actionability

Crosstab

			Cluster Number of Case					Total
			1	2	3	4	5	
Sesso	M	Count	16	18	27	12	8	81
		% within Sesso	19.8%	22.2%	33.3%	14.8%	9.9%	100.0%
		% within Cluster Number of Case	44.4%	40.9%	46.6%	31.6%	25.0%	38.9%
		% of Total	7.7%	8.7%	13.0%	5.8%	3.8%	38.9%
	F	Count	20	26	31	26	24	127
		% within Sesso	15.7%	20.5%	24.4%	20.5%	18.9%	100.0%
		% within Cluster Number of Case	55.6%	59.1%	53.4%	68.4%	75.0%	61.1%
		% of Total	9.6%	12.5%	14.9%	12.5%	11.5%	61.1%
Total		Count	36	44	58	38	32	208
		% within Sesso	17.3%	21.2%	27.9%	18.3%	15.4%	100.0%
		% within Cluster Number of Case	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
		% of Total	17.3%	21.2%	27.9%	18.3%	15.4%	100.0%

Agenda

- Behavioral Segmentation
  - Traditional Approach
- Hierarchical Methods
  - Methodological aspects
  - Application aspects

Behavioral Segmentation

Traditional protocol entails:

- The collection of significance assessment on or about specific features of the products/services surveyed

- The synthesis in underline variables through the Factor analysis

- The built-up of homogeneous groups through the Cluster analysis using the underlines variables obtained in the previous step

- The cross between the clusters and the social and demographic data of the target population

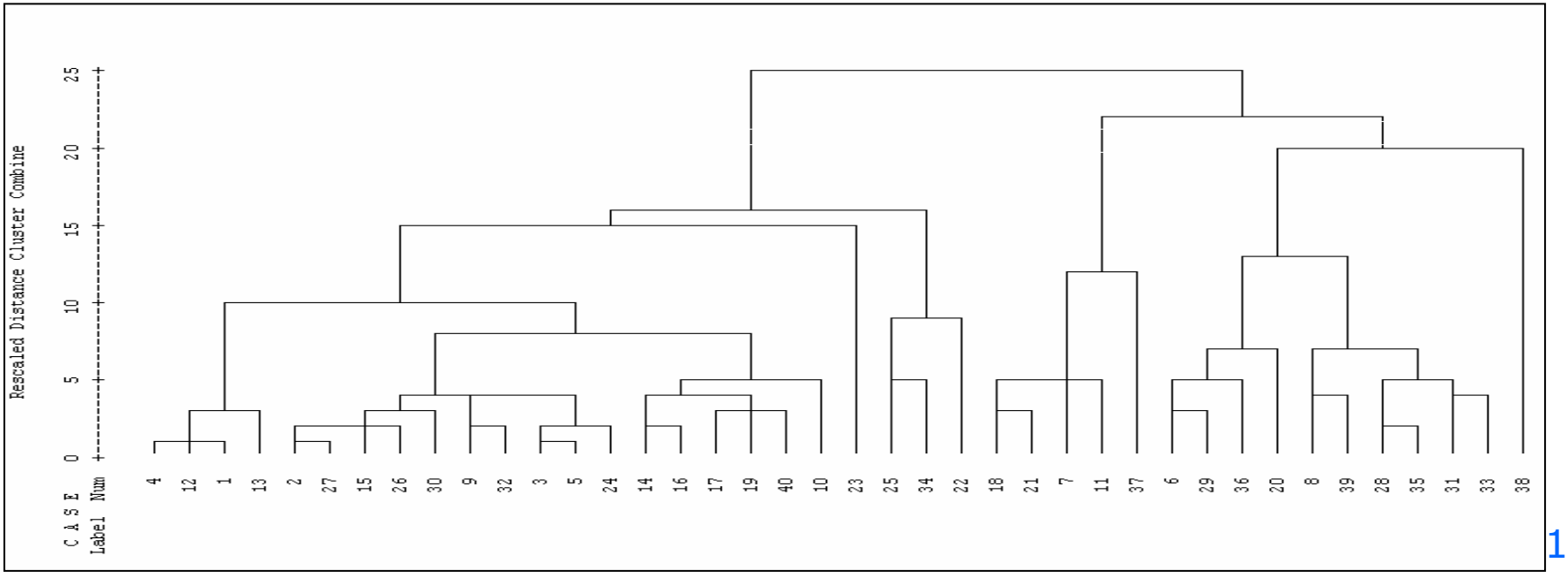
Hierarchical algorithms

- The hierarchical algorithms are developed sequentially, following a series of steps:
  - in every step two “units” (or groups of “units”) which are most homogeneous to each other according to the chosen distance measure (similarity) are joined.

- As soon as a “unit” is given to a specific cluster, this unit cannot be eventually allocated to a different cluster.

Hierarchical algorithms

- The most widespread algorithms are the agglomerative algorithms:
  - Let's start by presupposing that each unit is a cluster
  - We proceed by collecting "units" in increasingly larger groups
  - The aggregation process ends when all of the "units" are collected in just one cluster



Hierarchical algorithms

The major methodological steps in order to carry out a hierarchical cluster analysis are:

- Define a distance measure (similarity) between the pairs of units which we wish to classify
- Build up the Square Matrix **D** of the distances (similarity) among of all of the pairs of units
- Define a rule by which it is possible to decide which units and/or which groups should be joined/aggregated in every step of the algorithm process

Classification algorithm

- Now let's take a look at each step regarding the implementation of hierarchical cluster analysis:
  - *STEP 0:* We consider *n* "units" as split subsets of set *I*, that is as *n* primary clusters.
  - *STEP 1:* On the basis of the Square Matrix **D**, two units *i* and *j* are joined which have the shortest distance (denoted by d_{ij} of **D**).
    - By so doing the first "composite" cluster is set up.
    - Once this step is over, we therefore have *n*-2 primary clusters (the single units left) and the composite cluster "*i*,*j*".

Classification algorithm

- *STEP 2:* The Square Matrix **D** are updated, to keep track of the new aggregation
  - The columns related to the “units” *i and j* are eliminated.
  - A new column is added with the distances d_{ij},k between the composite cluster "*i*,*j*" and the *n*-2 remaining clusters.

- *STEPS from* 3 to *n*-1: The steps 1 e 2 are iterated until the *n* primary “units” get aggregated to a only one cluster with all of the primary units.

Grouping Criteria

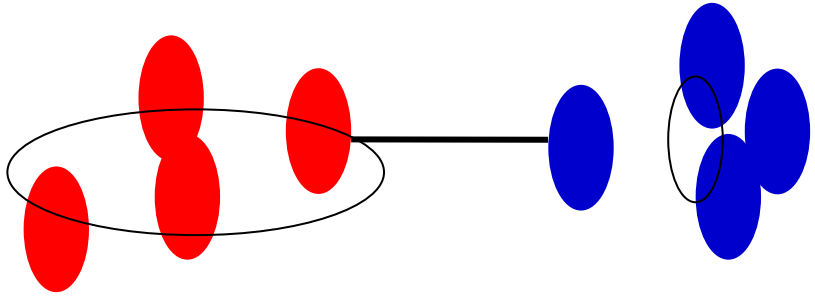
The aggregation among groups can be carried out by using several union criteria:

- Aggregation or Linkage Methods
- Centroid-based Methods
- Variance Methods

Aggregation Methods

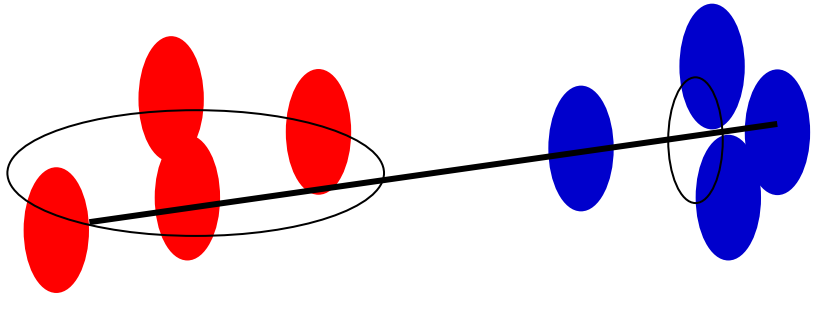
The aggregation methods operate on pairs of units and cause the aggregation among units based on the distance between them.

- Single Linkage
  - At every step of the aggregation process the units (or clusters) with the shortest distance are joined into a new cluster, whereas the distance between two groups is the actual distance between the two **nearest** elements in each cluster.



Aggregation Methods

- Complete Linkage
  - At every step of the aggregation process the units (or clusters) with the shortest distance are joined into a new cluster, whereas the distance between two groups is the actual distance between the two **furthest** elements in each cluster.



Centroid-based Methods

These methods use information considering all of the elements of the different clusters and not just the pairs of units

- Average Linkage
  - The distance between the clusters *Ci* and *Cj* is calculated as the average distance between each unit of the cluster C_i and each unit of the cluster C_j.

- Centroid
  - The distance between two clusters is calculated as the distance between the clusters centers.

Variance Methods

In order to identify the groups we may use criteria based on the maximization of the variability among the groups and minimization of the variability within the groups.

- Ward
  - For each cluster the means for all the variables are computed (centroid of the cluster).
  - The squared Euclidean distance between each unit and its cluster centroid is computed.
  - These distances are summed up by cluster.
  - At each step the two clusters (units), which produce the smallest increase in the overall sum of square within clusters distances, are joined.

This method is mostly efficient when the input variables are factors obtained by principal components method.

Centroid Method

numerical example

Input Data Matrix
5 observations and 2 variables

Obs	X1	X2
1	2	7
2	4	6
3	8	3
4	6	5
5	2	6

Objective of the analysis:
To classify the observations into groups according to an agglomerative criteria based on the Centroid Method

Squared Euclidean Distance Matrix between all the input observations

Obs	1	2	3	4	5
1	0	5	52	20	1
2		0	25	5	4
3			0	8	45
4				0	17
5					0

Cluster C1 => obs: 1,5

Centre C1=> X1: 2 - X2: 6.5

In the first Cluster we have obs. N°1 and obs N°5 because they have the smallest distance.
Once we create the Cluster, we calculate its center (the mean values of the input variables)

Centroid Method

numerical example

Input Data Matrix
4 observations and 1 Cluster: C1

Obs	X1	X2
C1	2	6,5
2	4	6
3	8	3
4	6	5

We substitute obs N°1 and obs N°5 with the centre values of Cluster C1.

Squared Euclidean Distance Matrix between Cluster C1 and the remaining observations

Obs	C1	2	3	4
C1	0	4,25	48,25	18,25
2		0	25	5
3			0	8
4				0

Cluster C2 => obs: C1,2

Centre C2=> X1: 3 - X2: 6.25

In the second Cluster (C2) we join cluster C1 and obs. N°2 because they have the smallest distance. Once we've created the new Cluster, we calculate its center (the mean values of the input variables)

Centroid Method

numerical example

Input Data Matrix
2 observations and 2 Clusters: C1 e C2

Obs	X1	X2
C2	3	6,25
3	8	3
4	6	5

We substitute cluster C1 and obs N°2 with the centre values of Cluster C2

Squared Euclidean Distance Matrix between Cluster C2 and the remaining observations

Obs	C2	3	4
C2	0	35,56	10,56
3		0	8
4			0

Cluster C3 => obs: 3,4

Centre C3=> X1: 7 - X2: 4

In the third Cluster (C3) we put obs. N°3 and obs. N°4 because they have the smallest distance. Once we've created the Cluster, we calculate its center (the mean values of the input variables)

Centroid Method

numerical example

Input Data Matrix
2 Clusters: C2 e C3

Obs	X1	X2
C2	3	6,25
C3	7	4

We substitute obs. N°3 and obs. N°4 with the center values of the input variables

Squared Euclidean Distance Matrix between Cluster C2 and Cluster C3

Obs	C2	C3
C2	0	21,06
C3		0

Cluster C4 => obs: C2, C3

Centre C4=> X1: 5 - X2: 5.125

In the last Cluster (C4) we join Cluster C3 and Cluster C4 and the classification process stops.

Output interpretation

- The main tool for the output interpretation of hierarchic cluster analysis is the *dendrogram,* which can be very effective to visualize the various steps of the aggregation process

- The horizontal line (vertical) shows the union of two clusters. The position of such horizontal lines in the diagram shows the distances to which these cluster are aggregated.

Centroid Method

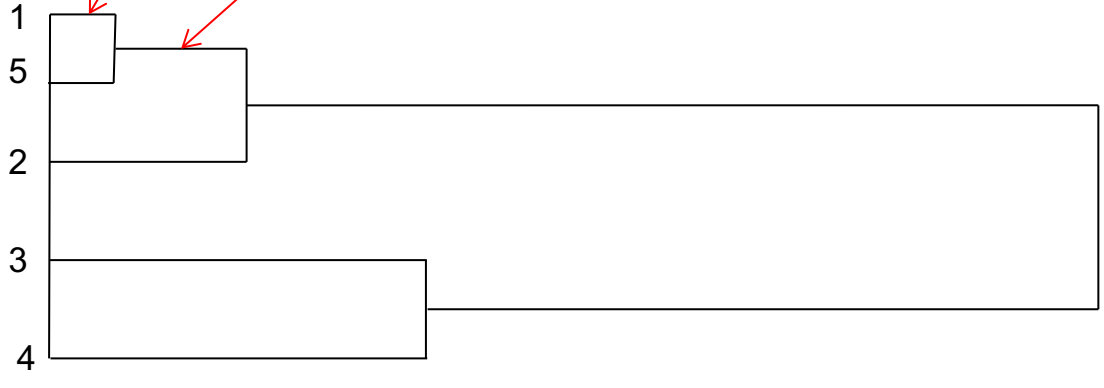
numerical example

Obs	1	2	3	4	5
1	0	5	52	20	1
2		0	25	5	4
3			0	8	45
4				0	17
5					0

Obs	C1	2	3	4
C1	0	4,25	48,25	18,25
2		0	25	5
3			0	8
4				0

Obs	C2	3	4
C2	0	35,56	10,56
3		0	8
4			0

Obs	C2	C3
C2	0	21,06
C3		0



Dendogram

Output interpretation

- The higher is the distance the longer will be the branches of the dendogram

- Cut the dendogram considering the length of the branches

Centroid Method

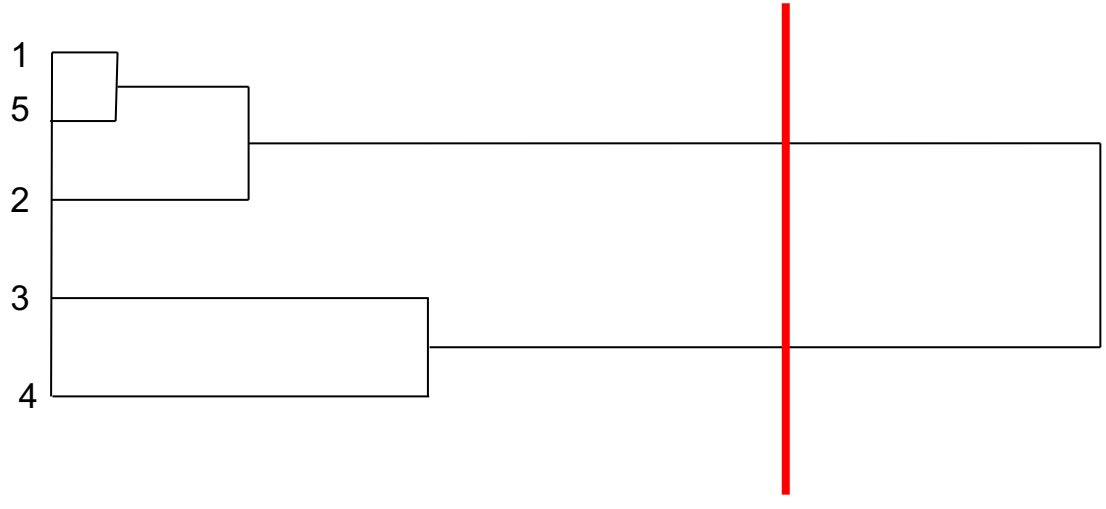
numerical example

Obs	1	2	3	4	5
1	0	5	52	20	1
2		0	25	5	4
3			0	8	45
4				0	17
5					0

Obs	C1	2	3	4
C1	0	4,25	48,25	18,25
2		0	25	5
3			0	8
4				0

Obs	C2	3	4
C2	0	35,56	10,56
3		0	8
4			0

Obs	C2	C3
C2	0	21,06
C3		0



Cut the Dendogram where the distances are the biggest ones

Centroid Method

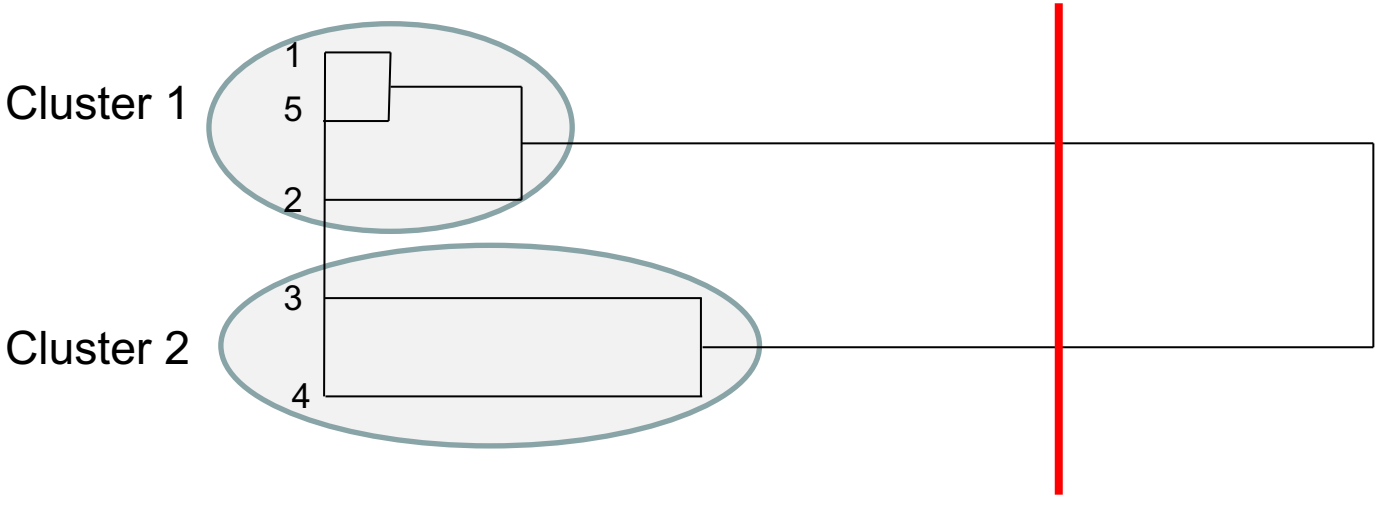
numerical example

Obs	1	2	3	4	5
1	0	5	52	20	1
2		0	25	5	4
3			0	8	45
4				0	17
5					0

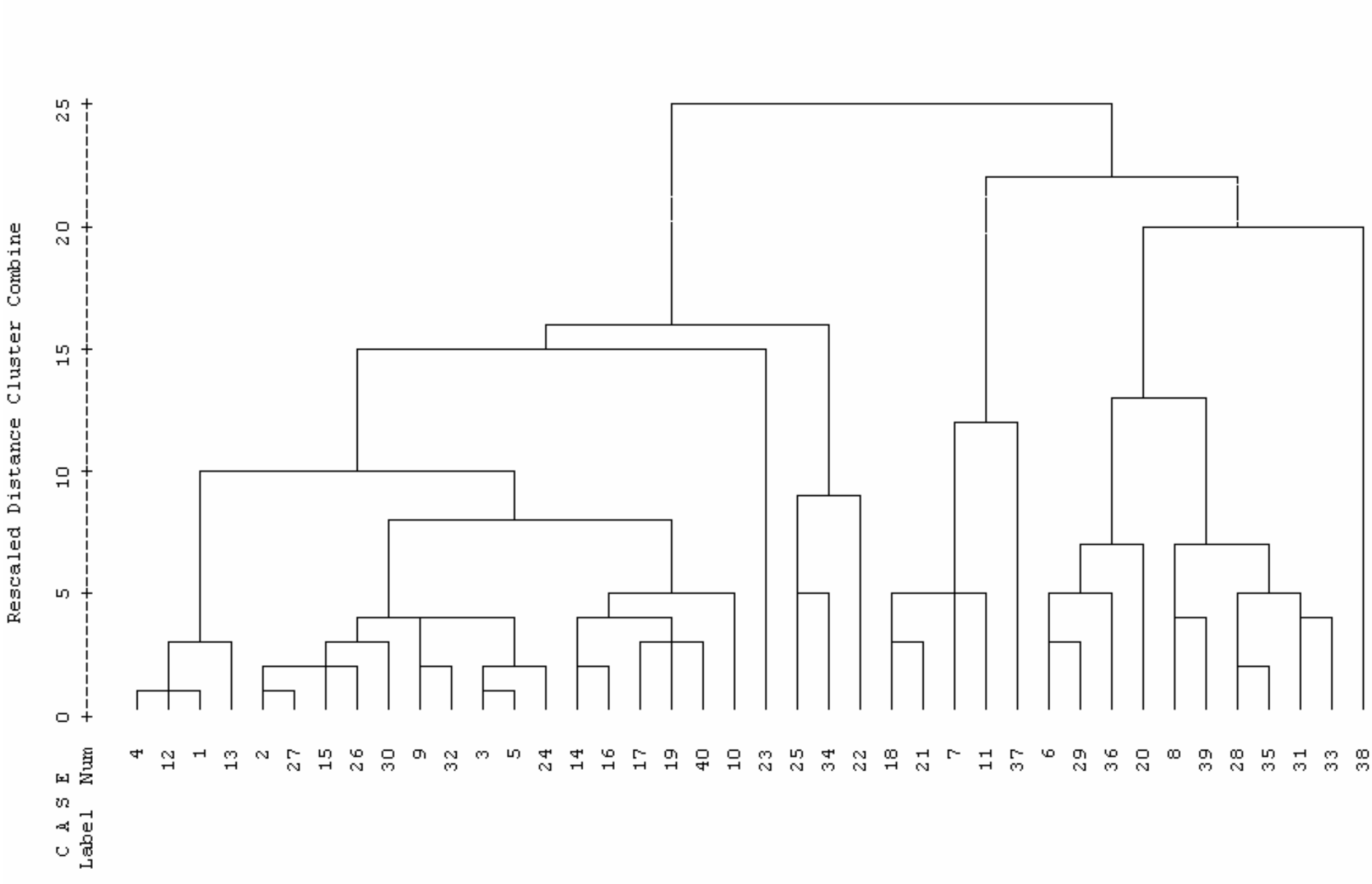
Obs	C1	2	3	4
C1	0	4,25	48,25	18,25
2		0	25	5
3			0	8
4				0

Obs	C2	3	4
C2	0	35,56	10,56
3		0	8
4			0

Obs	C2	C3
C2	0	21,06
C3		0



Dendrogram


Rescaled Distance Cluster Combine
0
5
10
15
20
25
CASE
Label Num
4
12
1
13
2
27
15
26
30
9
32
3
5
24
14
16
17
19
40
10
23
25
34
22
18
21
7
11
37
6
29
36
20
8
39
28
35
31
33
38

Dendrogram

- The horizontal line (vertical) shows the union of two clusters. The position of such horizontal lines in the diagram show the distances to which these cluster are aggregated.

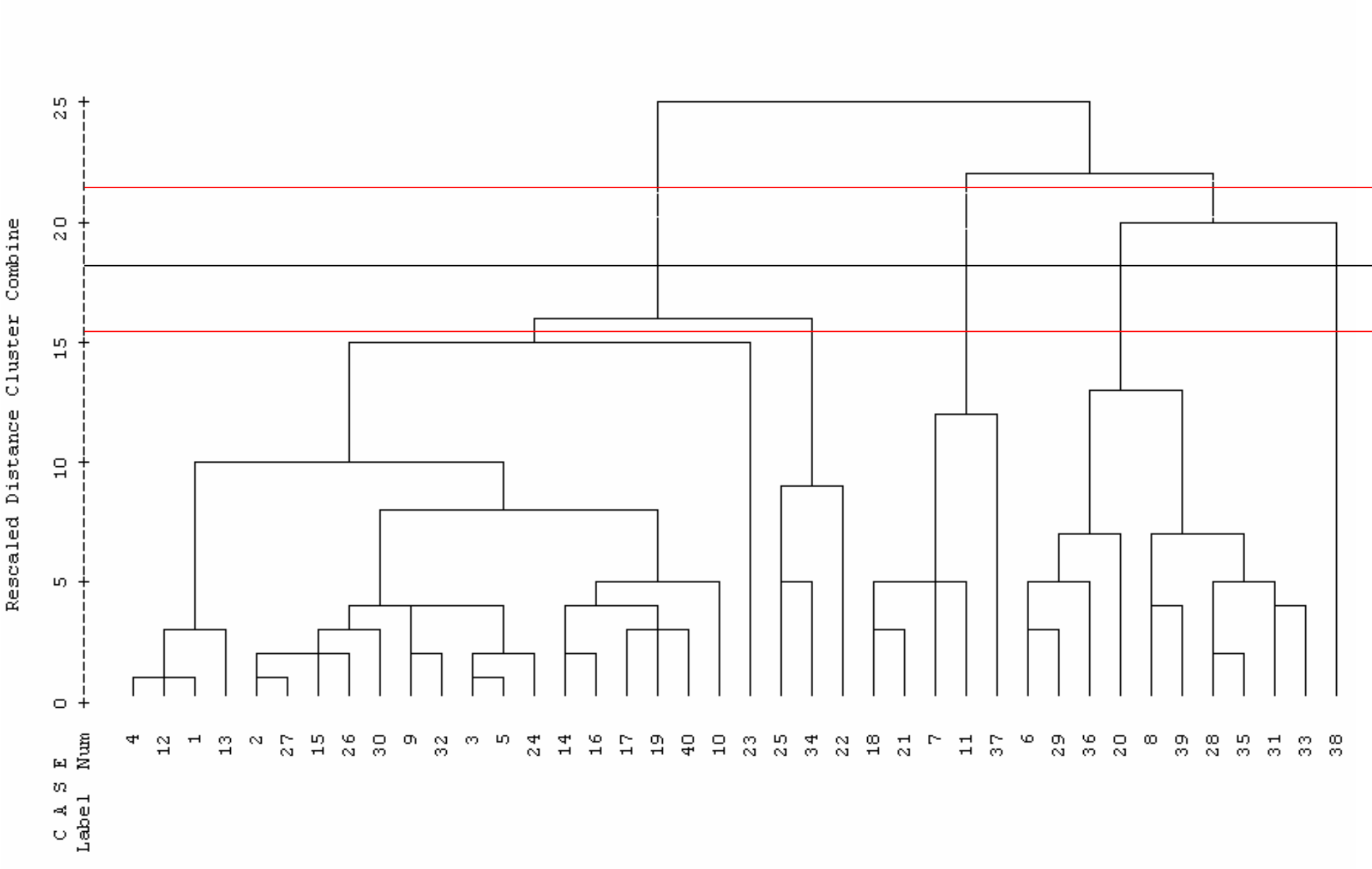
- For the sake of simplicity the distances are rearranged in the dendogram having the lowest distance equal to 1 and the highest distance equal to 25

Dendrogram

- The identification of a given number of clusters can be achieved by cutting through the dendogram horizontally (or vertically) and the clusters can be identified by moving from right to left

- In the example shown the clients who are aggregated in the initial steps of the procedure are:
  - client n° 4 with 12 and 1
  - client n° 2 with 27
  - client n° 3 with 5
  - In the final step the cluster with cases from 4 to 22 is aggregated to the cluster with the remaining cases from 18 to 38

Dendrogram


Rescaled Distance Cluster Combine
0
5
10
15
20
25
CASE
Label Num
4
12
1
13
2
27
15
26
30
9
32
3
5
24
14
16
17
19
40
10
23
25
34
22
18
21
7
11
37
6
29
36
20
8
39
28
35
31
33
38

Output interpretation

- In the dendogram shown in the example before we can see that third aggregation from below (which identifies 4 clusters) is at relative high distance between clusters compared to both the previous aggregation (which brings the clusters down to 5) and the following one (which brings the clusters down to 3)

- So we can say that the optimal solution in terms of groups seems to be 4

- However one of the 4 clusters consists only of one element (case 38) and therefore it could be subjected to presence of an outlier.

Output interpretation

- The goodness of the solution proposed through the analysis of the dendogram may be supported considering the analysis of the variance of the variables used to create the groups (ANOVA SPSS procedure)

- In case the number of units is relatively high (over 60) we can use the output which shows the agglomeration program.

Output interpretation

- It shows the complete and detailed sequence of the agglomeration procedure indicating:
  - the code of the cases aggregated step by step the so called (aggregated *cluster*),
  - the aggregation distance (*coefficients*),
  - the phase in which the case has already been aggregated into a previous step (Stage Cluster First Appears - S*tadio di formazione dei cluster*) and the next phase in which the newly formed cluster will again be aggregated (Next stage - *stadio successivo*)

Programma di agglomerazione

Stadio	Cluster accorpati		Coefficienti	Stadio di formazione del cluster		Stadio successivo
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	4	12	293.000	0	0	2
2	1	4	856.000	0	1	14
3	3	5	1809.000	0	0	6
4	2	27	3121.500	0	0	7
5	9	32	4871.500	0	0	18
6	3	24	6922.500	3	0	18
7	2	15	8993.333	4	0	8
8	2	26	11285.000	7	0	12
9	14	16	13988.500	0	0	21
10	28	35	16776.000	0	0	24
11	17	19	19829.000	0	0	15
12	2	30	23024.000	8	0	20
13	18	21	26386.500	0	0	25
14	1	13	29859.000	2	0	32
15	17	40	33343.333	11	0	21
16	6	29	37068.333	0	0	22
17	8	39	41330.833	0	0	29
18	3	9	45799.633	6	5	20
19	31	33	50646.633	0	0	24
20	2	3	55712.233	12	18	30
21	14	17	61014.200	9	15	27
22	6	36	66522.533	16	0	28
23	25	34	72175.533	0	0	31
24	28	31	77952.783	10	19	29
25	7	18	83790.283	0	13	26
26	7	11	90227.783	25	0	33
27	10	14	96926.317	0	21	30
28	6	20	105936.233	22	0	34
29	8	28	115213.650	17	24	34
30	2	10	124607.417	20	27	32
31	22	25	136331.750	0	23	36
32	1	2	148566.350	14	30	35
33	7	37	163900.450	26	0	38
34	6	8	180736.033	28	29	37
35	1	23	199798.100	32	0	36
36	1	22	219435.017	35	31	39
37	6	38	244787.244	34	0	38
38	6	7	272119.854	37	33	39
39	1	6	304518.575	36	38	0

Output interpretation

Flexible approach

- Input Scenario Cards: $X_{(n,p)}$
- Conjoint Analysis
- Cluster analysis => Hierarchical Algorithms
- Cross-Tabulation

Keywords

- Behavioral Segmentation
- Traditional Approach
  - Factor + Cluster
- Flexible Approach
  - Conjoint + Cluster
- Cluster Analysis
  - K Means Algorithm
  - Hierarchical Algorithms
- Distance Function
- Similarity Function

- K Means Algorithm
  - Number of Cases in each Cluster
  - Analysis of Variance (ANOVA) Chart
  - Final Cluster Centers Chart

- Hierarchical Algorithm
  - Distance Matrix
  - Aggregation Methods
    - Single Linkage
    - Complete Linkage
    - Average Linkage
    - Centroid
    - Ward
  - Dendogram

Text Book

Naresh K. Malhotra, "*Marketing Research – An Applied Orientation*", Pearson – Prentice Hall, 2010

- Chapter 20 – pag 660-687